Google Cloud
Console Platform

Estimating Covid infection rates in England. A look at administrative records, surveys, and Big Data

Applying the Total Survey Error, Total Error Framework, and Fit For Purpose to a crucial measurement topic

Mario Callegaro Ph.D.
User Experience Survey Research Scientist, Google Cloud London

January 15, 2021
Future of Survey Research Conference

Talk outline & presenter bio

Mario Callegaro is User Experience Survey Research Scientist, Google Cloud London.
He works any survey related projects within his organization. He also consults with numerous other internal teams regarding survey design, sampling, questionnaire design and online survey programming and implementation.

Mario has published a book on web surveys, edited a handbook on online panels, and recently is working on the topic of using surveys with Big Data, with a open access chapter published in 2018 in the Palgrave Handbook of Survey Research.

Acknowledgments: Yongwei Yang

1. Three frameworks to evaluate difference sources of population infection rates
2. Administrative records
3. Surveys
4. Big Data
5. Summary
6. References
7. Appendix

This talk focuses on measurement issue for prevalence rates of Covid-19 infections in England

This is a high-stake and super important topic of discussion, but at the same time I acknowledge the sensitivity and emotional toll that Covid has on everybody's lives

Three frameworks to evaluate difference sources of population infection rates

1

Total Survey Error (TSE)

(Biemer 2010)

- Sampling error
- Non sampling error

2

Total Error Framework (TEF) for Big Data

(Amaya, Biemer & Kinyon, 2020; Biemer & Amaya 2020)

- Identify appropriate data sources
- Extract, transform, load the data

3

Fit for purpose

(Baker et al, 2013)

Relevant utility concepts:

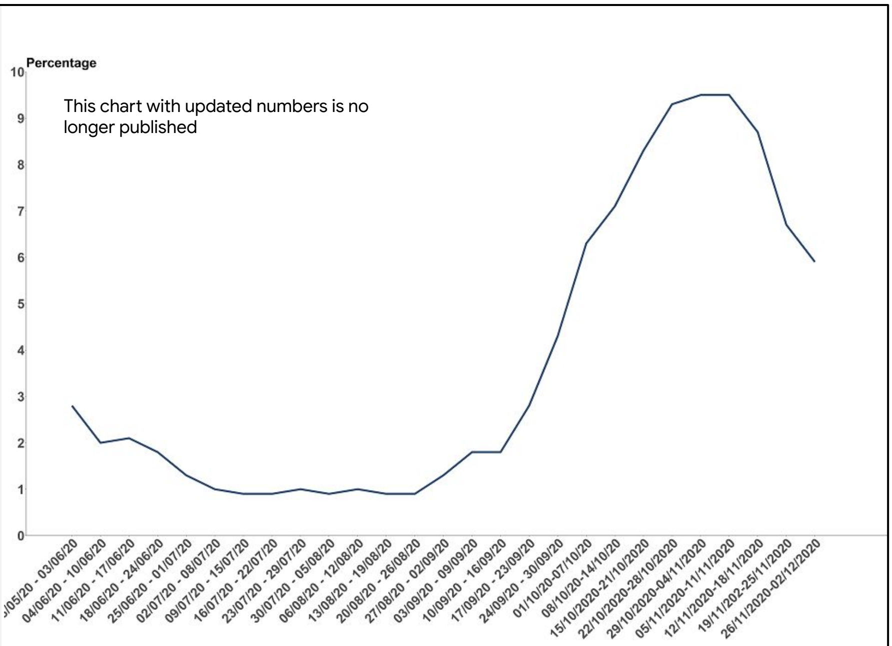
- Cost
- Accuracy
- Timeliness

Administrative Data

Percentage of people testing positive at least once for COVID-19 in each reporting week, England

Source:
National Health Service (NHS) Test and Trace December 10 report, Figure 4.

Link to methodology doc



This chart with updated numbers is no longer published

The chart shows the percent of people tested positive among all people tested that week from beginning of July 2020 to beginning of December 2020

