Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts

Authors: Sam Abbott *, Joel Hellewell *, Robin N Thompson, Katharine Sherratt, Hamish P Gibbs, Nikos I Bosse, James D Munday, Sophie Meakin, Emma L Doughty, June Young Chun, Yung-Wai Desmond Chan, Flavio Finger, Paul Campbell, Akira Endo, Carl A B Pearson, Amy Gimma, Tim Russell, CMMID COVID modelling group, Stefan Flasche, Adam J Kucharski, Rosalind M Eggo, Sebastian Funk

* contributed equally

Abstract

Background: Assessing temporal variations in transmission in different countries is essential for monitoring the epidemic, evaluating the effectiveness of public health interventions and estimating the impact of changes in policy.

Methods: We use case and death notification data to generate daily estimates of the time-dependent reproduction number globally, regionally, nationally, and subnationally over a 12 week rolling window. Our modelling framework, based on open source tooling, accounts for uncertain reporting delays, so that the reproduction number is estimated based on underlying latent infections and not reported cases or deaths.

Conclusions: This decision-support tool can be used to assess changes in virus transmission both globally, regionally, nationally, and subnationally. This allows public health officials and policymakers to track the progress of the outbreak in near real-time using an epidemioligcally valid measure. As well as providing regular updates on our website, we also provide an open source tool-set so that our approach can be used directly by researchers and policymakers on confidential data-sets. We hope that our tool will be used to support decisions in countries worldwide throughout the ongoing COVID-19 pandemic.

Keywords: Covid-19, SARS-CoV-2, surveillance, forecasting, time-varying reproduction number

Introduction

The coronavirus disease 2019 (COVID-19) pandemic that emerged in December 2019 has since spread to over 100 countries in every continent except Antarctica. While some information on the progress of an outbreak in a given country can be gained from the reported numbers of confirmed cases and deaths, these numbers can obscure changes in the underlying dynamics of the outbreak due to delays between infection and the eventual reporting of a case or death. Accounting for the uncertain delays from infection to symptom onset, and the uncertain delays from symptom onset to hospital admission, diagnostic testing or potential death, followed by further delays until data are recorded in official statistics, requires the use of specific statistical methods for handling right-truncated data [1–3],uncertainty, and the creation of a "nowcast" [4,5] (an estimate of the current number of newly infected or symptomatic cases).

A method for tracking the progress of an outbreak is to measure changes in the time-varying reproduction number (effective reproduction number), which represents the average number of secondary infections generated by each new infectious case [6–8]. This approach can be advantageous compared to monitoring numbers of newly reported or symptomatic cases since, in principle, reproduction number estimates reflect variations in transmission intensity. Due to the delays in disease progression, recorded numbers of newly notified or symptomatic cases will increase or decrease for a period after transmissibility has reduced or

increased, respectively. Monitoring changes in the time-varying reproduction can account for this delay and reveals variations in transmissibility that are not clear when using only reported cases.

This paper outlines the methods used to produce the website we have developed (https://epiforecasts.io/covid/) that presents real-time estimates and forecasts of reported cases by date of infection and the respective time-varying reproduction numbers globally, regionally, nationally and subnationally for Covid-19. This website relies on methods implemented in the `EpiNow2` R package and data aggregated in the `covidregionaldata` R package, both developed by the authors [9,10]. Our estimates overcome some of the limitations of naive implementations that derive estimates for the reproduction number directly from numbers of reported cases without adjusting (or with only partial adjustments) for the delay from infection to symptom onset or from onset to notification. Our approach also incorporates multiple sources of uncertainty that if excluded can bias estimates. The code that creates and updates the website is open source, and documented for use by others, allowing policymakers and researchers to run analyses using confidential data. The methods outlined in this paper and corresponding code base are under development, and new versions of this live article will be released alongside changes to the methods to create a record of the methodology used throughout the pandemic.

