Evaluating currently available open source tools and data for identifying temporal variation in transmission in China during the 2019-nCoV outbreak

S. Abbott (1), J. Hellewell (1), J. D. Munday (1), S. Flasche (1), CMMID nCoV team (1), S. Funk (1).
Note: this is preliminary analysis and has not yet been peer-reviewed.

Correspondence to: sam.abbott@lshtm.ac.uk

Introduction

- Background
- Detail
- What and aim

To identify changes in the reproduction number, and rate of spread during the course of the 2019-nCoV outbreak whilst highlighting potential biases and evaluating currently available open source tools and data.

Methods

Reporting delays were first estimated using a line-list of cases compiled from media and other reports [1]. Confirmed cases were used from the 22nd of January due to due to changes in reporting practices making cases reported before this point incomparable. The reporting delay was assumed to remain constant over time. We fitted a geometric, Poisson, and a negative binomial distribution to the observed delays and selected the best fit using the Chi-squared statistic. If no good fit was determined using a p-value threshold of 0.05, then the reporting delay was instead sampled from the empirical delays in the line-list. We then took 100 samples from the chosen delay distribution for each reported case and combined these with report dates to construct a distribution of onset dates for each case. To account for censoring, i.e. cases that have not yet been confirmed but will show up in the data at a later time, we used the `nowcast` function from the `surveillance` R package on a sub-sample of cases (1%) with a prediction lag of 1 day and a maximum delay of 18 days [2]. This sub-sampling was required to allow the analysis to be conducted in a reasonable time-frame with the compute resources available. We did not account for potential reporting biases that might occur due to changes in the growth rate of the outbreak over time.

We used the inferred number of cases to estimate the reproduction number on each day using the *EpiEstim* R package [4]. This uses a combination of the serial interval distribution and the number of observed cases to estimate the reproduction number at each time point [5,6], which were then smoothed using a 3-day time window. We tested three scenarios for the serial interval distribution: according to estimates from the early epidemic, with a mean of 7.5 days and standard deviation of 3.4 days [7]; SARS-like, with a mean of 8.4 days and standard deviation of 3.8 days [8]; and MERS-like, with a mean of 6.8 days and standard deviation of 4.1 days [9].

We estimated the rate of spread (r) using linear regression with time as the only exposure and logged cases as the outcome for the overall course of the outbreak [10]. The adjusted R^2 value was then used to assess the goodness of fit. In order to account for potential changes in the rate of spread over the course of the outbreak we used a 3-day sliding window to produce time-varying estimates of the rate of spread and the adjusted R^2. The doubling time was then estimated using $\ln(2)\frac{1}{r}$ for each estimate of the rate of spread.

We repeated all analyses using case counts by date of report, inferred case counts by date of onset, and inferred case counts by date of onset adjusted for reporting truncation [11,12]. We used our pipeline to both now-cast the current outbreak and to now-cast based on data available on the 27th of January, as well as the 11th of February. We then compared the now-cast from the 27th of January to available onset data and to data based on report date that has since become available [11,12].

We report the 95% confidence intervals for all measures using the 2.5% and 97.5% quantiles. Regulary updated nowcasts and estimates of the time-varying reproduction number and rate of spread are available here: *link to updatedable report*. All code for this analysis is available as an R package (*reference*).

Results

Confirmation delays

Our analysis of confirmation delays was based on 401 cases with reporting dates up to the 13th of February with the cutoff being the 22nd of January. The mean delay from onset to confirmation was 5.6 days with a standard deviation of 3.5 days. There is some indication that both the mean delay and the standard deviation of the delay have declined over time. In particular, there may have been a surge of confirmations, that possibly occurred a long time after symptom onset, after the National Health Commission Task Force visited Wuhan [13]. This potential trend over time and the potential bias it may introduce is the basis for our use of a cut-off when estimating the reporting delay. A time-varying reporting delay may be more appropriate but currently available data sources do not support this.