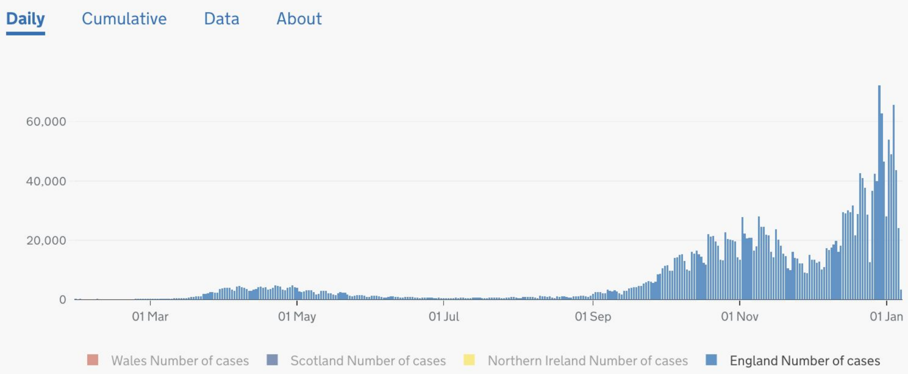
Applying the different frameworks to the Test and Trace data: strengths

Cost

Not super expensive

Timeliness

Data released on a daily basis



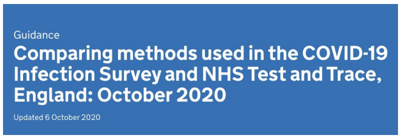
Screenshot as January 9, 2021

Number of people with at least one positive COVID-19 test result (either lab-reported or lateral flow device), by specimen date. Individuals tested positive more than once are only counted once, on the date of their first positive test. Data for the period ending 5 days before the date when the website was last updated with data for the selected area, highlighted in grey, is incomplete.
https://coronavirus.data.gov.uk/details/cases

Applying the different frameworks to the Test and Trace data: challenges

Coverage/sampling error

“Only testing people with COVID-19 symptoms underestimates infection rates in a population because many infected people show no or mild symptoms” (Ott, 2020)



Guidance

Comparing methods used in the COVID-19 Infection Survey and NHS Test and Trace, England: October 2020

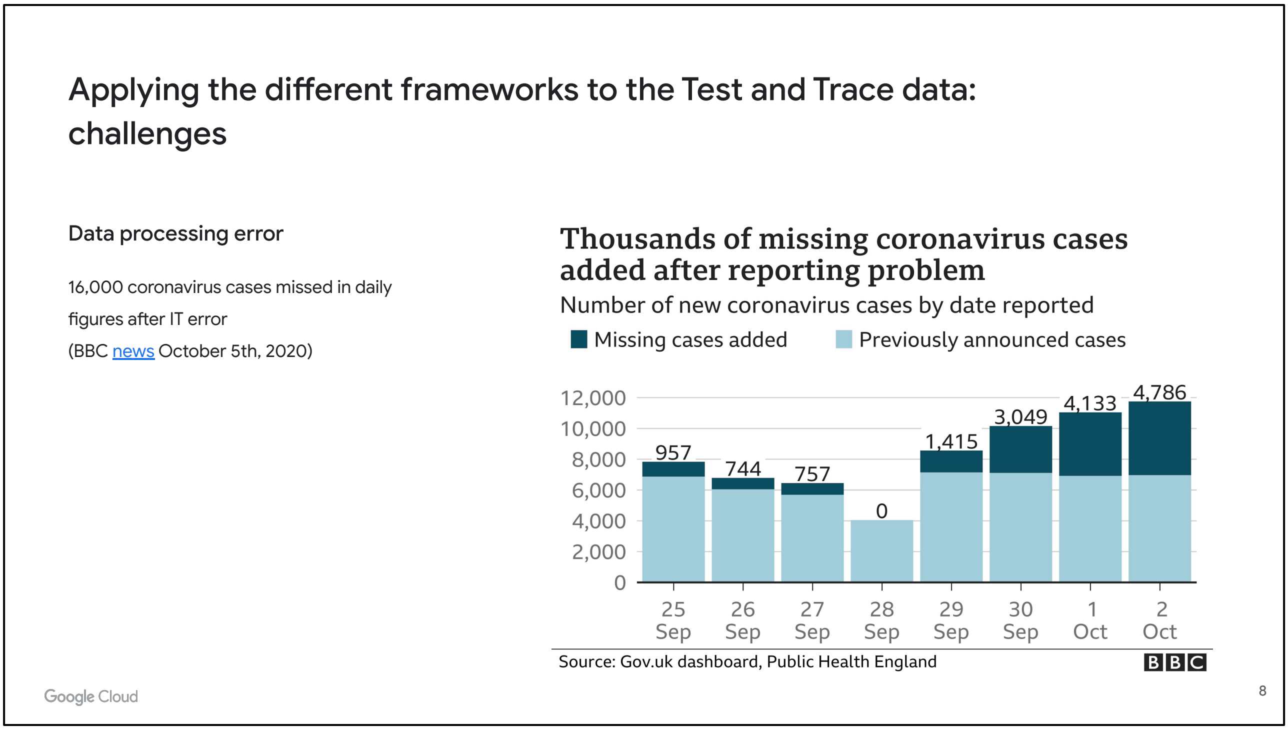
Updated 6 October 2020

Non probability sample (Department of Health and Social Care, 2020)

“The COVID-19 Infection Survey is based on a nationally representative survey sample [...] People are tested through NHS Test and Trace based on whether they are experiencing symptoms or if they are in a higher risk occupation or area. As not all populations have the same risk of infection, the population of those tested will not be nationally representative.”

