Office for National Statistics

COVID-19 Infection Survey: methods and further information

This methodology guide is intended to provide information on the methods used to collect the data, process it, and calculate the statistics produced from the COVID-19 Infection Survey.

Contact:
Kara Steel and Philippa Haughton
infection.survey.analysis@ons.gov.uk
+44 1633 651689

Release date:
24 August 2021

Next release:
To be announced

Table of contents

1 . COVID-19 Infection Survey

The coronavirus (COVID-19) pandemic is having a profound impact across the UK. In response to the pandemic, the COVID-19 Infection Survey measures:

- how many people across England, Wales, Northern Ireland and Scotland test positive for COVID-19 infection at a given point in time, regardless of whether they report experiencing symptoms
- the average number of new positive test cases per week over the course of the study
- the number of people who test positive for antibodies

The results of the survey contribute to the Scientific Advisory Group for Emergencies (SAGE) estimates of the rate of transmission of the infection, often referred to as "R". The survey also provides important information about the socio-demographic characteristics of the people and households who have contracted COVID-19.

The Office for National Statistics (ONS) is working with the University of Oxford, University of Manchester, Public Health England, Wellcome Trust, IQVIA and the Lighthouse laboratory at Glasgow to run the study, which was launched in mid-April 2020 as a pilot in England. We have expanded the size of the sample over August to October 2020 and from 31 October 2020 reported headline figures for all four UK nations.

This methodology guide is intended to provide information on the methods used to collect the data, process it, and calculate the statistics produced from the COVID-19 Infection Survey. We will continue to expand and develop methods as the study progresses, and we will publish an updated methodology guide when needed.

It can be read alongside:

- the COVID-19 bulletin, which gives weekly headline statistics
- the CIS Antibody and Vaccination data bulletin
- the Characteristics of people testing positive for COVID-19 bulletin
- the Quality and Methodology Information (QMI), which details the strengths and limitations of the data and methods used.
- the study protocol, which outlines the study design and rationale
- the study guide, which explains to participants what taking part in the study entails - we also provide translations of the study guide

2 . Study design: sampling

The sample for the survey in England, Wales and Scotland is drawn from the AddressBase, which is a commercially available list of addresses maintained by the Ordnance Survey. In Northern Ireland, the sample is selected by the Northern Ireland Statistics and Research Agency (NISRA) from people who have participated in NISRA and Office for National Statistics (ONS) surveys and have consented to be contacted again. This means that in all four countries only private households are included in the study. People living in care homes, other communal establishments and hospitals are not included.

We include children over the age of 2 years, adolescents and adults in the survey. Children are included because it is essential to understand the prevalence and the incidence of symptomatic and asymptomatic infection in children. This is particularly important for informing policy decisions around schools. Further information on the prevalence of coronavirus (COVID-19) in schools can be found in our latest release from the COVID-19 Schools Infection Survey.

Initially, adults aged over 16 years from around 20% of our household sample were asked to provide a blood sample as well as a swab sample. To monitor the impact of vaccination on individual and community immunity and infection, this was increased, and from February 2021 we have asked adults from a larger but still representative sample of households recruited to the study to give blood samples at their monthly visits. We also ask all individuals from any household where anyone has tested positive on a nose and throat swab to give blood samples. We now ask everyone to stay in the study until April 2022 (to have additional visits beyond their original 12-month study period). Blood samples are used to test for the presence of antibodies to the coronavirus (COVID-19).

The sample size has grown as the survey has expanded. At the start of the pilot stage of the study, we invited about 20,000 households in England to take part, anticipating that this would result in approximately 21,000 individuals from approximately 10,000 households participating. At the pilot stage of the study, all respondents to the COVID-19 Infection Survey were individuals who had previously participated in the Annual Population Survey, an ONS social survey, which means the number of ineligible addresses in the sample was substantially reduced.

Since August 2020, we expanded the survey to invite a random sample of households from the AddressBase. Fieldwork increased in England, and coverage of the study was extended to include Wales, Northern Ireland and Scotland. Survey fieldwork in Wales began on 29 June 2020 and we started reporting headline figures for Wales on 7 August 2020. Survey fieldwork in Northern Ireland began on 26 July 2020 and we started reporting headline figures for Northern Ireland on 25 September 2020. Survey fieldwork in Scotland began on 21 September 2020 and we started reporting headline figures for Scotland on 31 October 2020.

Ultimately, the swab target is to achieve approximately 150,000 individuals with swab test results at least every fortnight from October 2020 onwards in England, approximately 9,000 in Wales, approximately 5,000 in Northern Ireland and approximately 15,000 in Scotland (approximately 179,000 total across the UK). The blood target is to achieve up to 125,000 people with blood test results every month in England, and up to 7,500, 5,500 and 12,000 per month in Wales, Northern Ireland and Scotland respectively (approximately 150,000 in total across the UK). The absolute numbers reflect the relative size of the underlying populations.

More information about how participants are sampled can be found in the study protocol. We publish up-to-date information on sample size and response rates for all four countries in our technical dataset.

Response rates

Likelihood of enrolment decreases over time since the original invitation letter was sent, and so response rate information for those initially asked to take part at the start of the survey in England can be considered as final. We provide response rates separately for the different sampling phases of the study. These response rates along with commentary are found in the technical datasets.

Note response rates from different sampling phases are not comparable. The more recent response rates cannot be regarded as final since those who are invited are not given a time limit in which to respond, and because we aim to recruit households continuously to meet our fortnightly targets (rather than recruit everyone who registers immediately).

Technical table 2a: UK

Provides a summary of the total number of households registered and eligible individuals in registered households for the UK.

