Check for updates

METHOD ARTICLE

Inferring the number of COVID-19 cases from recently reported deaths [version 1; peer review: 2 approved]

Thibaut Jombart [1-3], Kevin van Zandvoort[1], Timothy W. Russell [1],
Christopher I. Jarvis[1], Amy Gimma [1], Sam Abbott [1], Sam Clifford[1],
Sebastian Funk [1], Hamish Gibbs[1], Yang Liu [1], Carl A. B. Pearson [1,4],
Nikos I. Bosse [1],
Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group,
Rosalind M. Eggo[1], Adam J. Kucharski[1], W. John Edmunds[1]

[1]Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK
[2]UK Public Health Rapid Support Team, London, UK
[3]MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, UK
[4]South African Centre for Epidemiological Modelling and Analysis, Stellenbosch University, Stellenbosch, South Africa

Abstract

We estimate the number of COVID-19 cases from newly reported deaths in a population without previous reports. Our results suggest that by the time a single death occurs, hundreds to thousands of cases are likely to be present in that population. This suggests containment via contact tracing will be challenging at this point, and other response strategies should be considered. Our approach is implemented in a publicly available, user-friendly, online tool.

Keywords

epidemics, outbreak, modelling, statistics, estimation, covid-19, SARS-CoV-2
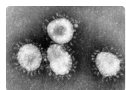
This article is included in the Coronavirus (COVID-19) collection.

Open Peer Review

Reviewer Status ✔ ✔

	Invited Reviewers	
	1	**2**
version 1 27 Apr 2020	✔ report	✔ report

1. **Benjamin M. Bolker** [iD], McMaster University, Hamilton, Canada

2. **Camille Pelat**, Santé publique France, Saint-Maurice, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Thibaut Jombart (thibautjombart@gmail.com)

Author roles: Jombart T: Conceptualization, Formal Analysis, Methodology, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **van Zandvoort K**: Data Curation, Formal Analysis, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Russell TW**: Data Curation, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Jarvis CI**: Data Curation, Formal Analysis, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Gimma A**: Data Curation, Formal Analysis, Software, Writing – Review & Editing; **Abbott S**: Data Curation, Formal Analysis, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Clifford S**: Data Curation, Validation, Writing – Review & Editing; **Funk S**: Conceptualization, Methodology, Supervision; **Gibbs H**: Data Curation, Formal Analysis, Investigation; **Liu Y**: Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **Pearson CAB**: Conceptualization, Formal Analysis, Methodology, Writing – Review & Editing; **Bosse NI**: Data Curation, Formal Analysis, Writing – Original Draft Preparation; **Eggo RM**: Conceptualization, Methodology, Writing – Review & Editing; **Kucharski AJ**: Conceptualization, Methodology, Writing – Review & Editing; **Edmunds WJ**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Jombart T, van Zandvoort K, Russell TW *et al.* **Inferring the number of COVID-19 cases from recently reported deaths [version 1; peer review: 2 approved]** Wellcome Open Research 2020, **5**:78 https://doi.org/10.12688/wellcomeopenres.15786.1

First published: 27 Apr 2020, **5**:78 https://doi.org/10.12688/wellcomeopenres.15786.1

Introduction

As the coronavirus-2019 (COVID-19,[1]) epidemic continues to spread worldwide, there is mounting pressure to assess the scale of epidemics in newly affected countries as rapidly as possible. We introduce a method for estimating cases from recently reported COVID-19 deaths. Results suggest that by the time the first deaths have been reported, there may be hundreds to thousands of cases in the affected population. We provide epidemic size estimates for several countries, and a user-friendly, web-based tool that implements our model[2].

Methods

Using deaths to infer cases

COVID-19 deaths start to be notified in countries where few or no cases had previously been reported[3]. Given the non-specific symptoms[4], and the high rate of mild disease[5], a COVID-19 epidemic may go unnoticed in a new location until the first severe cases or deaths are reported[6]. Available estimates of the case fatality ratio, i.e. the proportion of cases that are fatal (CFR,[7,8]), can be used to estimate the number of cases who would have shown symptoms at the same time as the fatal cases. We developed a model to use CFR alongside other epidemiological factors underpinning disease transmission to infer the likely number of cases in a population from newly reported deaths.