Methods

Data

We use daily counts of confirmed cases and deaths reported by the European Centre for Disease Control from the last 12 weeks for all analyses conducted at the national level [10,11]. To estimate the delay from symptom onset to reporting (once confirmed with a positive laboratory test), we use all cases from a publicly available linelist for which onset and notification dates are available [10,12]. This linelist combines all known linelist data from over 100 countries and at the time of writing. Countries are only included in the reported estimates if within the last 12 weeks they have fewer than 14 days with non-zero case counts. This restriction reduces the likelihood of spurious estimates for countries with limited transmission or case ascertainment.

For sub-national analyses, the data is aggregated using the `covidregionaldata` R package developed by the authors. Individual data sources are reported on the respective pages of our website. The data are fetched from government departments or from individuals who maintain a data source if no official data are available. Similarly to national estimates, subnational areas are only included if they report at least 14 days with non-zero cases in the last 12 weeks.

All analyses described below are run independently for each national or subnational entity under consideration. An automated timestamp is used to evaluate if data has been updated since the last time estimates were made in order to avoid repeatedly estimating based on the same data.

Delays between case onset and report

To estimate the reporting delay (i.e the delay between onset and case report or death) with appropriate uncertainty, we fit a log-normal distribution, using use the statistical modelling program stan [9,13], to 100 subsampled bootstraps (each with 250 samples drawn with replacement) of the available delay data. Accounting for left and right censoring occurring in the data as each date is rounded to the nearest day and truncated to the maximum observed delay. There was insufficient data available on the various reporting delays to estimate spatially- or temporally-varying delays whilst also accounting for the biases induced by the growth rate of reported cases, so they were considered to be static over the course of the epidemic.

This results in an onset to case report delay distribution with a mean of 2.6 days (standard deviation (SD): 1.2 days) and a standard deviation of 4.4 days (SD: 1.1 days) and an onset to death report delay distribution with a mean of 9.8 days (SD: 1.1 days) and a standard deviation of 2.1 days (SD: 1.1 days). For computational reasons the maximum allowed delay is set to be 30 days with truncation handled appropriately by `EpiNow2` [9].

As data may also be right truncated due to unrecorded delays (i.e the delay between a case report and its

appearance in an aggregated data set) we truncate all time-series to exclude the last 3 days of data, based on qualitative inspection of the stability of case counts in the data-sets used.

Estimating the time-varying reproduction number and nowcasting reported infections

We estimated the instantaneous reproduction number (R_t) using the `EpiNow2` R package (version 1.2.0) [9] on the last 12 weeks of available data, discarding estimates from the first 14 days globally, for United Nation regions, nationally, and subnationally for 10 countries. The instantaneous reproduction number represents the number of secondary cases arising from an individual showing symptoms at a particular time, assuming that conditions remain identical after that time, and is therefore a measure of the instantaneous transmissibility (in contrast to the case reproduction number - see Fraser (2007) [8] for a full discussion). `EpiNow2` implements a Bayesian latent variable approach using the probabilistic programming language Stan [13], which works as follows. The initial number of infections were estimated as a free parameter with a prior based on the initial number of cases, or deaths, respectively. For each subsequent time step, previous imputed infections were summed, weighted by an uncertain generation time probability mass function, and combined with an estimate of R_t to give the prevalence at time t [6,7,9]. We used a gamma prior for the reproduction number with mean 1 and standard deviation 1 reflecting our current belief that R_t is likely to be centered around 1 at this stage in the epidemic. This contrasts with our earlier approach which was to use a prior with a of mean 2.6 and standard deviation 2. This was based on early estimates for the basic reproduction number (R_0) from the initial stages of the outbreak in Wuhan [14,15] with long tails to allow for differences in the reproduction number between countries. The infection trajectories were then mapped to reported case counts by convolving over an uncertain incubation period and report delay distribution, and a negative binomial observation model combined with a multiplicative day of the week effect (with an independent effect for each day of the week) [16]. Temporal variation was controlled using an approximate Gaussian process [17] with a squared exponential kernel, with the length scale and magnitude estimated during the model fitting process. Each timeseries was fitted independently using Markov-chain Monte Carlo (MCMC). A minimum of 2 chains were used with a warmup of 500 each and 2000 samples post warmup. Convergence was assessed using the R hat diagnostic [13].