Technical table 2b: England

Provides a summary of the response rates for England, by the different sampling phases of the survey:

- Table A presents response rates for those asked to take part at the start of the survey, sampled from previous ONS studies
- Table B presents response rates for those invited from the end of May 2020, sampled from previous ONS studies

Tables A and B can be considered as relatively final as the likelihood of enrolment decreases over time:

- Table C presents response rates for those invited from the end of July 2020, from a randomly sampled list of addresses, where enrolment is continuing

Technical table 2c: Wales

Provides a summary of the response rates for Wales by the different sampling phases of the survey:

- Table A presents response rates for those invited from the end of June 2020, sampled from previous ONS studies
- Table B presents those asked to take part from the beginning of October 2020, from a randomly sampled list of addresses

Technical table 2d: Northern Ireland

Provides a summary of the response rates for Northern Ireland.

Technical table 2e: Scotland

Provides a summary of the response rates for Scotland.

Technical table 2f: swabs per day

Provides information on the number of swabs taken per day since the study began.

Attrition

To produce reliable and generalisable estimates, the survey sample should reflect the diversity of the population under investigation. For this reason, it is important we retain sample members who agree to participate for the duration of the study. For various reasons, some sample members are unreachable, withdraw their participation or drop out of the study. If those who drop out of the sample are significantly different from those who remain, it will affect researchers' ability to produce estimates that are generalisable to the target population. We monitor the number of people who drop out of the sample to mitigate the potential risks caused by attrition.

3 . Study design: data we collect

Nose and throat swab

We take nose and throat swabs to test for the presence of SARS-CoV-02, the virus that causes coronavirus (COVID-19). To do this, laboratories use a real-time reverse transcriptase polymerase chain reaction test (RT-PCR), not a lateral flow test. We ask everyone aged 2 years or older in each household to have a nose and throat swab, regardless of whether anyone is reporting symptoms or not. Those aged 12 years and older take their own swabs using self-swabbing kits, and parents or carers use the same type of kits to take swabs from their children aged between 2 and 11 years old. This is to reduce the risk to the study health workers and to respondents themselves.

We need to know more about how the virus is transmitted in individuals who test positive on nose and throat swabs; whether individuals who have had the virus can be re-infected symptomatically or asymptomatically; and about incidence of new positive tests in individuals who have not been exposed to the virus before.

To address these questions, we collect data over time. Every participant is swabbed once; participants are also invited to have repeat tests every week for the first five weeks as well as monthly. Initially this was for a period of 12 months. In May 2021, existing participants were invited to remain in the study until April 2022 and new participants were invited to take part in the study until this date.

The protocol offers more detailed information about when and how we collect data. Information about how we process nose and throat swabs is found in Section 4: Processing the data.

Blood sample

We collect blood samples from a randomly selected subsample of adults aged 16 years or older to test for antibodies, which help us to assess the number of people who have been infected in the past, and the impact of the vaccination programme at both the population and the individual level. Participants give 0.5 millilitres of blood using a capillary finger prick method undertaken by the participant and demonstrated by a specially trained fieldworker. The blood samples are taken at enrolment and then every month.

The protocol offers more detailed information about when and how we collect data. Information about how we process the blood sample data is found in Section 4: Processing the data. Blood tubes are kept in a cool bag during the day, and then sent to the University of Oxford overnight. Residual blood samples will be stored by the University of Oxford after testing where consent is given for this.

Survey data

We use the Coronavirus Infection Survey questionnaire to collect information from each participant, including those aged under 16 years. We collect information about their socio-demographic characteristics, any symptoms that they are experiencing, whether they are self-isolating, their occupation, how often they work from home, and whether the participant has come into contact with someone who they suspect has COVID-19. We also ask participants questions about their experiences of the pandemic, including questions about long COVID, whether participants have been vaccinated, how they travel to work, number of contacts with different amounts of physical and social distancing, and whether participants smoke.

Each participant in a household that agrees to participate are provided with an individual identifier. This allows for the differentiation of data collected between each household member.

Swabs and blood are labelled with a barcode, which is linked to the participant's individual identifier on the study database.

4 . Processing the data

Nose and throat swabs

The nose and throat swabs are sent to the Lighthouse laboratory at Glasgow. Here, they are tested for SARS-CoV-2 using reverse transcriptase polymerase chain reaction (RT-PCR). This is an accredited test that is part of the national testing programme. Swabs are discarded after testing. The virus genetic material from a selection of positive samples is sent for whole genome sequencing at Public Health England, to find out more about the different types of virus and variants of virus circulating in the UK.

If a test is positive, the positive result is linked to the date that the swab was taken, not to the date that the swab was analysed in the laboratory.

The RT-PCR test looks for three genes present in coronavirus: N protein, S protein and ORF1ab. Each swab can have one, two or all three genes detected. In the laboratories used in the survey, RT-PCR for three SARS-CoV-2 genes (N protein, S protein and ORF1ab) uses the Thermo Fisher TaqPath RT-PCR COVID-19 Kit, analysed using UgenTec FastFinder 3.300.5, with an assay-specific algorithm and decision mechanism that allows conversion of amplification assay raw data from the ABI 7500 Fast into test results with minimal manual intervention. Samples are called positive if at least a single N-gene and/or ORF1ab are detected (although S-gene cycle threshold (Ct) values are determined, S-gene detection alone is not considered sufficient to call a sample positive). We estimate a single Ct value as the arithmetic mean of Ct values for genes detected (Spearman correlation >0.98 between each pair of Ct values). More information on how swabs are analysed can be found in the study protocol.

The Cycle threshold (Ct) value is the number of cycles that each polymerase chain reaction (PCR) test goes through before a positive result is detectable. If there is a high quantity of the virus present, a positive result will be identified after a low number of cycles. However, if there is only a small amount of the virus present, then it will take more cycles to detect it. These values are used as a proxy for the quantity of the virus, also known as the viral load. The higher the viral load, the lower the Ct value. These values are helpful for monitoring the strength of the virus and for identifying patterns that could suggest changes in the way the virus is transmitting. We provide the Ct values of COVID-19 positive tests in the technical dataset. In our analysis, such as symptoms analysis, we define a ‘strong positive’ as a swab with a Ct value of less than 30.