Our approach involves two steps: first, reconstructing historic cases by assuming non-fatal cases are all undetected, and, second, model epidemic growth from these cases until the present day to estimate the likely number of current cases. We account for uncertainty in the epidemiological processes by using stochastic simulations for estimation of relevant quantities.

Two pieces of information are needed to reconstruct past cases: the number of cases for each reported death, and their dates of symptom onset. Intuitively, the CFR provides some information on the number of cases, as it represents the expected number of deaths per case, so that CFR^{-1} corresponds to the expected number of cases per death. In practice, the number of cases until the first reported death can be drawn from a Geometric distribution with an event probability equal to the CFR. Note that while our approach could in theory use different CFR for each case (to account for different risk groups), our current implementation uses the same CFR for all cases in a simulation. Dates of symptom onset are simulated from the distribution of the time from onset to death, modelled as a discretised Gamma distribution with a mean of 15 days and a standard deviation of 6.9 days[9].

Once past cases are reconstructed, we use a branching process model for forecasting new cases[10,11]. This model combines data on the reproduction number (R) and serial interval distribution to simulate new cases 'y_t' on day 't' from a Poisson distribution:

$$y_{t+1} \sim poisson(\lambda_t) \text{ with } \lambda_t = R \sum_{s \le t} y_s w(t-s)$$

where w(.) is the probability mass function of the serial interval distribution. More details on this simulation model can be found in Jombart et al.[11]. Optionally, this model can also incorporate heterogeneity in transmissibility using a Negative Binomial distribution instead of Poisson. The serial interval distribution was characterized as a discretized Lognormal distribution with mean 4.7 days and standard deviation 2.9 days[12]. We assume that past cases caused secondary transmissions independently (i.e. are not ancestral to each other), so that simulated cases for each death can be added. This assumption is most likely to be met when reported deaths are close in time. As the time between reported deaths increases, past cases may come from the same epidemic trajectory rather than separate, additive ones, in which case our method would overpredict epidemic size.

Further details on model design and parameters values are provided in Supplementary Material. Our approach is implemented in the R software[13] and publicly available as R scripts (see *Extended data*)[14], as well as in a user-friendly, interactive web-interface available at: https://cmmid.github.io/visualisations/inferring-covid19-cases-from-deaths[2].

Results

How many cases for a single death?

We first used our model to assess likely epidemic sizes when an initial COVID-19 death is reported in a new location. We ran simulations for a range of plausible values of R (1.5, 2 and 3) and CFR (1%, 2%, 3% and 10%), assuming a single death on the 1st March 2020[8]. 25,000 epidemic trajectories were simulated for each parameter combination. Simulations for an 'average severity' scenario[8] with $R = 2$ and CFR = 2% show that by the time a death has occurred, hundreds to thousands of cases may have been generated in the affected population (Figure 1). Results vary widely across other parameter settings, and amongst simulations from a given setting (Table 1), with higher R and lower CFR leading to higher estimates of the numbers of cases. However, a majority of settings give similar results to our 'average' scenario, suggesting that a single death is likely to reflect several hundreds of cases. Results were qualitatively unchanged when incorporating heterogeneity in the model using recent estimates[15], but prediction intervals were wider (*Extended data*).

Recently affected countries

We applied our approach to three countries which recently reported their first COVID-19 deaths (Spain, Italy, and France), using the same range of parameters as in the single-death analysis. In order to compare predictions to cases actually reported in these countries, projections were run until 4th March. Overall, predictions from the model using the baseline scenario ($R = 2$, CRF = 2%) were in line with reported epidemic sizes (Table 2). Results from other scenarios are presented in the *Extended data*. Actual numbers of reported cases fell within the 50% quantile intervals of simulations in all three countries Italy (median: 1 294 ; $QI_{50\%}$: [390 ; 3 034]; reported: 2 037), France (median: 592 ; $QI_{50\%}$: [177 ; 1 705]; reported: 190) and Spain, (median: 202 ; $QI_{50\%}$: [95 ; 823]; reported 202).