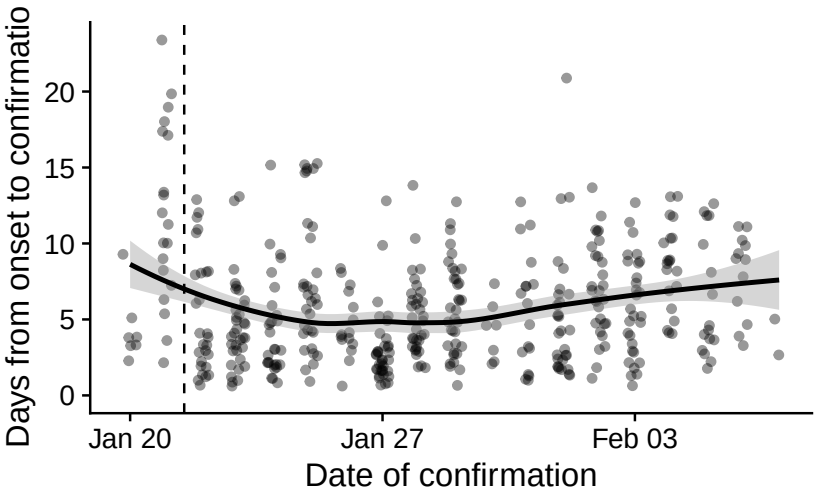

Figure 1: Delays from onset to confirmation, by date of confirmation. Every dot is one reported case with onset and confirmation dates reported. The black line indicates a loess-smoothed trend. Data to the left of the dotted line may be unreliable due to the spike in reporting on the 21st of January. Data prior to the 20th of January is not shown.

Adjusting the number of reported cases for confirmation delays

Our approach to estimating the number of cases by date of onset using cases by date of report and a delay distribution fit to the available data reproduced case counts by onset date well up to the 27th of January (Figure 2, A). After adjusting for truncation due to reporting delays (leading to a backlog of cases not yet reported) we found that the available onset data substantially underestimated the number of cases with symptom onsets each day (Figure 2, B).). An apparent decline in cases by onset date from the 25th of January was not matched by a corresponding later decline in the counts by date of report (Figure 2, B). Our adjusted estimates based on all currently available data indicate that the increase in cases by date of onset slowed in late January and early February, before speeding up until around the 4th of February when they began to slow again. However, our results are highly susceptible to the apparent change in case definition on

the 12th of February. The adjusted estimates predicted that cases by onset date would continue to decline when they in fact just over 13,000 cases were reported on day. We estimate that on the 11th of February there were 2700–16910 cases that developed symptoms.