RT-PCR from nose and throat swabs may be falsely negative, because of their quality or the timing of collection. The virus in nose and throat secretions peak in the first week of symptoms but may decline below the limit of detection in patients who present with symptoms beyond this time frame. For people who have been infected and then recovered, the RT-PCR technique provides no information about prior exposure or immunity. To address this, we also collect blood samples to test for antibodies.

Variants

The Alpha variant (B.1.1.7) of coronavirus (COVID-19) identified in the UK in mid-November 2020 has changes in one of the three genes that COVID-19 swab tests detect, known as the S-gene. This means that in cases compatible with this variant, the S-gene is not detected by the current test. Therefore Alpha (B.1.1.7) has the pattern ORF1ab+N (S-gene negative) in our variant analysis. Other variants, including both Delta (B.1.617.2) and Beta (B.1.351), are positive on all three genes, with the pattern ORF1ab+S+N. Almost all ORF1ab+S+N cases in the UK will now be the Delta variant, so this group is labelled “compatible with the Delta variant” in our analysis.

We try to read all letters of the virus' genetic material for every positive nose and throat swab with sufficient virus to do so (Ct less than 30). This is called whole genome sequencing. Sequencing is not successful on all samples that we test, and sometimes only part of the genome is sequenced. This is especially so for the higher Ct values, which are common in our data as we often catch people early or late in infection when viral loads tend to be lower (and hence Ct values are higher).

Where we successfully sequence over half of the genome, we use the sequence data to work out which type of variant is present in each virus. This method can tell us which variant might be responsible for any potential increase in cases, which are either “compatible with the Delta variant” or “compatible with the Alpha variant”. However, because we cannot get a sequence from every positive result, there is more uncertainty in these estimates.

These data are provided in our technical dataset using the international standard labels.

Sequencing analysis is produced by research partners at the University of Oxford and by Public Health England. Of particular note are Dr Katrina Lythgoe, Dr David Bonsall, Dr Tanya Golubchik, and Dr Helen Fryer.

More information on how we measure variants from positive tests on the survey can be found in our blog Understanding COVID-19 variants – What can the Coronavirus Infection Survey tell us?.

Blood samples

Blood samples are tested for antibodies, which are produced to fight the virus. We measure the presence of antibodies in the community population to understand who has had coronavirus (COVID-19) in the past, and the impact of vaccinations. It takes between two and three weeks after infection or vaccination for the body to make enough antibodies to fight the infection. Having antibodies can help to prevent individuals from getting infected again, or if they do get infected, they are less likely to have severe symptoms. Once infected or vaccinated, antibodies remain in the blood at low levels and can decline over time. The length of time antibodies remain at detectable levels in the blood is not fully known.

To test blood for antibodies, we use an ELISA for immunoglobulins IgG, based on tagged and purified recombinant SARS-CoV-2 trimeric spike (S) protein. From March 2021, we also test samples for IgG immunoglobulins against the nucleocapsid (N) protein to try to distinguish between those with immunity due to natural infection (which would be anti-S and anti-N positive) and vaccination (anti-S positive, but anti-N negative because vaccine produce antibodies to spike only).

The threshold for antibody positivity in the blood is 42 ng/ml nanograms per millilitre. A negative test result will occur if there are no antibodies or if antibody levels are too low to reach this threshold. It is important to draw the distinction between testing positive for antibodies and having immunity meaning having a lower risk of getting infected or infected again.

Following infection or vaccination, antibody levels can vary and sometimes increase but are still below the level identified as "positive" in our test, and other tests. This does not mean that a person has no protection against COVID-19, as an immune response does not rely on the presence of antibodies alone. We also do not yet know exactly how much antibodies need to rise to give protection. A person's T cell response will provide protection but is not detected by blood tests for antibodies. A person's immune response is affected by a number of factors, including health conditions and age. Additional information on the link between antibodies and immunity and the vaccine programme can be found in our blog What the ONS can tell you about the COVID-19 Vaccine programme.

The study protocol includes more information about swab and blood sample procedure and analysis.

Survey data

As in any survey, some data can be incorrect or missing. For example, participants and interviewers sometimes misinterpret questions or skip them by accident. We ran a pilot study before the full study, through which we have learnt how to improve the wording of the questions and the questionnaire structure. To minimise the impact of incorrect or missing data, we clean the data, by editing or removing data that are clearly incorrect. For example, when a participant leaves their job blank we take their previous answer instead, but only if they say they have not changed their job and we correct the misspelled names of countries that people say they have travelled to.

5 . Test sensitivity and specificity

Understanding false-positives and false-negative results

The estimates provided in the Coronavirus (COVID-19) Infection Survey bulletin are for the percentage of the private-residential population testing positive for coronavirus (COVID-19), otherwise known as the positivity rate. We do not report the prevalence rate. To calculate the prevalence rate, we would need an accurate understanding of the swab test's sensitivity (true-positive rate) and specificity (true-negative rate).

Our data and related studies provide an indication of what these are likely to be. To understand the potential impact, we have estimated what prevalence would be in two scenarios using different possible test sensitivity and specificity rates.

Test sensitivity

Test sensitivity measures how often the test correctly identifies those who have the virus, so a test with high sensitivity will not have many false-negative results. Studies suggest that sensitivity may be somewhere between 85% and 98%. A recent study considering tests in the Lighthouse labs estimated that this is most likely to be around 95%.

Our study involves participants self-swabbing under the supervision of a study healthcare worker. It is possible that some participants may take the swab incorrectly, which could lead to more false-negative results. However, research suggests that self-swabbing under supervision is likely to be as accurate as swabs collected directly by healthcare workers.

Test specificity

Test specificity measures how often the test correctly identifies those who do not have the virus, so a test with high specificity will not have many false-positive results.