Discussion

Several limitations need to be considered when applying our method. First, our approach only applies to the deaths of patients who have become symptomatic in the
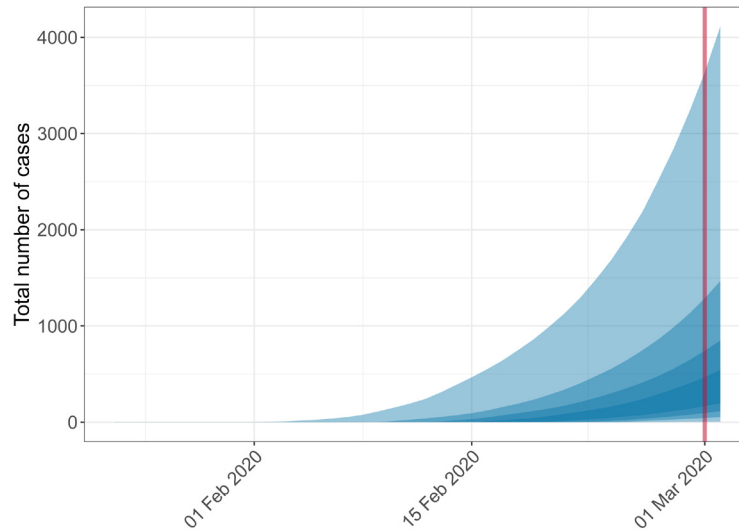
Figure 1. Example of simulated epidemic trajectories from a single death. This figure shows results of 200 simulations using a CFR of 2% and *R* of 2 based on a hypothetical situation where a single death occurred on the 1st March 2020, represented by the red line. Ribbons of different shades represent, from the lightest to the darkest, the 95%, 75%, 50% and 25% quantile intervals.

Table 1. Inferred number of cases for a single death. Inferred number of cases after detection of a single death under different values of the reproduction number, and case fatality ratio. We estimate the number of expected cases in the population at the day the death occurred, and present median, 50%, and 95% estimates of the quantile interval.

R	Median	Lower 95% Quantile Interval	Lower 50% Quantile Interval	Upper 50% Quantile Interval	Upper 95% Quantile Interval
CFR 1%					
1.5	252	5	102	596	2 572
2	519	9	174	1 477	8 325
3	1 733	37	541	7 461	138 624
CFR 2%					
1.5	132	2	52	294	1,110
2	276	5	93	780	5 694
3	964	19	300	4 174	49 137
CFR 3%					
1.5	75	2	27	191	757
2	181	4	60	465	2 515
3	719	7	173	3 100	89 909
CFR 10%					
1.5	29	0	10	65	219
2	46	0	15	136	1,020
3	245	2	63	983	30 708

location considered, which should usually be the case in places where traveler screening is in place. We also assume constant transmissibility (*R*) over time, which implies that behavior changes and control measures have not taken place yet, and that there is no depletion of susceptible individuals. Consequently, our method should only be used in the early stages of a new epidemic, where these assumptions are reasonable. Similarly, the assumption that each death reflects independent, additive

Table 2. Inferred number of cases for several countries assuming CFR of 2% and R of 2. All values are presented for the 4th of March 2020 for different countries. We present the predicted case counts as their median, 50%, and 95% estimates of the quantile interval. * First suspected death due to within country transmission.

Country	Date of first death*	Initial deaths	Reported cases	Median	Lower 95% Quantile Interval	Lower 50% Quantile Interval	Upper 50% Quantile Interval	Upper 95% Quantile Interval
Spain	4th March	1	202	263	8	95	823	7 829
Italy	26th Feb	1	2 037	1 294	33	390	3 034	19 487
France	21st Feb	1	190	592	10	177	1 705	7 501

epidemic trajectories is most likely to hold true early on, when reported deaths are close in time (e.g. no more than a week apart). Used on deaths spanning longer time periods, our approach is likely to overestimate epidemic sizes.

