Clinical Infectious Diseases

BRIEF REPORT

Unmasking the Actual COVID-19 Case Count

Samuel C. Kou,[1] Shihao Yang,[2] Chia-Jung Chang,[3] Teck-Hua Ho,[4] and Lisa Graver[5]

[1]Department of Statistics, Science Center, Harvard University, Cambridge, Massachusetts, USA, [2]Harvard Medical School, Boston, Massachusetts, USA, [3]State Street Corporation, Boston, Massachusetts, USA, [4]National University of Singapore, Singapore, and [5]Alvogen Inc, Pine Brook, New Jersey, USA

(See the Editorial Commentary by Faust on pages 2952–4.)

This report presents a novel approach to estimate the total number of COVID-19 cases in the United States, including undocumented infections, by combining the Centers for Disease Control and Prevention's influenza-like illness surveillance data with aggregated prescription data. We estimated that the cumulative number of COVID-19 cases in the United States by 4 April 2020 was > 2.5 million.

Keywords. undocumented infection; total case count; influenza-like illness; aggregated prescription data; CDC's influenza-like illness report.

During the COVID-19 pandemic, many infections with mild to no symptoms are not reported due to various factors, including limited testing [1, 2]. There is a critical need to estimate the true scale of the pandemic for hot-spot detection, resource allocation, and intervention planning. Existing modeling approaches use epidemiology data [2] and digital technology/data [3–5] to estimate the scale of COVID-19.

In this report, we present a novel approach to estimate the total number of COVID-19 cases, including undocumented infections, in the United States (US) by comparing data from the US Centers for Disease Control and Prevention (CDC) Outpatient Influenza-like Illness Surveillance Network (ILINet), which targets all influenza-like illness (ILI), overlapping with COVID-19, against the aggregated prescription data of oseltamivir [6], which targets influenza only.

Our model shows that current official numbers are severely underestimated: We estimate that by the week ending 21 March 2020, there were > 1.3 million total COVID-19 infections in the US and that by the week ending 4 April 2020, there were > 2.5 million total infections in the US.

METHODS

The CDC defines ILI as "fever and a cough and/or a sore throat without a known cause other than influenza" [7], which covers the common symptoms of COVID-19. CDC generates weekly reports on the ILI level [8] and conducts laboratorial influenza virologic surveillance.

Prior to mid-February 2020, these 2 surveillance measures moved in the same direction. Since mid-February, however, the 2 measures have diverged, with the difference between ILI and laboratory-confirmed influenza activities attributable to COVID-19 [7, 8]. If we can obtain an accurate measure for influenza level, we can then use the difference between the reported ILI level and the estimated influenza level to estimate the level of new COVID-19 cases on a weekly basis.

We used aggregated weekly prescription data of oseltamivir, prescribed to treat influenza A and B but not COVID-19, to estimate the influenza level. Specifically, we used a linear model to calibrate the CDC-reported ILI level to the oseltamivir prescription data from January 2010 to mid-February 2020, and then produced estimates for influenza activity for mid-February to early April 2020 (Figure 1).

RESULTS

Our estimated influenza level (blue line) closely matches the CDC-reported ILI level (Figure 1, black line) (correlation 0.974) prior to mid-February 2020, but significant gaps between the 2 levels (Figure 1, red and black lines) emerge after mid-February, which can be attributed to COVID-19. For the week ending 21 March 2020, we estimated that 47% of the reported ILI level could be from COVID-19, which corresponds to approximately 855 000 new symptomatic cases in the US. As the official confirmed number of new cases was 17 450 for that week [9], this result shows that there were > 800 000 unreported symptomatic cases. The figure also shows that the cumulative number of COVID-19 symptomatic cases in the US by the week ending 28 March 2020 was estimated to be > 2 million and that the cumulative number of symptomatic cases in the US by the week ending 4 April 2020 was estimated to be > 2.5 million.