Smith et al, 2020 study:
“Men and younger people were less likely to adhere to steps along the test, trace and isolate pathway”
“Key workers and people from minority ethnic backgrounds were less likely to identify common symptoms of COVID-19”

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/928684/S0732_CORSAIR_-_Adherence_to_the_test__trace_and_isolate_system.pdf



Applying the different frameworks to the Test and Trace data: challenges

Data processing error

16,000 coronavirus cases missed in daily figures after IT error

(BBC [news](https://www.bbc.co.uk/news/uk-54412581) October 5th, 2020)

Thousands of missing coronavirus cases added after reporting problem

Number of new coronavirus cases by date reported

Source: Gov.uk dashboard, Public Health England

BBC

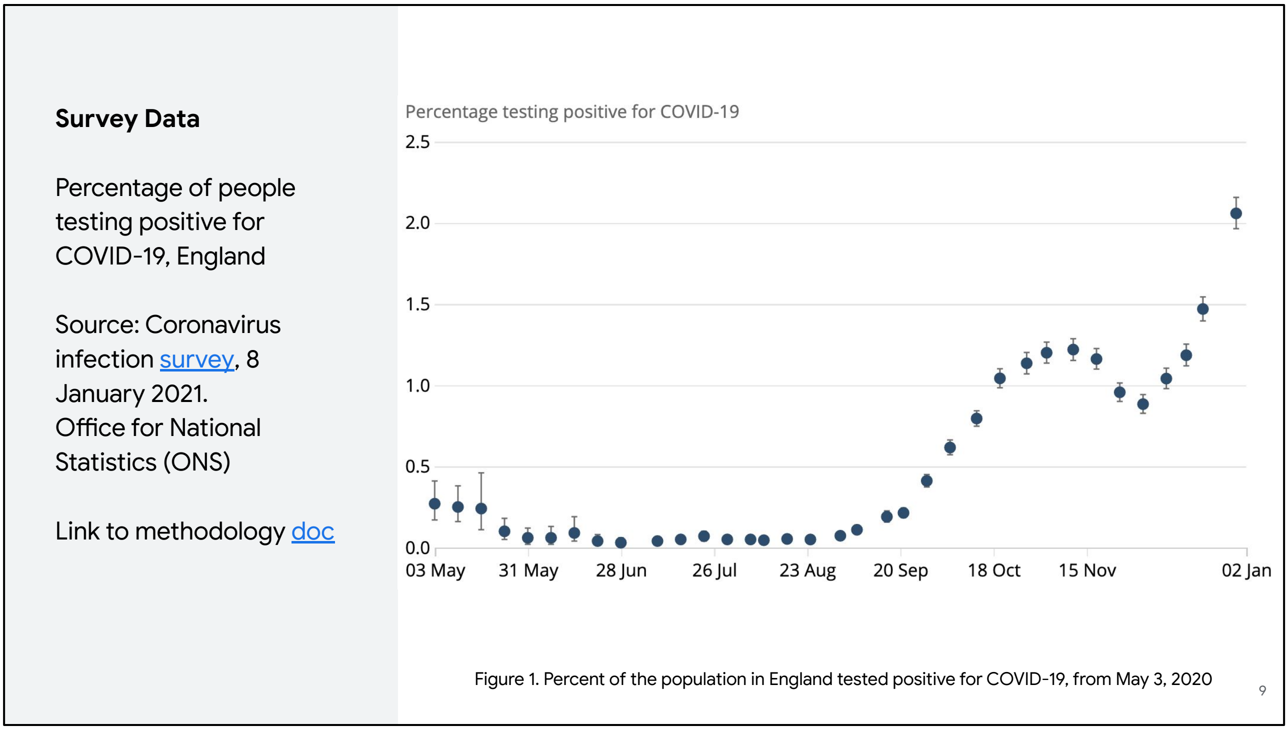
Google Cloud

8

https://www.bbc.co.uk/news/uk-54412581
https://theconversation.com/why-you-should-never-use-microsoft-excel-to-count-coronavirus-cases-147681

What happened:

