SENTINEL EVENT SURVEILLANCE TO ESTIMATE TOTAL SARS-COV-2 INFECTIONS, UNITED STATES

A PREPRINT

Andrew A. Lover*
Department of Biostatistics and Epidemiology
University of Massachusetts- Amherst
Amherst, MA
alover@umass.edu

Thomas McAndrew†
Department of Biostatistics and Epidemiology
University of Massachusetts- Amherst
Amherst, MA
mcandrew@umass.edu

March 17, 2020

ABSTRACT

Human infections with a novel coronavirus (SARS-CoV-2) were first identified via syndromic surveillance in December of 2019 in Wuhan China. Since identification, infections (coronavirus disease-2019; COVID-19) caused by this novel pathogen have spread globally, with more than 180,000 confirmed cases as of March 16, 2020. Effective public health interventions, including social distancing, contact tracing, and isolation/quarantine rely on the rapid and accurate identification of confirmed cases. However, testing capacity (having sufficient tests and laboratory throughput) to support these non-pharmaceutical interventions remains a challenge for containment and mitigation of COVID-19 infections.

We undertook a sentinel event strategy (where single health events signal emerging trends) to estimate the incidence of COVID-19 in the US. Data from a recent national conference, the Conservative Political Action Conference, (CPAC) near Washington, DC and from the outbreak in Wuhan, China were used to fit a simple exponential growth model to estimate the total number of incident SARS-CoV-2 infections in the United States on March 1, 2020, and to forecast subsequent infections potentially undetected by current testing strategies. Our analysis and forecasting estimates a total of **54,100** SARS-CoV-2 infections (80 % CI 5,600 to 125,300) have occurred in the United States to March 12, 2020.

Our forecast predicts that a very substantial number of infections are undetected, and without extensive and far-reaching non-pharmaceutical interventions, the number of infections should be expected to grow at an exponential rate.

Keywords Infectious disease surveillance · Emerging pathogens · Outbreak science · Epidemiology of epidemics

1 Introduction

To date, testing for COVID-19 cases (clinical and otherwise) has been limited in the US. Hospitals and healthcare facilities that do have tests have limited the testing criteria to those with highest potential risk, to optimize the use of scarce resources [1]. Where widespread testing has occurred, increased prevalences in high-risk groups have been found [2, 3].

Evidence-based public health programming depends on data-driven estimates of disease burden to develop interventions. As such, accurate estimate of the underlying burden of SARS-CoV-2 infection within the US can provide critical information for hospital administrators, county- and state-level Departments of Health, and the US Centers for Disease Control and Prevention (CDC) to optimize and prioritize resource allocation.

*https://www.umass.edu/sphhs/person/faculty/andrew-lover
†http://www.thomasmcandrew.com/

The Conservative Political Action Conference (CPAC), that took place Feb. 26-29, 2020 outside Washington DC is a large annual event that draws approximately 20,000 participants from across the entire US, thus serving as a representative national sample. A single person was reported as 'presumptive positive' for COVID-19 on Mar 7, 2020, and exposure was reported to be prior to the CPAC event [4].

The point incidence of COVID-19 in this defined cohort was used to estimate the national incidence during this period, and subsequent forecasting of current SARS-CoV-2 infections across the entire US. The number of attendees and number of confirmed cases were used to fit a simple exponential growth model to estimate the total COVID-19 infections per day throughout the United States up to March 15th, 2020.

2 Overview of Data and Methods

This analysis proceeded in three stages. Firstly, the point incidence of SARS-CoV-2 infection was calculated for the CPAC cohort. Secondly, With an assumption that this group was a near-random sample of Americans, this proportion (daily incidence rate) was then extrapolated to the entire US population (329 million persons) to provide an estimated national daily incidence on March 1, 2020. Finally, this estimate with appropriate error structures, was projected forward in time using epidemiological parameters from the epidemic in Wuhan, PRC to forecast cumulative infections on March 12, 2020.

Data for this analysis were complied from routine sources; press reports from CPAC; and parameters estimated from modeling studies for the early (exponential growth phase) of the epidemic in China. The number of attendees and the number of confirmed COVID-19 cases at CPAC was collected from press reporting. The epidemic doubling time was taken from previous work which estimated from data on the outbreak in Wuhan, China. This doubling time from the Wuhan data was estimated to be 6.4 days, (95 CrI 5.8–7.1), and we used a lognormal distribution for D following literature values for other doubling times [5, 6].