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) for all estimates. Sourced from [18] but refit 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) [19] rather than the incubation period used in the original study (code available here: https://github.com/seabbs/COVID19). This incubation period was also used to convolve from unobserved infections to unobserved onsets in the model. See [9] for further details on the approach.

Estimating the daily growth rate and doubling time

We estimated the rate of spread (r) by converting our R_t estimates using an approximation derived in [20]. The doubling time was then estimated by calculating $\ln(2)\frac{1}{r}$ for each estimate of the rate of spread.

Estimated change in daily cases

We defined the estimated change in daily cases to correspond to the proportion of reproduction number estimates for the current day that are below 1 (the value at which an outbreak is in decline). It was assumed that if less than 5% of samples were subcritical then an increase in cases was definite, if less than 20% of samples were subcritical then an increase in cases was likely, if more than 80% of samples were subcritical then a decrease in cases was likely and if more than 95% of samples were subcritical then a decrease in cases was definite. For countries/regions with between 20% and 80% of samples being subcritical we could not make a statement about the likely change in cases (defined as unsure).

The effect of changes in testing procedure

The results presented here are sensitive to changes in COVID-19 testing practices and the level of effort put into detecting COVID-19 cases, e.g. through contact tracing. For example, if numbers of incident infections remain constant but a country begins to find and report a higher proportion of cases, then an increasing value of the reproduction number will be inferred. This is because all changes in the number of cases are attributed to changes in the number of infections resulting from previously reported cases, and are not assumed to be a result of improved testing and surveillance. On the other hand, if a country reports a lower proportion of cases because a lower number of tests are performed (which can happen if reagents required for testing are no longer available, for example) or the surveillance system captures a lower proportion of infections, then the model will attribute this to a drop in the reproduction number that may not be a true reduction. In order for our estimates to be unbiased not all cases have to be reported, but the level of testing effort (and therefore the proportion of detected cases) must be constant [21]. This means that, whilst a change in testing effort will initially introduce bias, this will be reduced over time as long as the testing effort remains consistent from this point onwards.

Countries may also change the focus of their surveillance over the course of the outbreak. They may initially focus on identifying travellers returning from areas of known COVID-19 transmission and performing contact tracing on the contacts of known cases. As the outbreak evolves this may change to passive surveillance at hospitals. Here, the case definition may also change from tests based on polymerase chain reaction (PCR) to diagnoses based on symptoms and computed tomography (CT) scans. In the future, different kinds of COVID-19 tests may be deployed that could influence results, such as tests that detect both active and past infections.

Forecasting the reproduction number and case counts by date of infection

We forecast the time-varying effective reproduction number over a 14 day time horizon by assuming it remains the same as the last estimated R_t. The reproduction number forecast is then transformed into a case forecast using the `EpiNow2` model outlined in the previous section [9]. These forecasts are indicative only and should not be considered with a weight equal to the real-time estimates. Changes in contact rates, mobility, and public health interventions are not accounted for which may lead to significant inaccuracy.

Reporting

We report the median and 90% highest density credible intervals for all measures with 20%, 50% and 90% high density regions shown in figures. The analysis was conducted independently for all regions and is updated regularly as new data becomes available. To highlight the proportion of cases that have yet to be reported (due to correcting for right truncation), we show a cut-off in figures based on the mean of all delays. Values prior to this point are defined as estimates with values past this point being defined as estimates based on partial data. In reality, this is a continuum with estimates closer to now progressively being based on less data and therefore becoming increasing uncertain. All estimates are available as downloadable csvs under an open-source license for use elsewhere (https://github.com/epiforecasts/covid-rt-estimates/).