We know the specificity of our test must be very close to 100% as the low number of positive tests in our study over the summer of 2020 means that specificity would be very high even if all positives were false. For example, in the six-week period from 31 July to 10 September 2020, 159 of the 208,730 total samples tested positive. Even if all these positives were false, specificity would still be 99.92%.

We know that the virus was still circulating at this time, so it is extremely unlikely that all these positives are false. However, it is important to consider whether many of the small number of positive tests we do have might be false. There are two main reasons we do not think that is the case.

Symptoms are an indication that someone has the virus; but are reported in a minority of participants at each visit. We might expect that false-positives would not report symptoms or might report fewer symptoms (because the positive is false). Overall, therefore, of the positives we find, we would expect to see most of the false-positives would occur among those not reporting symptoms. If that were the case, then risk factors would be more strongly associated with symptomatic infections than without reported symptoms infections. However, in our data the risk factors for testing positive are equally strong for both symptomatic and asymptomatic infections.

Assuming that false-positives do not report symptoms, but occur at a roughly similar rate over time, and that amongst true-positives the ratio with and without symptoms is approximately constant, then high rates of if false-positives would mean that, the percentage of individuals not reporting symptoms among those testing positive would increase when the true prevalence is declining because the total prevalence is the sum of a constant rate of false-positives (all without symptoms) and a declining rate of true-positives (with a constant proportion with and without symptoms).

More information on sensitivity and specificity is included in Community prevalence of SARS-CoV-2 in England: Results from the ONS Coronavirus Infection Survey Pilotby the Office for National Statistics' academic partners. You can find additional information on cycle thresholds in a paper written by our academic partners at the University of Oxford.

The impact on our estimates

We have used Bayesian analysis to calculate what prevalence would be in two different scenarios, one with medium sensitivity and the other with low sensitivity. Table 1 shows these results alongside the weighted estimate of the percentage testing positive in the period from 6 September to 19 September 2022.

Scenario 1 (medium sensitivity, high specificity)

Table 1: The effects of test sensitivity on estimates

Reference period: 6 to 19 September 2020		95% credible interval	
		Lower	Upper
Estimated average percentage of the population who had COVID-19 (weighted)	0.22%	0.18%	0.26%
Prevalence rate in Scenario 1 (medium sensitivity, high specificity)	0.22%	0.17%	0.29%
Prevalence rate in Scenario 2 (low sensitivity, high specificity)	0.34%	0.24%	0.49%

Source: Office for National Statistics – Coronavirus (COVID-19) Infection Survey

Based on similar studies, the sensitivity of the test used is plausibly between 85% and 95% (with around 95% probability) and, as noted earlier, the specificity of the test is above 99.9%.

Scenario 2 (low sensitivity, high specificity)

To allow for the fact that individuals are self-swabbing, Scenario 2 assumes a lower overall sensitivity rate of on average 60% (or between 45% and 75% with 95% probability), incorporating the performance of both the test itself and the self-swabbing. This is lower than we expect the true value to be for overall performance but provides a lower bound.

The results show that when these estimated sensitivity and specificity rates are taken into account, the prevalence rate would be slightly higher but still very close to the main estimate presented in Section 2 of the Coronavirus (COVID-19) Infection Survey bulletin. This is the case even in Scenario 2, where we use a sensitivity estimate that is lower than we expect the true value to be. For scenario 2, prevalence is higher because this scenario is based on an unlikely assumption that the test misses 40% of positive results. For this reason, we do not produce prevalence estimates for every analysis, but we will continue to monitor the impacts of sensitivity and specificity in future.

Evaluation of the test sensitivity and specificity of five immunoassays for SARS-CoV-2 serology, including the ELISA assay used in our study, has shown that this assay has sensitivity and specificity (95% confidence interval) of 99.1% (97.8 to 99.7%) and 99.0% (98.1 to 99.5%) respectively; compared with 98.1% (96.6 to 99.1%) and 99.9% (99.4 to 100%) respectively for the best performing commercial assay.

6 . Analysing the data

The primary objective of the study is to estimate the number of people in the population who test positive for coronavirus (COVID-19) on nose and throat swabs, with and without symptoms.

The analysis of the data is a collaboration between the Office for National Statistics (ONS) and researchers from the University of Oxford and University of Manchester, Public Health England and Wellcome Trust. Our academic collaborators aim to publish an extended account of the modelling methodology outside the ONS bulletin publication in peer-reviewed articles, on topics including:

- Symptoms and SARS-CoV-2 positivity
- Community prevalence of SARS-CoV-2 in England
- Cycle threshold (Ct) values and positivity,
- the Alpha variant, identified in the UK in mid-November 2020
- Rates of seroconversion in NHS staff
- Risks of Coronavirus Transmission from community household data (PDF, 681KB)

A full list of articles and academic papers by our academic collaborators can be found on the Nuffield Department of Medicine website.

All estimates presented in our bulletins are provisional results. As swabs are not necessarily analysed in date order by the laboratory, we will not have received test results for all swabs taken on the dates included in the most recent analysis. Estimates may therefore be revised as more test results are included.

7 . Positivity rates

We use several different modelling techniques to estimate the number of people testing positive for SARS-CoV-2, the virus that causes coronavirus (COVID-19). As well as our headline figures, we provide estimates of the number of people testing positive for infection broken down by different characteristics (age, region and so on). This section provides further information on our modelling techniques.

Bayesian multi-level regression poststratification (MRP) model

A Bayesian multi-level regression post-stratification (MRP) model is used to produce our headline estimates of positivity on nose and throat swabs for each UK country as well as our breakdowns of positivity by region and age group in England. This produces estimated daily rates of people testing positive for COVID-19 controlling for a number of factors described in this section. Details about the methodology are also provided in the peer-reviewed paper from our academic collaborators published in the Lancet Public Health.