3 Results and Conclusions

Projecting this total infection burden forward (Fig. 1) using doubling time parameters from Wuhan gives a national cumulative incidence estimate of **54,145** infections (80 % CI 5,647 to 125,274) across the United States up until March 12, 2020. This total includes all infections- asymptomatic, subclinical and clinical cases.

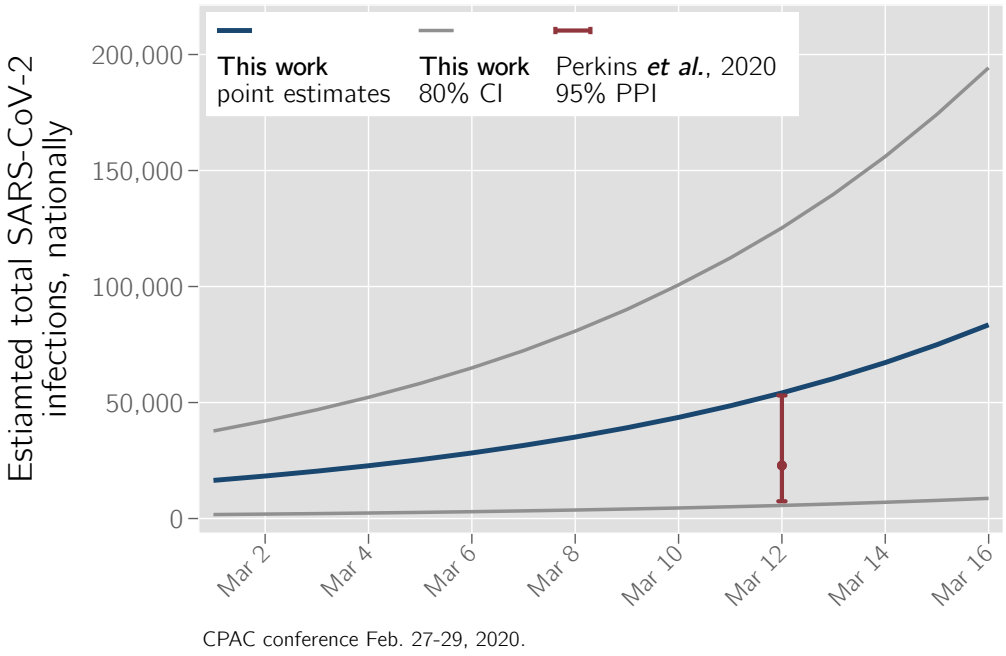

Figure 1: Projected cumulative national SARS-CoV-19 incidence, after March 1, 2020.

This forecasted cumulative number of infections is consistent with those of Perkins *et al.*, from independent data sources and with entirely different modeling frameworks, and greatly strengthens the validity of both sets of estimates [7]. The large number of predicted infections compared against the currently reported total confirmed COVID-19 cases in the United States suggest an important proportion of infections may be currently undetected, while entering a phase of potentially very rapid growth of the epidemic.

4 Limitations of this analysis

This analysis is not without limitations. Our forecasts are based on parameter estimates using best-available data at the time of analysis to inform public health policy (coined "outbreak science" [8]). Demographic data for CPAC attendees was not available, so we are unable to assess how representative this sentinel event may be for the entire US population. This analysis also relies on simplistic models that assumes a well-mixed population, and uses data derived from transmission scenarios that may not represent transmission within the US. The upper limit of total infections should be interpreted with caution, as small uncertainties early in the projections also accumulate exponentially. Finally, this analysis does not account for spatial heterogeneity and diverse population mixing patterns throughout the United States, which are a critical driver of observed transmission.

5 Detailed Methods

5.1 The model

We assumed that cumulative infections up to day t, denoted by I_t, followed an exponential growth model:

$$I_t = I_0\ 2^{t/D}, \tag{1}$$

where I_0 is the number of initial infections, t is the time, in days, since March 1, 2020, and D is the doubling time (in days). Estimates of I_0 and D are needed to forecast the cumulative number of infections at time t.

5.2 Estimating I_0

We estimated total incidence among the US population on March 1st, denoted by I_0, as follows,

$$I_0 = P \cdot \frac{I_0^{(\mathrm{CPAC})}}{P^{(\mathrm{CPAC})}},$$

where P denotes the size of the US population (329M, [9]), $I_0^{(\mathrm{CPAC})}$ the incidence among CPAC attendees and $P^{(\mathrm{CPAC})}$ the number of CPAC attendees. The parameter I_0 represents the assumed fraction of the US population that is infected with COVID19. The incidence estimated from CPAC was extrapolated to the US population based on the assumption of equal incidence in CPAC attendees and the US population.