- Companies that analysed the swab tests submitted their results as CSV files to Public Health England (PHE)
- PHE ingested these files into Excel XLS templates (Excel 2006 version, from 2007, Excel introduced the XLSX file format)
- The XLS file format has a limit of 65,536 rows and for this specific dataset of about 1,400 cases
- The extra cases were dropped from the template resulting in undercounting positive cases



Survey Data

Percentage of people testing positive for COVID-19, England

Source: Coronavirus infection survey, 8 January 2021. Office for National Statistics (ONS)

Link to methodology doc

Figure 1. Percent of the population in England tested positive for COVID-19, from May 3, 2020

Official estimates of the percentage of the population in England testing positive for the coronavirus (COVID-19) on nose and throat swabs from 3 May 2020

Applying the different frameworks to the ONS infection survey: strengths

Probability based sample

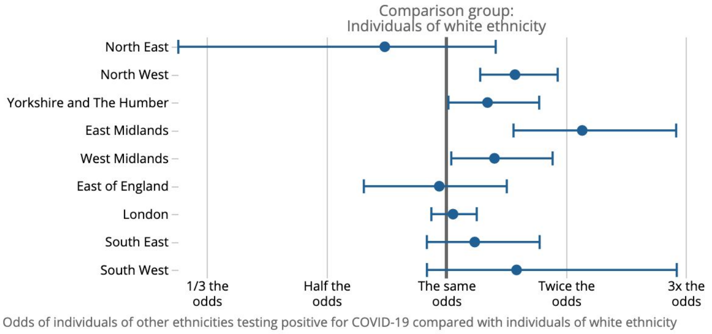
Non self selection of participants

Full demographics information on positive cases

Possibility to conduct further analysis as per chart below

Likelihood of testing positive varies by ethnicity and region, even when taking into account the sociodemographic factors

Source: Figure 8 from Coronavirus (COVID-19) Infection Survey: characteristics of people testing positive for COVID-19 in England and antibody data for the UK: December 2020



Applying the different frameworks to the ONS infection survey: challenges

Cost

This is a very expensive survey given its scope, sample size, and the cost of testing each respondents for COVID-19

Logistically challenging

The ONS started inviting 20K households who took part in a previous social survey

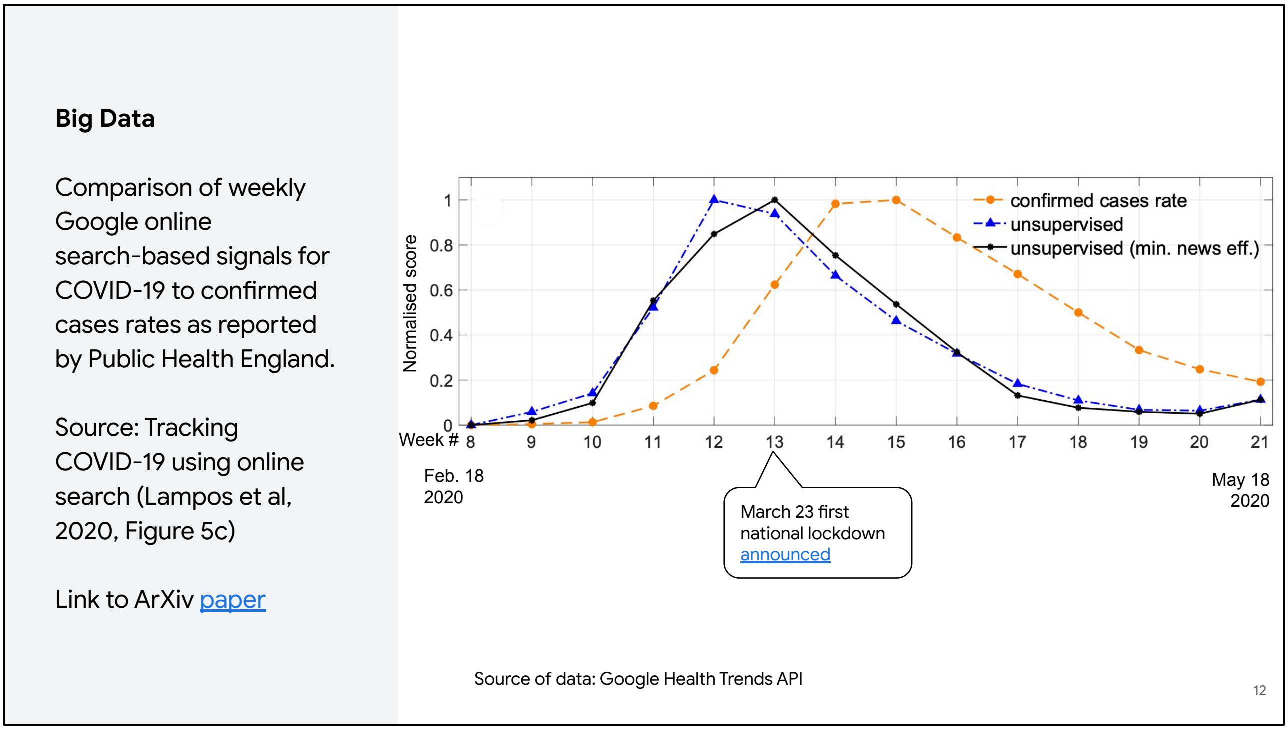
Then, an address based sample (ABS) have been used inviting 908K households since July 13