As the number of people testing positive (known as the positivity rate) is unlikely to follow a linear trend, time measured in days is included in the model using a non-linear function (thin-plate spline). Time trends are allowed to vary between regions by including an interaction between region and time. Given the low number of positive cases in the sample, the effect of time is not allowed to vary by other factors.

The models for the positivity rate for each country use all available swab data from participants from their respective country within time periods to estimate the number of people who are currently infected by COVID-19. We use a Bayesian multi-level generalised additive model with a complementary log-log link.

The COVID-19 infection survey is based on a nationally representative survey sample; however, some individuals in the original Office for National Statistics (ONS) survey samples will have dropped out and others will not have responded to the survey. To address this and reduce potential bias, the regression models adjust the survey results to be more representative of the overall population in terms of age, which is a fixed effect, and sex and region, which are as random intercepts (region is only adjusted for in the England model). This is called “post-stratification”. The regression models do not adjust for ethnicity, household tenure or household size, because we do not have the underlying denominators across the UK for these characteristics.

The data that are modelled are drawn from a sample, and so there is uncertainty around the estimates that the model produces. Because a Bayesian regression model is used, we present estimates along with credible intervals. These 95% credible intervals can be interpreted as there being a 95% probability that the true value being estimated lies within the credible interval. A wider interval indicates more uncertainty in the estimate.

Sub-regional estimates

Sub-regional estimates for England were first presented on 20 November 2020 and for Wales, Northern Ireland and Scotland on 19 February 2021. As sample sizes vary in local authorities, we pool local authorities together to create COVID-19 Infection Survey sub-regions in Great Britain and we used NHS Health Trusts for Northern Ireland. Sub-regional estimates are obtained from a spatial-temporal MRP model. This is on a similar basis to the dynamic Bayesian MRP model used for national and regional trend analysis that produces estimated daily rates of people testing positive for COVID-19 controlling for age and sex within sub-regions. Spatial-temporal in this context means the model borrows strength geographically and over time, meaning that the model implicitly expects rates to be more similar in neighbouring areas, and within an area over time. For our sub-regional analysis, we run two models: one for Great Britain and the other for Northern Ireland. Our academic partners from the University of Oxford have developed this spatiotemporal MRP methodology outside the ONS bulletin publication in a peer-reviewed article.

Initially for England, sub-regional estimates were produced using three-day groupings aggregated to a six-day period. However, because of falling numbers of positive cases and smaller sample sizes in some sub-regions, we have changed to seven-day groupings to provide more accurate estimates for all countries of the UK; these were presented for the first time on 12 February 2021.

Age analysis by category for England

We first presented our daily modelled estimates by age category for England on 11 September 2020 and refined our age categories on 2 October 2020. Our current age categories are:

- "age 2 years to school Year 6" includes those children in primary school and below
- "school Year 7 to school Year 11" includes those children in secondary school
- "school Year 12 to age 24 years" includes those young adults who may be in further or higher education
- age 25 to age 34 years
- age 35 to age 49 years
- age 50 to age 69 years
- age 70 years and above

Our current age categories separate children and young people by school age. This means that 11- to 12-year-olds have been split between the youngest age categories depending on whether they are in school Year 6 or 7 (birthday before or after 1 September). Similarly, 16- to 17-year-olds are split depending on whether they are in school Years 11 or 12 (birthday before or after 1 September). Splitting by school year rather than age at last birthday reflects a young person's peers and therefore more accurately reflects their activities both in and out of school.

The model used to produce our daily estimates by age category for England is similar to the model used to calculate our daily positivity estimates. We post-stratify the estimates so that results are adjusted to reflect the underlying population sizes.

We started publishing results from this updated model on 20 August 2021. Previously, the model presented the estimated level of infection using the East Midlands as a representative reference region. Therefore, previous results from our age category model are not comparable with national headline positivity estimates. The previous model also did not include the same interaction terms with time.

Methodology used to produce single year age over time estimates by UK country

To assess swab positivity over time by single year of age, we used generalised additive models (GAM) with a complementary loglog link and tensor product smooths. The latter allows us to incorporate smooth functions of age and time, where the effect of time is allowed to be different dependent on age. Tensor product smooths generally perform better than isotropic smooths when the covariates of a smooth are on different scales, for example, age in years and time in days.

The Restricted Maximum Likelihood (REML) criterion was used to optimize the smoothness of the curve given the observed data. The analyses are based on the most recent eight weeks of data on swab positivity among individuals aged 2 to 80 years. The effect of age and time are allowed to vary by region, but marginal probabilities and their confidence intervals are obtained for the whole of England. Separate models are run for England, Wales, Scotland and Northern Ireland.

8 . Incidence

The incidence of new infections (the number of new infections in a set period of time) helps us understand the rate at which infections are growing within the population and supports our main measure of positivity (how many people test positive at any time, related to prevalence) to provide a fuller understanding of the coronavirus (COVID-19) pandemic.

The incidence rate is different to the R number, which is the average number of secondary infections produced by one infected person and is produced by the Scientific Pandemic Influenza Group on Modelling (SPI-M), a sub-group of the Scientific Advisory Group for Emergencies (SAGE).

Current method for calculating incidence

We calculate the incidence of PCR-positive cases (related to the incidence of infection) from the Bayesian Multilevel Regression and Poststratification (MRP) model of positivity, using further detail from our sample. Because we test participants from a random sample of households every day, our estimate of positivity is unbiased providing we correct for potential non-representativeness due to non-participation by post-stratifying for age, sex and region.

We use information from people who ever test positive in our survey (from 1 September 2020) to estimate how long people test positive for. We apply information from this group to the whole of the sample and produce an estimate for incidence for the whole of the household population. We estimate the time between the first positive test and the last time a participant would have tested positive (the "clearance" time) using a statistical model. We do this accounting for different times between visits.