The initial relative incidence $I_0^{(s)}/P$ was assumed to follow a beta distribution

$$\begin{aligned} I_0^{(s)} &= P \cdot \theta^{(s)}, \\ \theta^{(s)} &\sim \mathcal{B}\left(I_0^{(\mathrm{CPAC})} + I_0^{(\mathrm{CPAC})}, P^{(\mathrm{CPAC})} - I_0^{(\mathrm{CPAC})} + 1 - \frac{I_0^{(\mathrm{CPAC})}}{P^{(\mathrm{CPAC})}}\right). \end{aligned}$$

with a prior over θ, the probability of infection, equal to a $\mathcal{B}\left(I_0^{(CPAC)}, 1 - \frac{I_0^{(\mathrm{CPAC})}}{P^{(\mathrm{CPAC})}}\right)$. The prior is generated by having observed one sample from the general population with incidence "as observed in the CPAC population"

5.3 Estimating D

The doubling time of the disease was taken from previous work using data on the outbreak in Wuhan, China, and was estimated to be 6.4 days, (95 CrI 5.8–7.1) and lognormally distributed [5, 6]

$$\log\left(D^{(s)}\right) \sim N\left(\log(6.4), (1.3/4)^2\right).$$

5.4 Model-based projections

Point estimates for projected incidence were obtained directly using Eq. 1 and the point estimates of doubling time and initial incidence. Uncertainty in the number of infections was obtained by taking $30,000$ Monte Carlo samples of I_0 and D according to their probability densities, and constructing a probability mass function for trajectories, I_t,

$$p(I_t = I) \approx N^{-1} \sum_{i,j} \mathbb{1}_I \left(I_{0i}\, 2^{t/D_j} \right).$$

Specifically, the 80% projection interval for I_t is given by the 10^{th} and 90^{th} percentiles of samples $\{I_t^{(s)}, I_t^{(2)}, \ldots, I_t^{(S)}\}$. Analysis was performed using R [10] for a 16-day window after March 1, in 1-day intervals. Standard 95% and 80% confidence intervals were constructed across the forecast period.

6 Acknowledgements

The authors are grateful for feedback and insightful suggestions from Nick Reich and Leontine Alkema (UMass-Amherst).

References

[1] Jinnong Zhang, Luqian Zhou, Yuqiong Yang, Wei Peng, Wenjing Wang, and Xuelin Chen. Therapeutic and triage strategies for 2019 novel coronavirus disease in fever clinics. *The Lancet Respiratory Medicine*, 2020.

[2] US Centers for Disease Control and Prevention. CDC announces additional COVID-19 infections. Library Catalog: www.cdc.gov.

[3] Robinson Meyer, Erin Kissane, and Alexis Madrigal. The COVID Tracking Project. https://covidtracking.com/.

[4] J. Edward Moreno. CPAC attendee tests positive for coronavirus, March 2020. Library Catalog: thehill.com.

[5] Joseph T. Wu, Kathy Leung, and Gabriel M. Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*, 395(10225):689–697, February 2020. Publisher: Elsevier.

[6] Dong K. Kim. Statistical methods for estimating doubling time in *in vitro* cell growth. *In Vitro Cellular & Developmental Biology. Animal*, 33(4):289–293, 1997. Publisher: [Springer, Society for In Vitro Biology].

[7] T. Alex Perkins, Sean M. Cavany, Sean M. Moore, Rachel J. Oidtman, Anita Lerch, and Marya Poterek. Estimating unobserved SARS-CoV-2 infections in the United States. March 2020.

[8] Caitlin Rivers, Jean-Paul Chretien, Steven Riley, Julie A Pavlin, Alexandra Woodward, David Brett-Major, Irina Maljkovic Berry, Lindsay Morton, Richard G Jarman, Matthew Biggerstaff, et al. Using "outbreak science" to strengthen the use of models during epidemics. *Nature Communications*, 10(1):1–3, 2019.

[9] US Census Bureau. Monthly population estimates for the United States (na-est2019-01).

[10] R Core Team. R: A language and environment for statistical computing, 2013.

Revision History

Revision	Date	Author(s)	Description
1.0	17.1.20	AAL, TMc	created
1.1	17.1.20	AAL	revised plot legends