Timeliness

Data are released every 2 weeks

Potential nonresponse bias

Household response rates of the ABS is around 13%



Big Data

Comparison of weekly Google online search-based signals for COVID-19 to confirmed cases rates as reported by Public Health England.

Source: Tracking COVID-19 using online search (Lampos et al, 2020, Figure 5c)

Link to ArXiv paper

March 23 first national lockdown announced

Source of data: Google Health Trends API

https://fullfact.org/health/coronavirus-lockdown-hancock-claim/

Applying the different frameworks online search data: strengths

Timeliness

Nowcasting and futurecasting

Search data preceded confirmed cases by 1 to 2 weeks

Relatively cost effective

Once the search terms are tested and the model has been tested the tool can be used at a relatively low cost

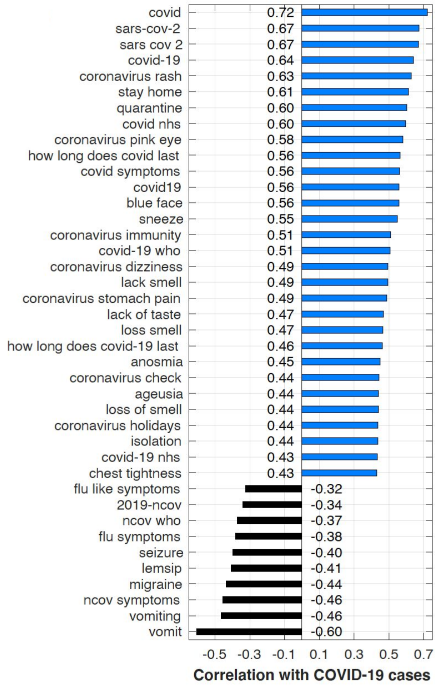
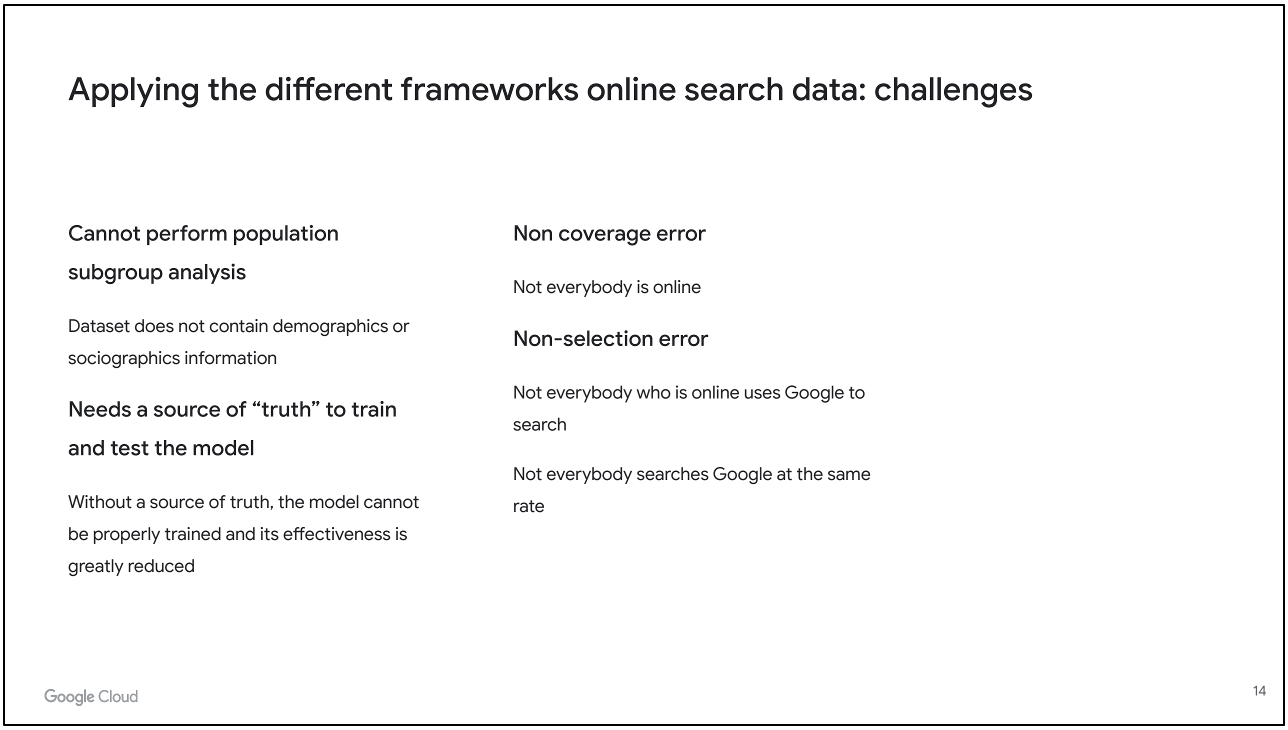