With these clearance time estimates we can then model backwards, deducing when new positives occurred in order to generate the positivity estimate. This method uses a deconvolution approach developed by Joshua Blake, Paul Birrell and Daniela De Angelis at the MRC Biostatistics Unit and Thomas House at the University of Manchester. Posterior samples from the MRP model over the last 100 days are used in this method.

Clearance time is the length of time that an individual remains positive. We use an estimate of the distribution of clearance times derived from the COVID-19 Infection Survey, which varies by the date a participant first tests positive. The distribution of clearance times is estimated by modelling the time from an individual's first positive test in the COVID-19 Infection Survey. Only first positive tests from 1 September 2020 onwards are included in this model, given the very low rates of positivity observed over the summer of 2020, and the small numbers before this time.

Clearance time considers the sequence of positive and negative test results of an individual:

- the clearance time for individuals testing negative, following a positive test, is modelled as occurring at some point between their last positive and first negative test
- intermittent negatives (consisting of three or fewer consecutive negative tests) between positive tests within 120 days of their previous positive test are ignored as this is considered a single episode of infection in that period, following World Health Organization guidance
- new positives that occur more than 120 days after an individual's previous positive test are treated as a new episode of infection providing the participant has one or more immediately preceding negative tests, as are new positives that occur after four consecutive negatives
- participants who are last seen positive are censored at their last positive test

The estimated distribution of clearance times is modelled using flexible parametric interval censored survival models, choosing the amount of flexibility in the model based on the Bayesian Information Criterion. We allow the distribution of clearance times to change according to the date a participant first tests positive.

There is a bias in estimating the clearance distribution because the analysis used to estimate how long a person stays positive only starts from their first positive test. Since (most) people will have become positive on an earlier day, this will bias the clearance curves downwards (making the estimates too short). However, there is another bias due to the survey missing positive episodes entirely if they are short. This means that our dataset has fewer short positive episodes than in the population as a whole, and that the sample used to run the survival analysis is biased towards people with longer positive episodes. This will bias the clearance curves upwards (making the estimates too long). We include whether the first positive a participant had in the survey was their first test in the study, and if not, how many days their last negative test was previously as explanatory variables. There is no evidence that either of these variables are associated with clearance time, and we have therefore used the overall estimate.

The estimate of the incidence of PCR-positive cases (relating to the incidence of infection) is produced by combining a posterior sample from the Bayesian MRP positivity model with the estimated distribution of the clearance times, allowing for the fact that some people will remain positive for shorter or longer times than others. Once the distribution of clearance is known, we compute a deterministic transformation (known as deconvolution) of the posterior of the positivity. The resulting sample gives the posterior distribution of the incidence of PCR-positive cases.

We calculate incidence estimates based on the MRP positivity model for the entire period of data in the MRP positivity model but we present it excluding the first two weeks. This is to avoid boundary effects (at the start of the positivity model, infections will have happened at various points in the past).

The official estimate of incidence is the estimate from this model at the reference date. The reference date used for our official estimates of incidence is 14 days before the end of the positivity reference day. This is necessary as estimates later than this date are more subject to change as we receive additional data.

This method of estimating incidence enables us to estimate incidence for Wales, Scotland and Northern Ireland, as well as for England, as we can assume the same clearance distribution across all countries.

Figure 1 compares the previously published official estimates of incidence for England between 4 September 2020 and 28 November 2020 (points with credible intervals in chart) to an indicative incidence estimate based on the new method (solid line in chart).

Figure 1: Comparison of official estimates of incidence based on the old model compared with estimates based on the new model

Estimated numbers of new PCR-positive COVID-19 cases in England, based on nose and throat swabs with modelled estimates from 4 September 2020 to 28 November 2020

Download the data

.xlsx

Previous method for incidence

Estimates for incidence from 13 July 2020 to 28 November 2020 considered every day that each participant was in the study from the date of their first negative test to the earlier of their latest negative test or the greater of seven days before or halfway between their last negative test and first positive test, which are called days at risk (for a new positive test in the study). Each new positive was considered to represent an infection starting at the mid-point between the day of the test and the previous negative swab or at seven days before the day of the test, whichever was closest to the first positive test. This is because we do not know the exact point when the infection occurred and infections only last so long. We excluded everyone whose first swab test in the study was positive, so this method looked at new positives found during the study. Each week our incidence model used a Bayesian Multilevel Regression and Poststratification (MRP) model (log link) with thin plate splines to produce a smooth estimate of incidence over the preceding eight weeks of data. The model censored follow up at the start of the reference week. The most recent official estimate of incidence was defined as the estimate on this last day included in the model. The week after, the next estimate was produced using a week of entirely new data, augmented data for the preceding week (due to additional test results being received for the previous reference week), and the same previous data back to seven weeks (eight weeks data in total). Official estimates were not revised using estimates from later models.

We started recruiting participants on 26 April 2020 and started repeating tests on 1 May 2020. Therefore, only data from 11 May 2020 onwards were included in the incidence model, as to be included in the incidence analysis at least two repeated swab test visits are required.

Why was a new method needed?

When enrolled on the survey, participants are swabbed weekly for five weeks and then move to monthly swabbing. Until mid-November 2020, most visits (at least 60% to 75%) were from participants being swabbed weekly providing us with regular and timely updates on the number of new positive tests and the "time at risk". However, because we recruited a large number of people in August to September 2020, the proportion swabbed monthly increased during November 2020. Consequently, the assumption that new positives in the study represented (almost) all new infections was not sustainable, because of the longer gap between visits. Further, our estimates of “days at risk" in the final four weeks in the model increasingly under-estimated time at risk in this period because of the increasing numbers on monthly visits who had not yet had their next visit. As a result, the series became inconsistent. The method of estimation therefore needed changing to account for the pattern of monthly tests.

The new method, detailed above, needed to account for the fact that most survey respondents are on monthly visits and the period of time between tests is long enough to miss a significant proportion of infections.