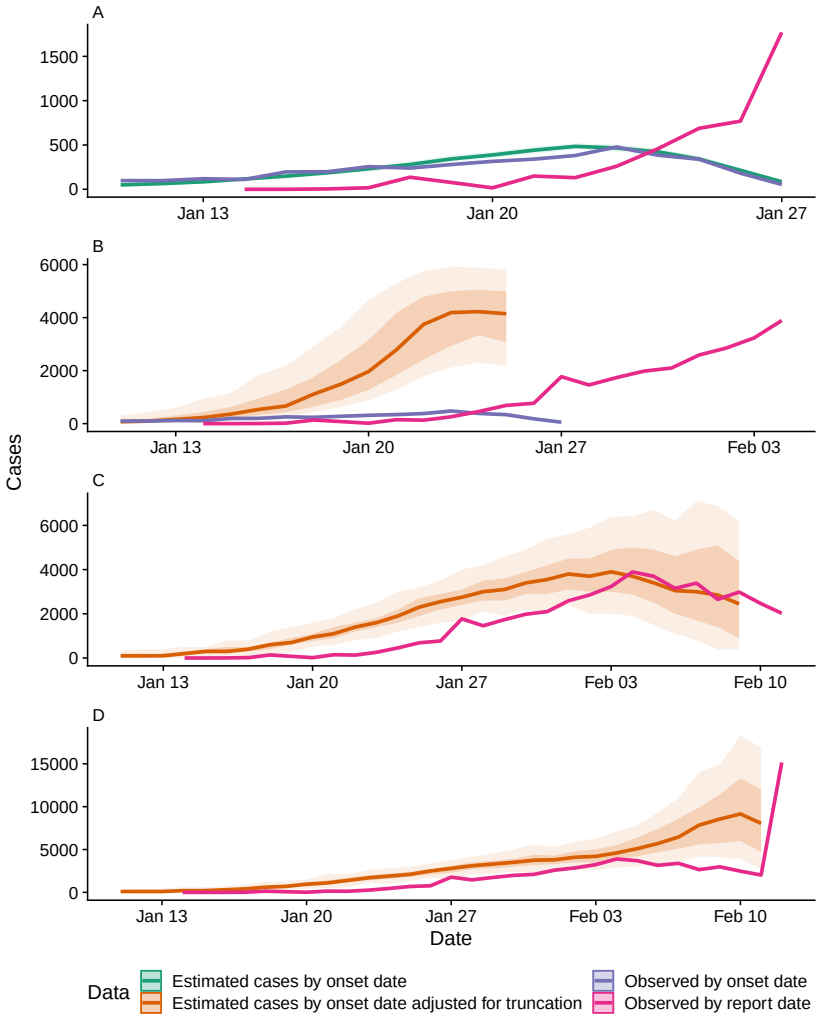

*Figure 2: A.) Comparison of cases by report date (up to the 27th of January), onset date (until the 27th of January) [11], and our estimated cases by onset date (based on reported data up to the 27th of January). B.) Comparison of cases by report date (up to the 5th of February), onset date (until the 27th of January) [11], and our estimated cases by onset date adjusted for truncation (based on reported data up to the 27th of January). C.) Comparison of cases by report date (up to the 11th of February), and our estimated cases by onset date adjusted for truncation (based on reported data up to the 11th of February). D.) Comparison of cases by report date and our estimated cases by onset date adjusted for truncation (based on data up to the **13th of February.*

Time-varying reproduction number

Our analysis indicates that the time-varying reproduction number declined from the 14th of January until the beginning of February, regardless of the serial interval scenario used (Figure 3). However, as discussed in the previous section, recent changes in the case definition has led to an increase in the number of estimated cases by onset, which has in turn resulted in the most recent estimates of the reproduction number increasing. Time-varying reproduction number estimates based solely on reported cases are consistently larger but also show the same decreasing trend over the course of the outbreak. Time-varying reproduction number estimates based on unadjusted case counts by onset date were initially higher than those estimated using

adjusted case counts by onset date. The impact of the adjustment is apparent in the most recent estimates of the reproduction number, with those based on unadjusted case counts being lower than those based on adjusted case counts. As expected, uncertainty in the time-varying reproduction number estimates is initially high and then decreases as the number of cases increases. Initially, the assumed serial interval scenario has a large impact on estimates of the basic reproduction number but this effect weakens over time with the latest estimates for the basic reproduction being 1.1–2.6 with a serial interval as estimated in the early epidemic, 1.1–2.4 with a serial interval assumed to be MERS-like, and 1.2–2.8 with a serial interval assumed to be SARS-like. The peak (based on the unadjusted case counts by onset) reproduction number estimated for each scenario was 6.8–13.2 with a serial interval as estimated in the early epidemic, 4.9–8.8 with a serial interval assumed to be MERS-like, and 8.4–16.3 with a serial interval assumed to be SARS-like.

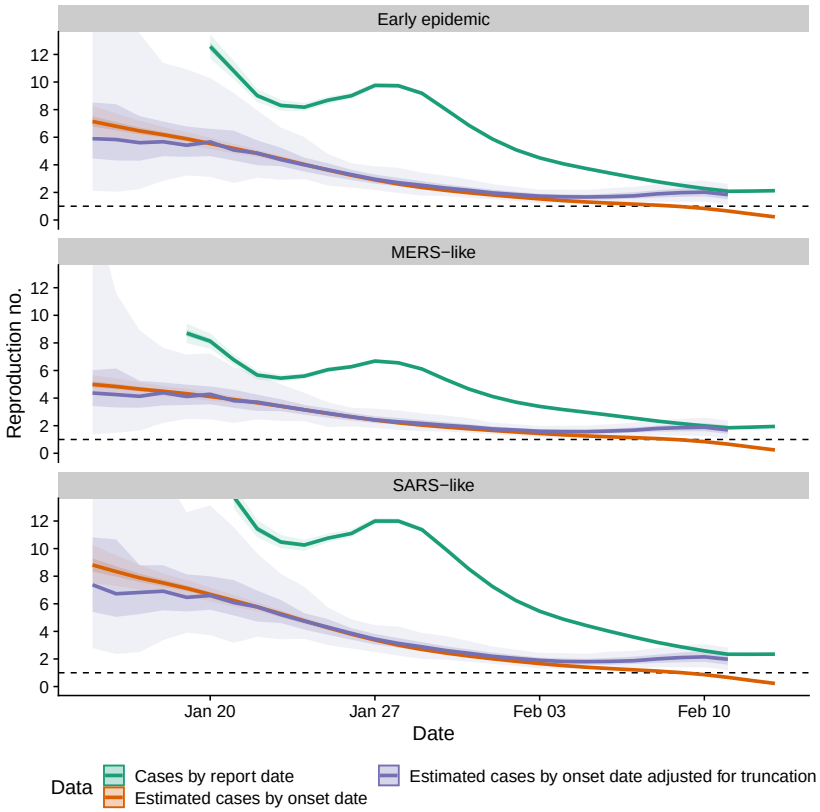

Figure 3: Time-varying reproduction number for different assumed serial interval distributions using both estimated cases by onset date and cases by report date