Website, summarised estimates, and interactivity

We use open-source Rmarkdown templates and the distill framework to generate webpages summarising these estimates [22,23]. The `RtD3` package is used to provide interactive visualisations of all estimates [24]. Estimates by country are provided on a dedicated static page along with global, and regional, summaries. More detailed subnational estimates are available for over 10 countries in an flexible framework into which additional subnational estimates will be added as more data becomes available.

Discussion

We provide a centralised resource, which generates comparable daily estimates of the time-varying reproduction number and a daily nowcast of the number of cases newly infected derived using a standardised

method. We account for the delay between infection and case notification and include all sources of quantifiable uncertainty. This resource may be useful for policymakers to track the progression of the COVID-19 outbreak and evaluate the effectiveness of intervention measures. As new data become available, we will include sub-national estimates for additional countries, and provide additional support for public health agencies or researchers interested in applying our methods to their data.

There are several advantages associated with our approach. Firstly, reported case counts are the only data required, which allows our approach to be used in a wide variety of contexts. Secondly, we apply the same methodology to all countries. This means that estimates can be compared without having to consider differences in the underlying methodology (even if differences in testing should still be accounted for as discussed below). Finally, we have constructed our approach using open source tools and all of our code, raw data, and results are available online and developed with other users in mind. This means our approach can be applied by others to non-public data and be fully evaluated by end users.

Our approach is also subject to several limitations. Firstly, the model requires that the proportion of infections that are notified is constant. In other words, it requires consistency in the focus of the surveillance method, level of effort spent on testing, and case definition. Yet it is often the case that the level of under-reporting in a country changes over the course of an outbreak [21]. However, it should be noted that any changes in surveillance testing procedures will only bias the estimates temporarily if they begin to remain consistent again after they have changed. How long the bias remains in the reproduction number estimates will depend on the serial generation time and delay distributions, as well as the lengthscale of the gaussian process used in the reproduction number estimation process. The impact of testing and other reporting biases vary between measures of transmission (test positive cases, hospital admissions, test positive deaths). For this reason we include estimates based on reported deaths and provide tooling to allow estimates to be produced for alternative datasets. In theory, estimates from disparate sources should be comparable using our approach, however if they in fact represent different sub-populations then there may be variation between them that can potentially be usefully interpreted.

In addition, the model is limited by how representative the delay that we use from infection to notification distribution is for a given location. As there is limited data to assess this, we estimate a bootstrapped global delay distribution using the combined data from every country. In particular, the delay from onset to notification can especially impact the upscaling of cases by date of onset that accounts for cases that have onset but not yet been reported. If the true delay from onset to notification for a given country is shorter than our global delay, then we will overestimate onset case numbers, and vice versa for true delays longer than the distribution we used. Additionally, estimates of the reporting delay distribution are known to be biased early in an epidemic and may vary over time [25]. However, our use of a bootstrapped subsampling approach mitigates these issues by allowing multiple delay distributions based on the observed data to be considered at the cost of increasing uncertainty in our estimates.

Our model is also limited by the data avaliable to us. For example, the publically available linelists contain little data on the importation status of cases. This means that cases counts may be biased upwards by attributing imported cases to local transmission. This bias is particularly problematic when case counts are low. Unfortunately, in the absence of data, this issue can only be explored via scenario analysis.

As more data becomes available, future work should look to refine the distributions used for generation time, incubation period, and the report delay. There is also the potential to extend the present model to account for chainges in the delay from onset to notification over the course of an outbreak though additional data would need to be available for this to be possible. Finally, there is scope to explore how outbreak dynamics that differ among particular sub-populations, such as high-risk COVID-19 patients, can bias overall reproduction number estimates. This may be achieved by comparing reproduction number estimates from disparate data sources such as test positive cases, hospital admissions, and test positive deaths.