9 . Antibody and vaccination estimates

We present estimates of antibody positivity and, to provide context for these antibody estimates, we also present estimates of vaccine uptake in the population. Antibody positivity is measured by antibodies to the spike (S) protein. Vaccine uptake is tracked over all visits over time. We validate our self-reported vaccination data in England with data from the National Immunisation Management Service (NIMS), which is the System of Record for the NHS coronavirus (COVID-19) vaccination programme in England. The equivalent of NIMS is currently not included for other countries, so vaccination estimates for Wales, Northern Ireland and Scotland are produced only from Coronavirus (COVID-19) Infection Survey self-reported records of vaccination.

Current method for antibody and vaccination estimates

Modelled antibody and vaccine estimates use a spatial-temporal Integrated Nested Laplace Approximation (INLA) model with post-stratification. Post-stratification is a method to ensure the sample is representative of the population that can be used with modelled estimates to achieve the same objective as weighting. This estimation method is also used to produce sub-regional estimates for swab positivity and is like the multi-level regression model and post-stratification in the way that it uses Bayesian inference to derive an estimate. Spatial-temporal in this context means the model borrows strength geographically and over time. For both antibody and vaccine estimates, we run two separate models: one for Great Britain and the other for Northern Ireland. This reflects the geography of the four countries as Northern Ireland does not share a land border with Great Britain; the geo-spatial model incorporates physical land distance between regions. All models are run on surveillance weeks (a standardised method of counting weeks from the first Monday of each calendar year to allow for the comparison of data year after year and across different data sources for epidemiological data).

The antibodies model for Great Britain is currently run at a regional level and includes ethnicity, vaccine priority age groups, and sex. The antibody model for Northern Ireland is a temporal model (no spatial component) due to lower sample size, and accounts for sex and age in wider groups (16 to 24, 25 to 34, 35 to 49, 50 to 69, 70 years and over).

The vaccines model for Great Britain is currently run at a subregional level and includes ethnicity, vaccine priority age groups, and sex. The vaccine model for Northern Ireland is also run at a subregional level due to a higher number of participants with information about vaccine uptake. The model accounts for sex and age in wider groups (16 to 24, 25 to 34, 35 to 49, 50 to 69, 70 years and over).

To assess antibody positivity over time by single year of age (similar to single year of age swab positivity) we use generalised additive models (GAM) with a complementary log-log link and tensor product smooths, with a spline over study day and age at visit. The analyses are based on the most recent eight weeks of data on antibody positivity among individuals aged 16 years and over. The number of participants aged over 85 years is relatively small so we recode these participants to be aged 85 years, which is a standard technique to reduce outlier influence. Separate models are run for England, Wales, Scotland and Northern Ireland.

From June 2021, we reduced potential bias in our antibody estimates by removing the participants from our analyses who had consented to antibody testing after being invited because an individual in the household had previously tested positive for COVID-19 on a nose and throat swab (under protocol 2.1).

Our research partners at the University of Oxford have published several academic articles on coronavirus (COVID-19) antibodies and vaccinations:

- Impact of Delta on viral burden and vaccine effectiveness
- Antibody response to SARS-CoV-2 vaccines
- Total Effect Analysis of Vaccination on Household Transmission
- Anti-spike antibody response to natural SARS-CoV-2 infection in the general population
- Impact of vaccination on new SARS-CoV-2 infections in the United Kingdom

Previous method for producing 28-day weighted antibodies estimates by country

From 23 October 2020 we presented weighted monthly estimates for the number of people testing positive for antibodies to SARS-CoV-2 and from 3 February 2021 we presented these for rolling 28-day periods in our fortnightly antibody article.

We needed to weight the estimates to reduce potential bias. Although the study is based on a nationally representative survey sample, some individuals will have dropped out and others will not have responded to the study. For England and Wales, we applied weighting to ensure the responding sample is representative of the population in terms of age (grouped), sex, region, and ethnicity. For Northern Ireland and Scotland, we adjusted for age (grouped), sex and region. This is because ethnicity is already well represented in the survey for these devolved administrations.

Why was a new method needed?

The estimates represented an average of a 28-day period. With the speed of the vaccination roll-out, antibody estimates increased rapidly in specific age groups within each 28-day period. This change in the underlying antibody positivity meant we would be continuously underestimating the antibody positivity for each country, region and age group. Therefore, we moved to Integrated Nested Laplace Approximation modelling and post-stratification from 30 March 2021.

10 . Weighting

The 14-day estimates of the number of people who have coronavirus (COVID-19) are based on weighted data to ensure the estimates are representative of the target population in England, Wales, Northern Ireland and Scotland. The study is based on a nationally representative survey sample; however, some individuals in the original Office for National Statistics (ONS) survey samples will have dropped out and others will not have responded to the study.

To address this and reduce potential bias, we apply weighting to ensure the responding sample is representative of the population in terms of age (grouped), sex and region. This is different from the modelled estimates, which use a different method to adjust for potential non-representativeness of the sample through multi-level regression post-stratification (described in Section 7: Positivity rates).

We used to present weighted estimates for antibodies, but now produce post-stratified modelled estimates.

Confidence intervals for estimates

The statistics are based on a sample, and so there is uncertainty around the estimate. Confidence intervals are calculated so that if we were to repeat the survey many times on the same occasion and in the same conditions, in 95% of these surveys the true population value would be contained within the 95% confidence intervals. Smaller intervals suggest greater certainty in the estimate, whereas wider intervals suggest uncertainty in the estimate.

Confidence intervals for weighted estimates are calculated using the Korn-Graubard method to take into account the expected small number of positive cases and the complex survey design. For unweighted estimates, we use the Clopper-Pearson method as the Korn-Graubard method is not appropriate for unweighted analysis.