Figure 4a. Top-30 positively and top-10 negatively correlated search queries with confirmed COVID-19 cases in four English speaking countries (US, UK, Australia, and Canada)



Applying the different frameworks online search data: challenges

Cannot perform population subgroup analysis

Dataset does not contain demographics or sociographics information

Needs a source of “truth” to train and test the model

Without a source of truth, the model cannot be properly trained and its effectiveness is greatly reduced

Non coverage error

Not everybody is online

Non-selection error

Not everybody who is online uses Google to search

Not everybody searches Google at the same rate

In UK Google trends provides data by the 4 regions and by the 16 major cities

There are other analytical challenges involving search trend data. In the consumer attitude research (e.g., brand sentiment research), for example, we know that simply using the time series of a specific search term about a product may not be a good predictor for changes in brand sentiment. Instead, cutting the sample underlying the time series by different filters (e.g., web search vs. YT search, general search vs. shopping search) and "normalizing" the time series against different "benchmarks" (e.g., against a more popular competitive product, against a generally popular query, against a seasonal popular query). Similar approaches may be worthwhile to explore for pandemic forecasts.

What method is the best to measure incidence rates of Covid 19?
Fit for purpose and Rich Data are the silver lining

Start with the general research question

What sampling decisions do you make that can affect your ability to generalize?
(Skip Lupia: “Looking to the Future” Jan 14, 2020)

What level of measurement precision do you need?

Do you need subgroup and geographics analysis?

How timely do you need the data to be?

What is the budget for the study?

Rich Data

“The inclusion of multiple complementary indicators that enable accurate and efficient quantification of the target constructs and their relationships”
(Callegaro & Yang, 2018, p.186)

References I

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a Big Data world: Adapting the TSE framework to Big Data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Biemer, P. P., & Amaya, A. (2021). Total error framework for found data. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big Data meets survey science. A collection of innovative methods* (pp. 133–161). Wiley.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Callegaro, M., & Yang, Y. (2018). The role of surveys in the era of “Big Data.” In D. L. Vannette & J. A. Krosnick (Eds.), *Palgrave Handbook of survey research* (pp. 175–192). Palgrave. https://link.springer.com/chapter/10.1007/978-3-319-54395-6_23

Covid: 16,000 coronavirus cases missed in daily figures after IT error. (2020, October 5). *BBC News*. https://www.bbc.com/news/uk-54412581

Department of Health and Social Care. (2020, October 6). *Guidance. Comparing methods used in the COVID-19 infection survey and NHS test and trace, England: October 2020. Updated 6 October 2020*. Link to webpage

Georgiou, A. V. (2020). Letter to the Editors. Defeating the pandemic requires high quality and ethical official statistics. *Journal of Official Statistics*, *36*(4), 729–736. https://doi.org/10.2478/jos-2020-0036

Greater London Authority. (2020, October 23). *Coronavirus (COVID-19) Mobility Report*. https://data.gov.uk/dataset/4a475119-698e-49f3-afcc-8739558e1696/coronavirus-covid-19-mobility-report

References II

Lampos, V., Majumder, M. S., Yom-Tov, E., Edelstein, M., Moura, S., Hamada, Y., Rangaka, M. X., McKendry, R. A., & Cox, I. J. (2020). Tracking COVID-19 using online search. *ArXiv:2003.08086 [Cs]*. http://arxiv.org/abs/2003.08086

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, *368*(6490), 489–493. https://doi.org/10.1126/science.abb3221