Contact tracing has been shown to be an efficient control measure when imported cases can be detected early on [16], in addition to permitting the estimation of key epidemiological parameters[12]. When the first cases reported in a new location are mostly deaths, however, our results suggest that the underlying size of the epidemic would make control via contact tracing extremely challenging. In such situations, efforts focusing on social distancing measures such as school closures and self-isolation may be more likely to mitigate epidemic spread.

Data availability
Underlying data
All data underlying the results are available as part of the article and no additional source data are required.

Extended data
Zenodo: Extended data for: Inferring the number of COVID-19 cases from recently reported deaths. http://doi.org/10.5281/zenodo.3733289[14].

This project contains the file 'extended_data' (PDF), which contains supplemental information and methodological details regarding the model described in this article.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability
The Shiny app using the model is available at: https://cmmid.github.io/visualisations/inferring-covid19-cases-from-deaths.

Source code and R scripts available at: https://github.com/thibautjombart/covid19_cases_from_deaths.

Archived code at time of publication: http://doi.org/10.5281/zenodo.3733047[2].

License: Code is available under an MIT License; other documentation is available under a CC-BY 4.0 License.

References

1. WHO: **Novel Coronavirus – China**. 2020; [cited 2020 Mar 4].
    Reference Source

2. Jombart T, Abbott S, Gimma A, *et al.*: **Inferring COVID-19 cases from recent death. (Version 1.0.0).** *Zenodo*. 2020.
    http://www.doi.org/10.5281/zenodo.3733047

3. World Health Organization: **Coronavirus disease 2019 (COVID-19). Situation Report 34**. 2020.
    Reference Source

4. World Health Organization: **Pneumonia of unknown cause--China**. Geneva: WHO. 2020.
    Reference Source

5. Lauer SA, Grantz KH, Bi Q, *et al.*: **The incubation period of 2019-nCoV from publicly reported confirmed cases: estimation and application**. *medRxiv*. 2020.
    Publisher Full Text

6. De Salazar PM, Niehus R, Taylor A, *et al.*: **Estimating underdetection of internationally imported COVID-19 cases**. *medRxiv*. 2020.
    Reference Source

7. Wilson N, Kvalsvig A, Barnard LT, *et al.*: **Estimating the Case Fatality Risk of COVID-19 using Cases from Outside China**. *medRxiv*. 2020.
    Publisher Full Text

8. MIDAS-network: **COVID-19 parameter estimates**. [cited 2020 Mar 5].
    Reference Source

9. 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.** *J Clin Med*. 2020; **9**(2): pii: E538.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Nouvellet P, Cori A, Garske T, *et al.*: **A simple approach to measure transmissibility and forecast incidence.** *Epidemics*. 2018; **22**: 29–35.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. Jombart T, Jarvis CI, Mesfin S, *et al.*: **The cost of insecurity: from flare-up to control of a major Ebola virus disease hotspot during the outbreak in the Democratic Republic of the Congo, 2019.** *Euro Surveill*. 2020; 25(2): 1900735.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Nishiura H, Linton NM, Akhmetzhanov AR: **Serial interval of novel coronavirus (2019-nCoV) infections**. *medRxiv*. 2020.
    Publisher Full Text

13. R Core Team: **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing; 2020.
    Reference Source

14. Jombart T: **Extended data for: Inferring the number of COVID-19 cases from recently reported deaths.** *Zenodo*. 2020.
    http://www.doi.org/10.5281/zenodo.3733289

15. Riou J, Althaus CL: **Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020.** *Euro Surveill*. 2020; **25**(4): 2000058.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

16. Bernard Stoecklin S, Rolland P, Silue Y, *et al.*: **First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020.** *Euro Surveill*. 2020; **25**(6): 2000094.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✓ ✓

✓ **Camille Pelat**

Santé publique France, Saint-Maurice, France

This article describes a statistical modeling method for estimating the number of COVID-19 cases from the first reported deaths in a defined location. The described methodology can provide useful information for decision making, especially as a Shiny app has been developed for facilitating quick application of the method by public health practitioners, and the R code has been made available.