Time-varying rate of spread and doubling time

Overall, we estimate that the overall rate of spread for the 2019-nCoV is currently 0.11–0.14 which translates to a doubling time of 4.8–6.4 days. This estimate was derived using adjusted case counts by onset date, in comparison we found that the rate of spread was 0.022–0.047 (with a doubling time of 15–32) when unadjusted case counts by onset date were used. Using cases counts by report date we estimate that the rate of spread of the outbreak is 0.25–0.27 with a doubling time of 2.6–2.7 days. The overall estimate had a lower adjusted R^2 than the estimate based on cases by date of report indicating a worse fit for the exponential model (0.77–0.89 vs 0.94–0.94). Our initial estimates of the rate of spread had a large degree of uncertainty but our findings indicate a continual decrease until the beginning of February. This finding was also susceptible to the change in case definition on the 12th of February with previous now-casts estimating a continual decline in the rate of growth (Figure 4). The quality of the fit for the exponential model to the estimated cases by onset data has followed a similar trend. The rate of spread was consistently higher when assessed using cases by report date alone and showed less evidence of a decrease over time. When estimated using unadjusted case counts by onset alone, we found that the rate of spread continued to decline.

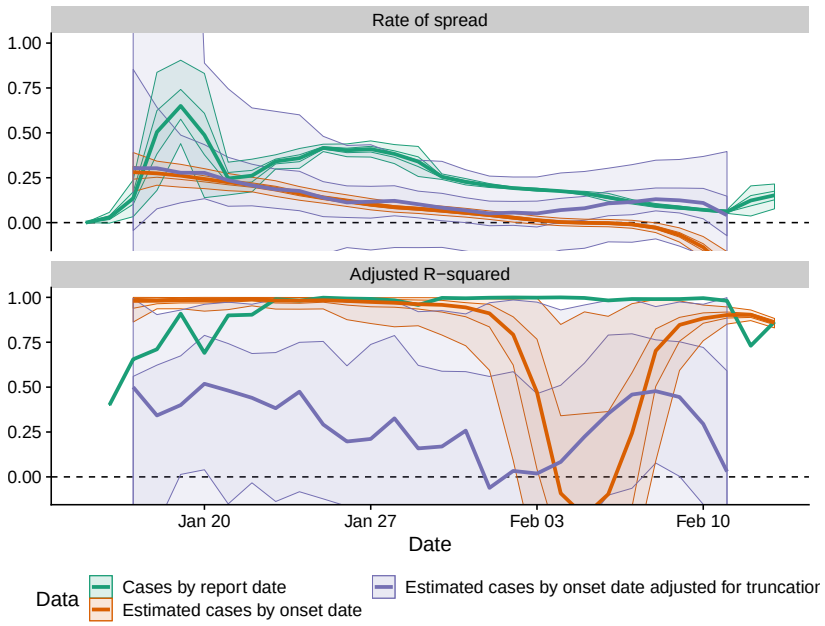

Figure 4: Time-varying rate of spread and adjusted R-squared using a 7-day sliding window.

Discussion

We found that our approach could reproduce observed cases by date of onset when these data were available. After adjusting for truncation due to reporting delays we found that on the 11th of February there were 2700–16910 cases that developed symptoms. We found that our approach was highly susceptible to changes in case definitions with the update to the case definition on the 12th of February leading to a change in our findings from a decrease in cases by date of onset since early February to an increase through to the 10th of February. In all serial interval scenarios considered we found that the reproduction number declined from the 14th of January through to the beginning of February. Based on data available on the 13th of February we estimate the current reproduction to be 1.1–2.6 with a serial interval as estimated in the early epidemic, 1.1–2.4 with a serial interval assumed to be MERS-like, and 1.2–2.8 with a serial interval assumed to be SARS-like. However, these findings were again susceptible to changes in the case definition with earlier nowcasts forecasting lower reproduction numbers. Estimates for the time-varying rate of spread followed similar trends as those observed in the reproduction number as estimated by the `EpiEstim` R package [4] and were similarly impacted by changes in the case definition. Overall, we found an estimate of 0.11–0.14 for the rate of spread which translates to a doubling time of 4.8–6.4 days. Our findings should not be overinterpreted due to the numerous issues we have identified with the data and have not been able to fully account for.