Our approach, providing real-time estimates of the reproduction number, serves as a valuable tool for decision makers looking to track the course of COVID-19 outbreaks. The nowcasts explicitly account for delays, using the same methodology across all countries and sub-national regions. These reproduction number estimates may also be used to ascertain the likely outbreak trajectory if no policy interventions are made. They can also provide real-time feedback on whether transmission is decreasing following a particular intervention, or

whether it is increasing following the relaxing or lifting of current intervention measures. We hope that our website and the related toolkit will provide a valuable resource for devising strategies to contain COVID-19 outbreaks worldwide.

Data availability

Latest data: https://github.com/epiforecasts/covid-rt-estimates

Archived data at the time of publication: https://doi.org/10.5281/zenodo.3841818

License: MIT

Software availability

Development

- Website (Front-end): https://github.com/epiforecasts/covid
- *EpiNow2* R package (R estimation, data processing, visualisation and reporting): https://github.com/epiforecasts/EpiNow2
- *covidregionaldata* R package (data aggregation and processing): https://github.com/epiforecasts/covidregionaldata
- *RtD3* R package (interative visualisation): https://github.com/epiforecasts/RtD3

Archived at the time of publication

- Website: https://doi.org/10.5281/zenodo.3841818
- *EpiNow2* R package [9]: https://doi.org/10.5281/zenodo.3957489
- *covidregionaldata* R package [10]: https://doi.org/10.5281/zenodo.3957539
- *RtD3* [24]: https://doi.org/10.5281/zenodo.4011841

License: MIT

Acknowledgements

This project was enabled through access to the **MRC eMedLab Medical Bioinformatics infrastructure**, supported by the **Medical Research Council** (MR/L016311/1). Additional compute infrastructure and support was provided by the **Met office**. We thank Venexia Walker for comments on a version of this draft. The following authors were part of the Centre for Mathematical Modelling of Infectious Disease 2019-nCoV working group. Each contributed in processing, cleaning and interpretation of data, interpreted findings, contributed to the manuscript, and approved the work for publication: Samuel Clifford, Mark Jit, Stéphane Hué, Eleanor M Rees, Petra Klepac, Damien C Tully, Rachel Lowe, Kathleen O'Reilly, Nicholas G. Davies, Quentin J Leclerc, Arminder K Deol, Gwenan M Knight, C Julian Villabona-Arenas, Fiona Yueqian Sun, Emily S Nightingale, Alicia Rosello, Adam J Kucharski, Yang Liu, Billy J Quilty, Matthew Quaife, Jon C Emery, Katherine E. Atkins, Simon R Procter, W John Edmunds, Megan Auzenbergs, Christopher I Jarvis, David Simons, Kiesha Prem, Graham Medley, Thibaut Jombart, Charlie Diamond, Anna M Foss, Rein M G J Houben, Kevin van Zandvoort, Georgia R Gore-Langton.

Funding

The following funding sources are acknowledged as providing funding for the named authors. Alan Turing Institute (AE). This research was partly funded by the Bill & Melinda Gates Foundation (NTD Modelling Consortium OPP1184344: CABP). DFID/Wellcome Trust (Epidemic Preparedness Coronavirus research programme 221303/Z/20/Z: CABP). This research was partly funded by the Global Challenges Research Fund (GCRF) project 'RECAP' managed through RCUK and ESRC (ES/P010873/1: AG). HDR UK (MR/S003975/1: RME). Nakajima Foundation (AE). UK DHSC/UK Aid/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 (ITCRZ 03010: HPG). UK MRC (MC_PC 19065: RME). Wellcome Trust (206250/Z/17/Z: TWR; 208812/Z/17/Z: SFlasche; 210758/Z/18/Z: JDM, JH, NIB, SA, SFunk, SRM).