Office for National Statistics. *COVID-19 in England and antibody data for the UK - Office for National Statistics*. Retrieved December 28, 2020. Link to webpage

Office for National Statistics. *Coronavirus (COVID-19) Infection Survey, UK - Office for National Statistics*. Retrieved January 9, 2021. Link to webpage

Office for National Statistics. *COVID-19 Infection Survey (Pilot): Methods and further information*. Retrieved December 28, 2020. Link to webpage

Ott, A. (2020). *Monitoring wastewater for COVID-19*. https://post.parliament.uk/monitoring-wastewater-for-covid-19/

Smith, L. E., Potts, H. W., Amlôt, R., Fear, N. T., Michie, S., & Rubin, J. G. (2020). *Adherence to the test, trace and isolate system: Results from a time series of 21 nationally representative surveys in the UK (the COVID-19 Rapid Survey of Adherence to Interventions and Responses [CORSAIR] study)*. Link to pdf

Stephens-Davidowitz, S. (2020, April 5). Opinion. Google searches can help us find emerging Covid-19 outbreaks. *The New York Times*. https://www.nytimes.com/2020/04/05/opinion/coronavirus-google-searches.html

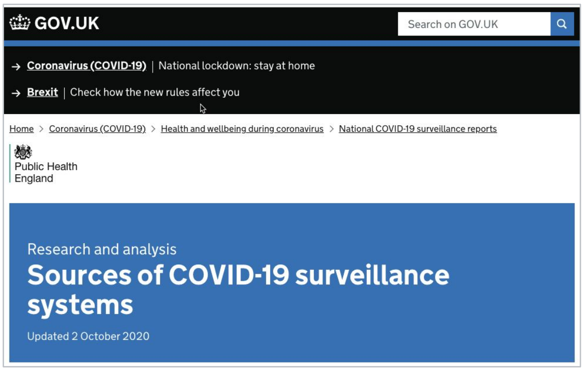
Weekly statistics for NHS Test and Trace (England) and coronavirus testing (UK): 26 November to 2 December. GOV.UK. Retrieved December 28, 2020. Link to webpage

Appendix

Other ways to measure Covid infection rates & Covid related topics
Slides from the January 5, 2021 Government press conference

Other ways to measure Covid infection rates & Covid related topics

- Wastewater monitoring program (Ott, 2020)
- Mobility data to measure the effects of movement restrictions (Greater London Authority, 2020)
- Survey data to estimate compliance with preventive rules such as washing hands, wearing masks… (Office for National Statistics 2020)
- Social network analysis



→ Coronavirus (COVID-19) | National lockdown: stay at home

→ Brexit | Check how the new rules affect you

Home > Coronavirus (COVID-19) > Health and wellbeing during coronavirus > National COVID-19 surveillance reports

Public Health England

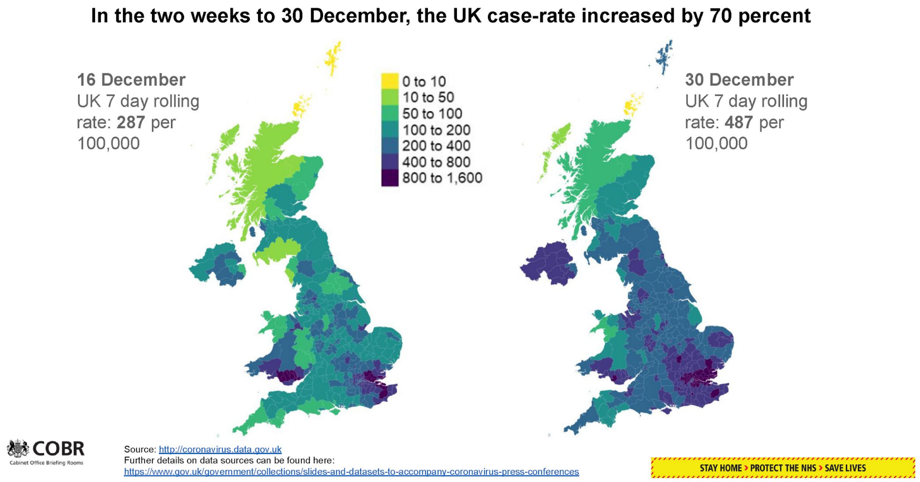
Research and analysis

Sources of COVID-19 surveillance systems

Updated 2 October 2020

What sources does the UK Government use to make decisions?