Our study combines open source tools and data to produce estimates of the number of cases by date of symptom onset, the time-varying reproduction number, and the time-varying rate of spread. It is available as an R package allowing this approach to be built upon by others and to be updated as the current outbreak continues. We identified numerous issues with the available data that impacted our ability to produce reliable estimates of key epidemiological parameters. We found that changes in the underlying case definition impacted our findings dramatically. This made it difficult to draw definite conclusions. If data were available using a consistent case definition it may have been possible to account for this. However, we were able to produce estimates of the number of cases per day by onset date that accounted for reporting bias. This means that our findings can be more readily used to assess the current trajectory of the outbreak than case counts based on date of report or unadjusted case counts based on date of onset, although this can only be done if it is assumed that the reporting delay and case definition is static. As we have shown, when this is not the case our adjusted case counts by onset date were highly variable from day-to-day. We used a publically available line-list to estimate the delay distribution [1]. Unfortunately this data resource appears to be no longer updated and in the absence of other publically available line-lists on which to base

our delay distribution we were unable to explore varying it by time. However, as our analysis is publically available, and easily reproducible, it may be the case that others with access to propriety data may be able to expand on our findings using a time-varying delay distribution. This may account for some sensitivity caused by the changing case definitions.

The 2019-NCoV outbreak is currently ongoing. Our analysis is available as an automated pipeline (in the form of an R package), meaning that it can be regularly updated and expanded upon, by ourselves or others as new data becomes available. We based our estimates of the reporting delay on a line-list produced by aggregating media reports [1]. This appears to be no longer be maintained meaning that changes in reporting practices over time can not be captured. Fortunately our approach is line-list agnostic and our analysis may be repeated once new data becomes available. Unfortunately our approach to dealing with onset dates being truncated, which was based on the `surveillance` R package [2], was not robust to missing data and was computationally expensive. These limitations meant that we had to introduce an initial ad-hoc sampling step to generate a complete sample line-list and that we then had to downsample this line-list to make the analysis feasible in real-time. Further developments of these approaches may make them more robust in future outbreaks.

This study highlights the impact of a changing case definition and reporting delay on estimates of epidemiological parameters. We were able to account for some of the biases present in the publically available data. This allowed us to estimate that 2700–16910 cases developed symptoms on the 11th of February and that the latest estimate of the reproduction number was 1.1–2.6. However, these findings cannot be overinterpreted as we were not able to adjust for a changing case definition. This study has evaluated currently available open source tools and found that additional work is needed to make them robust to data qaulity issues during ongoing outbreaks. We have made our work available as an R package so that it can be expanded upon, either using new methodology or additional data sources. We have also produced a dynamic report using our pipline that will continue to be updated, and developed as the outbreak continues.

References

1 Xu B, Gutierrez B, Hill S *et al.* Epidemiological Data from the nCoV-2019 Outbreak: Early Descriptions from Publicly Available Data. 2020.

2 Meyer S, Held L, Höhle M. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software* 2017;**77**:1–55. doi:10.18637/jss.v077.i11

3 R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria:: R Foundation for Statistical Computing 2019. https://www.R-project.org/

4 Cori A. *EpiEstim: Estimate time varying reproduction numbers from epidemic curves*. 2019. https://CRAN.R-project.org/package=EpiEstim

5 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

6 Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology* 2004;**160**:509–16. doi:10.1093/aje/kwh255

7 Li Q, Guan X, Wu P *et al.* Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*;**0**:null. doi:10.1056/NEJMoa2001316

8 Lipsitch M. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science* 2003;**300**:1966–70.

9 Cauchemez S, Nouvellet P, Cori A *et al.* Unraveling the drivers of mers-cov transmission. *Proceedings of the National Academy of Sciences* 2016;**113**:9081–6. doi:10.1073/pnas.1519235113

10 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

11 Epidemic update and risk assessment of 2019 novel coronavirus. 28 january, 2020. http://www.chinacdc.cn/yyrdgz/202001/P020200128523354919292.pdf

12 Yu G. *NCov2019: Stats of the '2019-nCov' cases*. 2020.

13 Wu P, Hao X, Lau EHY *et al.* Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in wuhan, china, as at 22 january 2020. *Eurosurveillance* 2020;**25**. doi:https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000044