11 . Confidence intervals and credible intervals

Simple explanations of confidence and credible intervals have been provided in previous sections, nevertheless, there is still some question about the difference between these two intervals. Whether we use credible or confidence intervals, depends upon the type of analysis that is conducted.

Earlier in the article, we mentioned the positivity model is a dynamic Bayesian multi-level regression post stratification model. This type of analysis produces credible intervals that are used to show uncertainty in parameter estimates, because this type of analysis directly estimates probabilities. While, for the 14-day positivity estimates confidence intervals are provided because this is a different type of analysis using what are called frequentist methods. The use of confidence and credible intervals is a direct consequence of the type of statistics used to make sense of the data: frequentist statistics or Bayesian statistics respectively.

The difference between credible intervals and confidence intervals are associated with their statistical underpinnings; Bayesian statistics are associated with credible intervals, whereas confidence intervals are associated with frequentist (classical) statistics. Both intervals are related to uncertainty of the parameter estimate, however they differ in their interpretations.

With confidence intervals, the probability the population estimate lies between the upper and lower limits of the interval is based upon hypothetical repeats of the study. For instance, in 95 out of 100 studies, we would expect that the true population estimate would lie within the 95% confidence intervals. While the remaining five studies would deviate from the true population estimate. Here we assume the population estimate is fixed and any variation is due to differences within the sample in each study. Whereas credible intervals aim to estimate the population parameter from the data we have directly observed, instead of an infinite number of hypothetical samples. Credible intervals estimate the most likely values of the parameter of interest, given the evidence provided from our data. Here we assume the parameter estimates can vary based upon the knowledge and information we have at that moment. Essentially, given the data we have observed there is a 95% probability the population parameter falls within the interval. Therefore, difference between the two concepts is subtle: the confidence interval assumes the population parameter is fixed and the interval is uncertain. Whereas, credible intervals assume the population parameter is uncertain and the interval is fixed.

12 . Statistical testing

Where we have done analysis of the characteristics of people who have ever tested positive for coronavirus (COVID-19), we have used pairwise statistical testing to determine whether there was a significant difference in infection rates between pairs of groups for each characteristic.

The test produces p-values, which provide the probability of observing a difference at least as extreme as the one that was estimated from the sample if there truly is no difference between the groups in the population. We used the conventional threshold of 0.05 to indicate evidence of genuine differences not compatible with chance, although the threshold of 0.05 is still marginal evidence. P-values of less than 0.001 and 0.01 and 0.05 are considered to provide relatively strong and moderate evidence of genuine difference between the groups being compared respectively.

Any estimate based on a random sample rather than an entire population contains some uncertainty. Given this, it is inevitable that sample-based estimates will occasionally suggest some evidence of difference when there is in fact no systematic difference between the corresponding values in the population as a whole. Such findings are known as "false discoveries". If we were able to repeatedly draw different samples from the population, then, for a single comparison, we would expect 5% of findings with a p-value below a threshold of 0.05 to be false discoveries. However, if multiple comparisons are conducted, as is the case in the analysis conducted within the Infection Survey, then the probability of making at least one false discovery will be greater than 5%.

Multiplicity can occur at different levels. For example, in the Infection Survey we have:

- two primary outcomes of interest -- positivity for current infection based on a swab test and positivity for previous infection based on a blood test
- several different exposures of interest (for example, age and sex)
- several exposures with multiple different categories (for example, working location)
- repeated analysis over calendar time

Consequently, the p-values used in our analysis have not been adjusted to control either the familywise error rate (FWER, the probability of making at least one false discovery) or the false discovery rate (FDR, the expected proportion of discoveries that are false) at a particular level. Instead, we focus on presenting the data and interpreting results in the light of the strength of evidence that supports them.

13 . Geographic coverage

Since 20 November 2020, we have presented modelled estimates for the most recent week of data at the sub-national level for England and for Wales, Northern Ireland and Scotland since 19 February 2021. To balance the granularity with the statistical power, we have grouped together groups of local authorities into COVID-19 Infection Survey sub-regions. The geographies are a rule-based composition of local authorities, and local authorities with a population over 200,000 have been retained where possible. For our Northern Ireland sub-regional estimates, our CIS sub-regions are NHS Health Trusts instead of groups of local authorities. The boundaries for these COVID-19 infection Survey sub-regions can be found on the Open Geography Portal.

14 . Analysis feeding into R

The statistics produced by analysis of this survey contribute to modelling, which predicts the reproduction number (R) of the virus.

R is the average number of secondary infections produced by one infected person. The Scientific Pandemic Influenza Group on Modelling (SPI-M), a sub-group of the Scientific Advisory Group for Emergencies (SAGE), has built a consensus on the value of R based on expert scientific advice from multiple academic groups.

15 . Uncertainty in the data

The estimates presented in this bulletin contain uncertainty. There are many sources of uncertainty, but the main sources in the information presented include each of the following.

Uncertainty in the test (false-positives, false-negatives and timing of the infection)

These results are directly from the test, and no test is perfect. There will be false-positives and false-negatives from the test, and false-negatives could also come from the fact that participants in this study are self-swabbing. More information about the potential impact of false-positives and false-negatives is provided in "Sensitivity and Specificity analysis".

The data are based on a sample of people, so there is some uncertainty in the estimates

Any estimate based on a random sample contains some uncertainty. If we were to repeat the whole process many times, we would expect the true value to lie in the 95% confidence interval on 95% of occasions. A wider interval indicates more uncertainty in the estimate.

Quality of data collected in the questionnaire

As in any survey, some data can be incorrect or missing. For example, participants and interviewers sometimes misinterpret questions or skip them by accident. To minimise the impact of this, we clean the data, editing or removing things that are clearly incorrect. In these initial data, we identified some specific quality issues with the healthcare and social care worker question responses and have therefore applied some data editing (cleaning) to improve the quality. Cleaning will continue to take place to further improve the quality of the data on healthcare and social care workers, which may lead to small revisions in future releases.