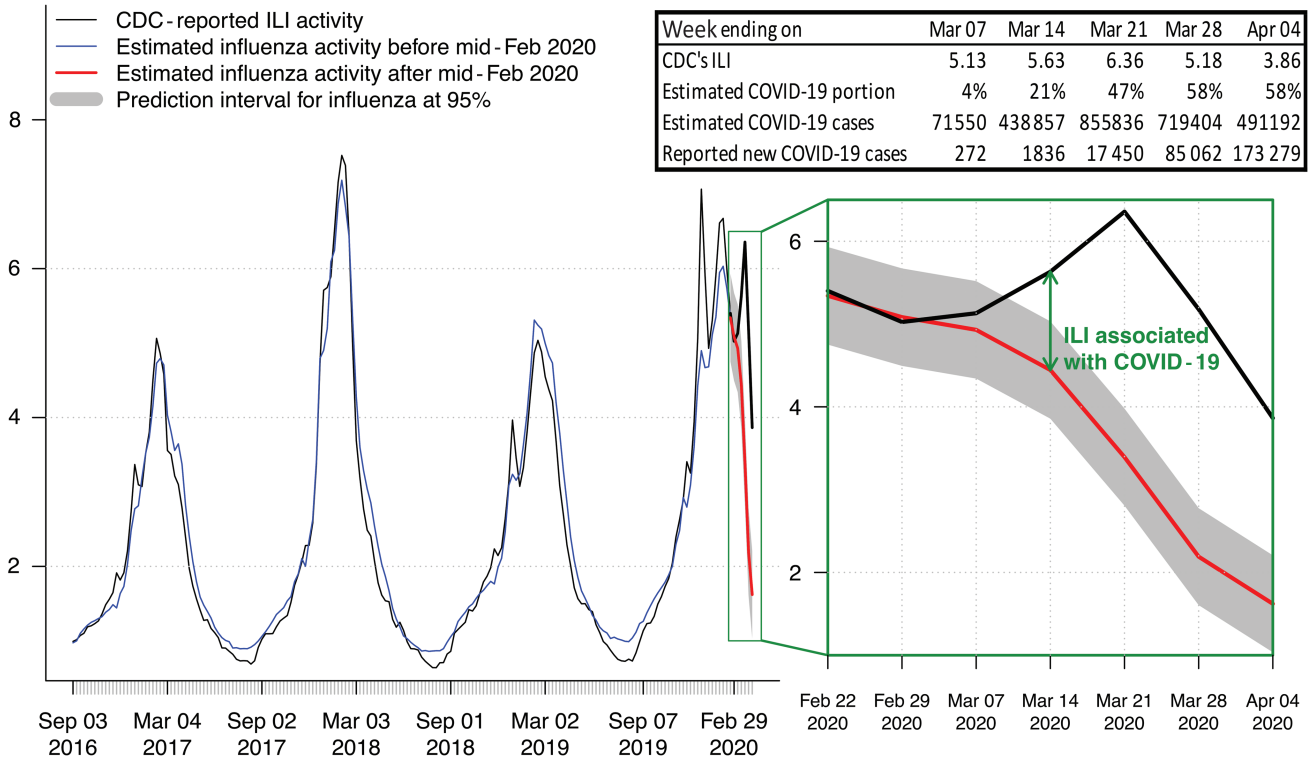
Our results show that the official numbers are severely underestimated, a conclusion that appears to be supported by a recent large-scale screening study covering > 6% of the Icelandic population [10] and another antibody survey study in Santa Clara County, California (although the study was

Received 22 April 2020; editorial decision 11 May 2020; accepted 12 May 2020; published online May 15, 2020.

Correspondence: S. C. Kou, Department of Statistics, 1 Oxford St, Cambridge, MA 02138, USA (kou@stat.harvard.edu).

Clinical Infectious Diseases® 2020;71(11):2949–51

DOI: 10.1093/cid/ciaa580



Week ending on	Mar 07	Mar 14	Mar 21	Mar 28	Apr 04
CDC's ILI	5.13	5.63	6.36	5.18	3.86
Estimated COVID-19 portion	4%	21%	47%	58%	58%
Estimated COVID-19 cases	71550	438857	855836	719404	491192
Reported new COVID-19 cases	272	1836	17 450	85 062	173 279

