Some challenges of sparse data necessitating strong assumptions in investigating early COVID-19 disease

EClinicalMedicine 000 (2020) 100499

Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

Commentary

Some challenges of sparse data necessitating strong assumptions in investigating early COVID-19 disease

John Norrie

Edinburgh Clinical Trials Unit, Usher Institute, Edinburgh EH16 4UX, UK

ARTICLE INFO

Article History:
Received 25 July 2020
Accepted 26 July 2020
Available online xxx

In this study published in EClinicalMedicine, Du et al. [1] use a novel approach to estimate unseen COVID-19 cases early in the pandemic, when neither awareness of the disease nor suitable testing was available. By retrospectively testing samples from patients seeking treatment for seasonal flu, then calculating the 'COVID-19-to-influenza positives ratio' (CIPR) of SARS-CoV-2 positives to flu positives, and applying the CIPR to observed flu cases, they extrapolate the likely unseen COVID-19 cases.

This method requires many strong assumptions, and generates imprecise estimates given few observed infective events, and is subject to several different selection biases. Such estimates are needed, since for a new virus, accurate assessment of onset date and early transmission dynamics are difficult. These early data are needed to understand pandemic development, and for predicting onset and containing new infective waves in space and time, with localised outbreaks probable. Therefore, we must understand the influence of both strong assumptions and sparse data on model outputs and interpretation.

The authors consider two locations – the original epicentre, Wuhan, China and more recently Seattle, USA. The influence of these strong assumptions and sparse data is most readily seen in Wuhan [2], where 26 adults presenting with influenza-like-illness (ILI) had 4 SARS-CoV-2 and 7 flu positives. This estimated 1386 (95% credible interval (CrI) 420–3793) symptomatic COVID-19 cases (adults >30) in 2-weeks from 30/12/2019. For Seattle [3], 25 SARS-CoV-2 and 442 flu positives from 2353 (299 children, 2054 adults) reporting acute respiratory illnesses (ARI) gives corresponding estimates 2268 (95%CrI 498–6069; children) and 4367 (95%CrI 2776–6526; adults) in 2-weeks from 24/02/2020.

These estimates extrapolate from small numbers (in Wuhan, single figures), generating very wide 95% credible intervals. The Bayesian approach used is appropriate for rare events, allowing incorporation of external information, assuming the 'priors' can be elicited convincingly [4]. The strong assumption is that undetected SARS-CoV-2 to flu positives ratio is constant over the estimation period. However, flu is seasonal [5], whereas COVID-19 seasonality is unknown. Since estimated COVID-19 cases are a scalar multiple of observed flu infections, this assumption is critical. The 2-week estimation period selected should reduce bias from discordant seasonality in the two infections. However, even short estimation periods show high variability. In Wuhan [3] the week following the 2-weeks used showed five SARS-CoV-2 and zero flu infections. So, including this 3rd week, the non-Bayes ratio increases from 4/7 (0.57) to 9/7 (1.29), over double. Along with possible flu reduction from pandemic containment measures [6], this all underlines the fragility of these reported estimates.

In addition, in Wuhan [3], from 54 samples aged < 30, there were zero SARS-CoV-2 and 30 flu positives. The authors chose not to use these data, only estimating symptomatic COVID-19 cases in over 30′s in Wuhan; in Seattle they could estimate for children and adults. Additional to temporal concerns, assumptions are necessary around spatial applicability of the CIPR. Across 13 Wuhan districts, with just 4 SARS-CoV-2 positives, at least 9 districts must have had 0 positives detected. We would be sceptical applying estimates from these data to the whole of China; so, what is reasonable spatial extrapolation? The authors have assumed the ratio applies to all 13 Wuhan districts. The observed district zeros could be within-sampling variability given the estimated ratio, or could indicate no COVID-19 infection in those districts. Both are consistent with these sparse data [7].

A further interpretational challenge is diagnostic test misclassification for both SARS-CoV-2 and flu. Both numerator and denominator of the ratio could have false positives & negatives. Early SARS-CoV-2 RT-PCR tests [8] had modest sensitivity (~75%) with better specificity, with throat swabs having lower sensitivity than nasal samples. Likewise, rapid influenza diagnostic tests (RIDTs) [9] have low to moderate sensitivity (50–70%) with better specificity (90–95%). So false negatives will be more common than false positives in both, but it is the ratio of these misclassifications that matters. Interestingly, in Seattle it was ARI rather than ILI (Wuhan) that was the treatment seeking behaviour, raising the additional complexity of needing to test for multiple respiratory conditions.