Administrative Data. January 5 press conference



In the two weeks to 30 December, the UK case-rate increased by 70 percent

16 December
UK 7 day rolling
rate: **287** per
100,000

30 December
UK 7 day rolling
rate: **487** per
100,000

- 0 to 10
- 10 to 50
- 50 to 100
- 100 to 200
- 200 to 400
- 400 to 800
- 800 to 1,600

COBR
Cabinet Office Briefing Rooms

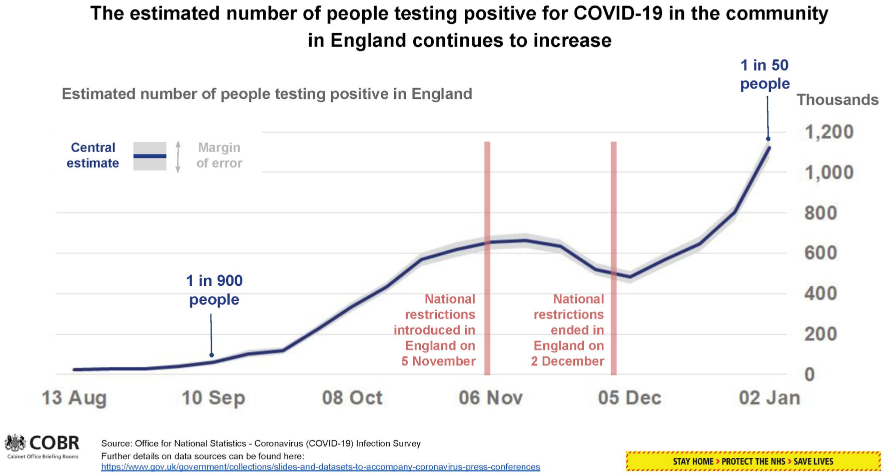
Source: http://coronavirus.data.gov.uk
Further details on data sources can be found here:
https://www.gov.uk/government/collections/slides-and-datasets-to-accompany-coronavirus-press-conferences

STAY HOME > PROTECT THE NHS > SAVE LIVES

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949818/2021-01-5_COVID-19_Press_Conference_Slides__for_publication_.pptx.pdf

What sources does the UK Government use to make decisions?

Survey Data. January 5 press conference



The estimated number of people testing positive for COVID-19 in the community in England continues to increase

COBR
Cabinet Office Briefing Rooms

Source: Office for National Statistics - Coronavirus (COVID-19) Infection Survey
Further details on data sources can be found here:
https://www.gov.uk/government/collections/slides-and-datasets-to-accompany-coronavirus-press-conferences

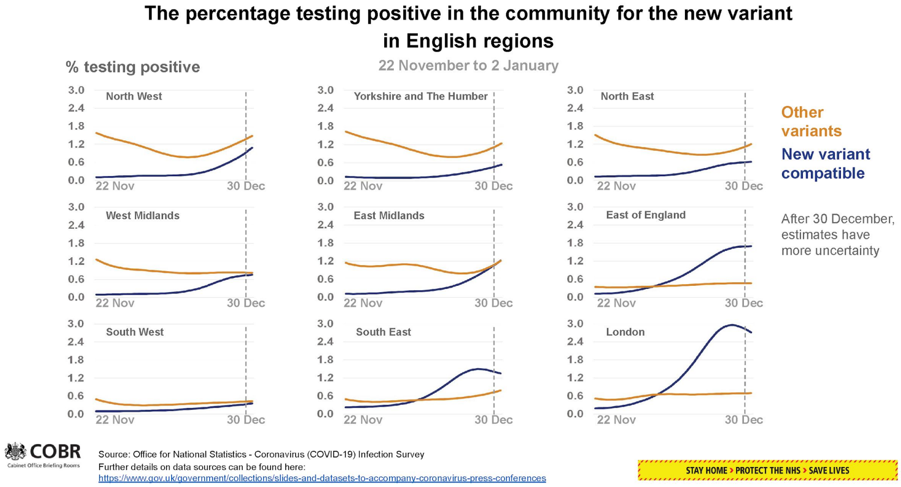
STAY HOME > PROTECT THE NHS > SAVE LIVES

Google Cloud

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949818/2021-01-5_COVID-19_Press_Conference_Slides__for_publication_.pptx.pdf

What sources does the UK Government use to make decisions?

Survey Data. January 5 press conference



The percentage testing positive in the community for the new variant in English regions

22 November to 2 January

After 30 December, estimates have more uncertainty

Source: Office for National Statistics - Coronavirus (COVID-19) Infection Survey
Further details on data sources can be found here:
https://www.gov.uk/government/collections/slides-and-datasets-to-accompany-coronavirus-press-conferences

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949818/2021-01-5_COVID-19_Press_Conference_Slides__for_publication_.pptx.pdf