Introduction:
- I would be interested to see in the text a few words about how many (and which) countries found themselves in the situation of observing no COVID-19 case before the first deaths were reported. The reference provided (number 2) is not really specific about this point.

Methods:
- The statistical method is well described and seems sound.

- I have a minor comment: in practice, published estimates of the CFR and R will be used as input parameters for the model. These estimates are derived from samples and are usually published with a certain measure of uncertainty, typically the standard deviation or a confidence interval. My understanding is that this estimation uncertainty on these input parameters is not taken into account in the prediction model: instead, the CFR and R are held constant for all simulations drawn with a set of parameter.

- Taking into account the uncertainty on these input parameters may lead to even greater prediction intervals, but may reflect more completely the uncertainty about the total number of cases given the current knowledge about the disease at a certain point in time. This could be done, for example, by drawing the CFR in a Beta distribution with a and b derived from the published mean and sd instead of holding it constant. In the Shiny app, the user could provide the confidence interval.

Discussion:

○ It would be interesting if the authors could comment on the availability of other published methods developed for inferring the number of cases based on reported deaths. If such methods exist, how do they compare in their approach and results with the proposed one? What are the comparative strengths of the proposed method?

○ typo error: "theunderlying size…"

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: biostatistics, public health

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 19 May 2020

https://doi.org/10.21956/wellcomeopenres.17312.r38539

✓ **Benjamin M. Bolker** 🆔
Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada

This is a useful, technically correct, and clearly written contribution.

Could the authors comment on how much extra mileage one gets/advantages of this approach

relative to simply saying that the current number of cases is approximately equal to 1/CFR? That is, does one have to reconstruct the past history to know how much trouble one is currently in?

What would the effect of a heterogeneous CFR be? (I believe this would correspond e.g. to a 'beta-Geometric distribution', unless one instead wanted to treat it as a finite mixture of probabilities for discrete risk categories).

It would be nice to have a little more detail (i.e. a few sentences) on the simulation procedure. I see how to get from CFR and deaths to a total number of preceding cases, and how to simulate times of symptom onset for the observed deaths. It's not completely obvious to me how to get from there to 'history of past cases' (i.e. incidence over time); does one run the renewal process backward in time? Or use branching-process theory to find the time distribution of symptom onset of the index case given the current size of the epidemic?

Please clarify "We assume that past cases caused secondary transmissions independently (i.e. are not ancestral to each other), so that simulated cases for each death can be added." Does this mean that you assume that all observed deaths are from separate lineages/transmission chains? (The last sentence of the paragraph suggests that, but the initial statement could probably be clearer.) (Does this assumption even matter if we are in the branching-process regime?).

I appreciate that the authors are trying to keep things simple, and thus the scenario-based approach (try the model for a range of CFR/R values and see what is implied) is useful. I note that the confidence intervals are already very wide (that's part of the point), but there are several quantities that are treated as known (delay distribution, serial interval distribution); I wonder how sensitive the results are to these assumptions (probably not much - I'm guessing that with R specified they might only change the timing, not the numbers). Given that the authors are already basing the answers on 25,000 solutions, it might not be too hard to construct point estimates and intervals based on a prior/uncertainty distribution of R and CFR (rather than constructing separate scenarios), and allowing for uncertainty in the delay and serial distributions.

Minor comments/typos:
  ◦ Intro, line 1; methods, l. 7: extra comma inside parens before superscript refs?)

  ◦ "use [a] different CFR for each case"

  ◦ "parameters" values

  ◦ "theunderlying"

  ◦ "schoolclosures"

  ◦ In tables 1 and 2 consider stating "2.5% quantile, 25% quantile, 50% quantile, 97.5% quantile" (rather than lower/upper x 95%/50%) ?

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ecology, evolution, epidemiological modeling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.