The estimation of the date of first COVID-19 infection used a model incorporating the epidemic doubling rate, taken from a separate study [10], and author's estimated COVID-19 infections across the districts, with uncertainty expressed as 95% credible intervals generated by Monte Carlo resampling. We again see the influence of

DOI of original article: http://dx.doi.org/10.1016/j.eclinm.2020.100479.
E-mail address: j.norrie@ed.ac.uk

https://doi.org/10.1016/j.eclinm.2020.100499

small numbers, with the 95% credible interval for this date of first onset stretching to 7 weeks for Wuhan (from late October to mid-December 2019), while for Seattle, with more data, around 3 weeks (from late December 2019 to mid-January 2020).

Nonetheless, despite all these challenges, the authors have developed a novel and useful approach to estimate important unknowns, including the onset date of local outbreaks. Such estimates inform transmission models, debated by governments and their critics, when assessing the rapidity and adequacy of public health response to outbreak control. It is important to understand model limitations, appreciating the 95% credible intervals only reflect the estimated precision under these strong assumptions. Further validation is important in subsequent COVID-19 waves, with larger samples, better tests, and more accurate flu statistics available, and model extension to include co-infections in winter surges. In the meantime, these innovative methods are welcome, but should be used cautiously, understanding the fragility of estimates to sparse data and strong assumptions.

Declaration of Competing Interests

Professor John Norrie is employed by the University of Edinburgh, and as Chair of the Medical Research Council / National Institute of Health Research (MRC/NIHR) Efficacy and Mechanisms Evaluation (EME) Funding Committee.

References

[1] Du Z, Javan E, Nugent C, Cowling BJ, Meyers LA. Using the COVID-19 to influenza ratio to estimate early pandemic spread in Wuhan, China and Seattle, US. EClinicalMedicine 2020. doi: 10.1016/j.eclinm.2020.100479.
[2] Kong W-H, Li Y, Peng M-W, Hong D-G, Yang X-B, Wang L, et al. SARS-CoV-2 detection in people with influenza-type illness. Nat Microbiol 2020;5:675–8.
[3] Chu HY, Englund JA, Starita LM, Famulare M, Brandsetter E, Nickerson DA, et al. Early detection of COVID-19 through a citywide pandemic surveillance platform. N Engl J Med 2020. doi: 10.1056/NEJMc2008646.
[4] Hampson LV, Whitehead J, Eleftheriou D, Brogan P. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. Stat Med 2014;33:4186–201.
[5] Lofgren Eric, Fefferman NH, Naumov YN, Gorski J, Naumova EN. Influenza Seasonality: underlying Causes and Modeling Theories. J. Virol. May 2007;81(11):5429–36. doi: 10.1128/JVI.01680-06.
[6] Chan K-H, Lee P-W, Chan CY, et al. Monitoring respiratory infections in covid-19 epidemics. BMJ 2020;369:m1628.
[7] Leitgöb H. Analysis of Rare Events. In: Atkinson P, Delamont S, Cernat A, Sakshaug JW, Williams RA, editors. SAGE research methods foundations; 2019. doi: 10.4135/9781526421036863804.
[8] Watson J., Whiting P.F., Brush J.E.Interpreting a COVID-19 test/ BMJ2020;369: m1808 doi: 10.1136/bmj.m1808(Published 12 May 2020)
[9] https://www.cdc.gov/flu/professionals/diagnosis/overview-testing-methods.htm, accessed 20 July 2020.
[10] Du Zhanwei, Wang Lin, Caucherucz Simon, Xu Xiaoke, Wong Xianwen, Cowling Benjamin, et al. Risk for transportation of coronavirus disease from Wuhan to other cities in China. Emerg Infectious Dis J 2020;26:1049.

Please cite this article as: J. Norrie, Some challenges of sparse data necessitating strong assumptions in investigating early COVID-19 disease, EClinicalMedicine (2020), https://doi.org/10.1016/j.eclinm.2020.100499