References

1 Linton NM, Kobayashi T, Yang Y *et al.* Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of clinical medicine* 2020;**9**.

2 Cori A, Donnelly CA, Dorigatti I *et al.* Key data for outbreak evaluation: Building on the ebola experience. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2017;**372**.

3 Mizumoto K, Kagaya K, Zarebski A *et al.* Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2020;**25**.

4 Donker T, Boven M van, Ballegooijen WM van *et al.* Nowcasting pandemic influenza a/h1n1 2009 hospitalizations in the netherlands. *European journal of epidemiology* 2011;**26**:195–201.

5 Kassteele J van de, Eilers PHC, Wallinga J. Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained p-spline smoothing. *Epidemiology (Cambridge, Mass)* 2019;**30**:737–45.

6 Cori A, Ferguson NM, Fraser C *et al.* A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* 2013;**178**:1505–12. doi:10.1093/aje/kwt133

7 Thompson RN, Stockwin JE, Gaalen RD van *et al.* Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics* 2019;**29**:100356. doi:https://doi.org/10.1016/j.epidem.2019.100356

8 Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS one* 2007;**2**:e758.

9 Abbott S, Hellewell J, Hickson J *et al.* EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters. - 2020;**-**:–. doi:10.5281/zenodo.3957489

10 Abbott S, Sherratt K, Bevan J *et al.* Covidregionaldata: Subnational data for the covid-19 outbreak. - 2020;**-**:–. doi:10.5281/zenodo.3957539

11 Disease Prevention EC for, Control. Download today's data on the geographic distribution of covid-19 cases worldwide. 2020. www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

12 Xu B, Gutierrez B, Hill S *et al.* Epidemiological data from the nCoV-2019 outbreak: Early descriptions from publicly available data. http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337

13 Stan Development Team. RStan: The r interface to stan. 2020.http://mc-stan.org/

14 Imai N, Cori A, Dorigatti I *et al.* Report 3: Transmissibility of 2019-nCoV. https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-2019-nCoV-transmissibility.pdf

15 Abbott S, Hellewell J, Munday J *et al.* The transmissibility of novel coronavirus in the early stages of the 2019-20 outbreak in wuhan: Exploring initial point-source exposure sizes and durations using scenario analysis. *Wellcome open research* 2020;**5**:17.

16 Endo A, Abbott S, Kucharski AJ *et al.* Estimating the overdispersion in covid-19 transmission using outbreak size outside china [version 1; peer review: 1 approved]. *Wellcome open research* 2020;**5**.

17 Riutort-Mayol G, Bürkner P-C, Andersen MR *et al.* Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. 2020.http://arxiv.org/abs/2004.11408

18 Ganyani T, Kremer C, Chen D *et al.* Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance* 2020;**25**.

19 Lauer SA, Grantz KH, Bi Q *et al.* The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine* 2020;**172**:577–82.

20 Park SW, Champredon D, Weitz JS *et al.* A practical generation-interval-based approach to inferring the strength of epidemics from their speed. *Epidemics* 2019;**27**:12–8. doi:https://doi.org/10.1016/j.epidem.2018.12.002

21 Russell TW. Using a delay-adjusted case fatality ratio to estimate under-reporting. https://cmmid.github.io/topics/covid19/severity/global_cfr_estimates.html

22 Xie Y, Allaire JJ, Grolemund G. *R markdown: The definitive guide*. Boca Raton, Florida:: Chapman; Hall/CRC 2018. https://bookdown.org/yihui/rmarkdown

23 Allaire J, Iannone R, Xie Y. *Distill: R markdown format for scientific and technical writing*. 2020. https://github.com/rstudio/distill

24 Gibbs H, Abbott S, Funk S. RtD3: Rt visualization in d3. *Zenodo* 2020;**-**:–. doi:10.5281/zenodo.4011841

25 Britton T, Scalia Tomba G. Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society, Interface* 2019;**16**.