Figure 1. The estimated influenza level before and after mid-February 2020. Prior to mid-February 2020, our estimated influenza level (blue line) closely matches the Centers for Disease Control and Prevention (CDC)–reported influenza-like illness (ILI) level (black line), but significant gaps between the 2 levels (red and black lines) emerge after mid-February, which can be attributed to COVID-19. To estimate the COVID-19 weekly case counts shown in the figure, we used the ILI total counts reported in ILINet, the reported 8.5% sampling rate of ILINet, and the reported 50% ± 8% rate of persons with symptomatic ILI seeking medical care for their illness. For the reported rates, see https://www.cdc.gov/flu/about/burden/preliminary-in-season-estimates.htm and https://www.cdc.gov/flu/about/burden/how-cdc-estimates.htm.

cautioned for its design and potential sampling bias) [11]. Our study targeted symptomatic COVID-19 cases as we used the CDC-reported percentage of patients with symptomatic illness who would seek medical care in our estimation. Therefore, if we consider the substantial presymptomatic and asymptomatic cases revealed by the Icelandic study [10], the total number of COVID-19 infections in the US is likely to be even higher than our estimates.

DISCUSSION

Our estimation method is simple and intuitive. It contrasts the CDC-reported ILI level with the estimated influenza level from influenza-specific prescription data to obtain an estimate of the COVID-19 level. Our approach innovatively combined the traditional syndromic surveillance system with big data from pharmacy prescriptions. It provides a feasible solution for estimating unreported COVID-19 cases with mild symptoms.

One limitation of our model is that the estimate might become more conservative through time due to administrative/government interventions. Toward the start of April, the syndromic surveillance system ILINet got more and more affected by the changes in the healthcare system, including increased use of telemedicine, the recommendation to limit hospital visits to only severe illness, and tightened social distancing. These changes affect the total number of hospital visits, patients' inclination to seek outpatient healthcare, and doctors' medication prescription. Thus, our estimates in early to mid-March could be more accurate as these changes had not yet taken place, and our estimate would serve as a lower bound for the symptomatic cases of COVID-19 in later weeks.

Our study indicates the feasibility to estimate COVID-19 case count using multiple data sources. This approach can be used in conjunction with approaches utilizing digital data sources for COVID-19 case estimation [12, 13]. COVID-19 presents an unprecedented challenge. Conquering it requires unprecedented levels of collaboration and data sharing across government agencies, research institutes, and the private sector.

Note

Potential conflicts of interest. L. G. is directly employed by Alvogen. All other authors report no potential conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Nature. Coronavirus latest: confirmed cases cross the one-million mark. Nature **2020**. Available at: https://www.nature.com/articles/d41586-020-00154-w. Accessed 4 April 2020.

2. Li R, Pei S, Chen B, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science **2020**; 368:489–93.
3. McKendry RA, Rees G, Cox IJ, et al. Share mobile and social-media data to curb COVID-19. Nature **2020**; 580:29.
4. Buckee CO, Balsari S, Chan J, et al. Aggregated mobility data could help fight COVID-19. Science **2020**; 368:145–6.
5. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med **2020**; 542:125.
6. IQVIA. IQVIA National Prescription Audit™. April 2020.
7. Centers for Disease Control and Prevention. US Influenza Surveillance System: purpose and methods. **2020**. Available at: https://www.cdc.gov/flu/weekly/overview.htm. Accessed 14 April 2020.
8. Centers for Disease Control and Prevention. US Influenza Surveillance System: past weekly surveillance reports. **2020**. Available at: https://www.cdc.gov/flu/weekly/pastreports.htm. Accessed 14 April 2020.
9. European Centre for Disease Prevention and Control. Situation update worldwide, as of 6 April 2020. Available at: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. Accessed 6 April 2020.
10. Gudbjartsson DF, Helgason A, Jonsson H, et al. Spread of SARS-CoV-2 in the Icelandic population. N Engl J Med **2020**. doi:10.1056/NEJMoa2006100.
11. Bendavid E, Mulaney B, Sood N, et al. COVID-19 antibody seroprevalence in Santa Clara County, California. medRxiv [Preprint]. Posted 30 April 2020. Available at: https://dx.doi.org/10.1101/2020.04.14.20062463. Accessed 9 May 2020.
12. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci U S A **2015**; 112:14473–8.
13. Stephens-Davidowitz S. Google searches can help us find emerging Covid-19 outbreaks. **2020**. Available at: https://www.nytimes.com/2020/04/05/opinion/coronavirus-google-searches.html. Accessed 21 April 2020.