The Data Science of COVID-19 Spread: Some Troubling Current and Future Trends

Rex Douglass [*], Thomas Leo Scherer , Erik Gartzke

1 University of California, San Diego, La Jolla,CA, 92093

* Corresponding author: rexdouglass@gmail.com

Abstract: One of the main ways we try to understand the COVID-19 pandemic is through time series cross section counts of cases and deaths. Observational studies based on these kinds of data have concrete and well known methodological issues that suggest significant caution for both consumers and produces of COVID-19 knowledge. We briefly enumerate some of these issues in the areas of measurement, inference, and interpretation.

Introduction

The SARS-COV-2 global pandemic has exposed weaknesses throughout our institutions, and the sciences are no exception. Given the deluge of official statistics and 300+ new COVID-19 working papers posted each day[1], it is imperative for both consumers and producers of COVID-19 knowledge to be clear on what we do and do not know. In this brief review, we enumerate ways that data science has highlighted these weaknesses and is helping to address them.

In terms of understanding where we are, how we got here, and what is likely to follow, here are some things we need to know. We need to know the rate of spread of COVID-19 in a population R, over time R_t, across different political and demographic communities R_{ct}, and prior to any non-pharmaceutical interventions R_{c0}. We need to know how many cases of active infection exist in a community I_{ct} and how many of those infections resulted in death D_{ct}. We need to know the causal effect of interventions X_{ct} on say rate of spread, between the observed treated populations R_{ct1} and the counterfactual populations had they not been treated R_{ct0}. To do so, we need some plausible causal identification strategy that allows us to account for the fact that interventions are themselves chosen and implemented in response to changes in R_{ct}, and that many outside factors likely drive both R_{ct} and X_{ct} simultaneously. These unknowns give rise to fundamental problems of measurement, inference, and interpretation.

Measurement

For the first several months of the pandemic and still in most countries now, there is no direct measure of I_{ct}. Very few countries have implemented an ideal regularly timed national survey like the U.K.'s Office for National Statistics COVID-19 infection survey (Pouwels et al. 2020). More typically, we are reliant on serological estimates of Cumulative Infections CI_{ct} that measure the presence of antibodies indicative of infection at some point in the past. These are also still rare, and they have false positive

[1] "COVID-19 Primer." Accessed August 17, 2020. https://covid19primer.com/.

rates that make them inappropriate for populations with low infection rates (Peeling et al. 2020).

More commonly available are Confirmed Cases CC_{ct}. COVID-19 tests are administered in a jurisdiction, and positive results are anonymized, tabulated, reported, and aggregated by increasingly nested bureaucracies. These bureaucracies are concerned primarily with releasing legally required contemporary measurements and not maintaining consistent historical time series. This has resulted in the world's largest, most desperate scavenger hunt to scrape, transcribe, and translate counts disseminated in oral briefings, public websites, PDFs, and even static images (Alamo et al. 2020). Teams from every country are working in often uncoordinated and duplicated efforts to compile government reporting into consistent panel data; these teams include newspapers (Sun et al. 2020), nonprofits (USAFacts 2020), large private companies (Zhang, Donthini, and Source 2020; Wolf, Ary, and Firooz 2020), consortiums of volunteers (Yang et al. 2020; Zohrab et al. 2020; Group 2020), and Wikipedians.

The resulting ecosystem of panel datasets vary in spatial and temporal coverage, have little metadata about sources or changing definitions, and generally do not handle revisions to past counts from reporting sources. Direct comparisons between sources reveal worrying disagreements and temporal artifacts like reporting delays, seasonalities, discontinuities, and sudden revisions in counts both upwards and downwards (Wang et al. 2020). It is not obvious how to correctly account for these problems or adjudicate between conflicting sources without a clear ground truth. There also is no permanent archive of the raw source material meaning reconstructing the full chain of evidence may no longer be possible.

Likewise, we do not have direct measures of Deaths D_{ct} but only Confirmed Deaths CD_{ct}. CD_{ct} suffers from all of the problems of Confirmed Cases CC_{ct} except for possibly less under-reporting depending on if the person died at home or in medical care. Choosing CD_{ct} as the lesser of two evils, many projects attempt to take plausible values of the Infected Fatality Rate $IFR = D/I$ to back out an estimation for I_{ct} (Meyerowitz-Katz and Merone 2020). Others have turned to estimating Excess Deaths ED_{ct}, which is a number proportional to the number of total deaths reported in an area above what would be expected given the number of deaths reported in previous years (Weinberger et al. 2020). ED_{ct} is also not a direct estimate of D_{ct} as it can include deaths that were not caused by COVID-19 directly, e.g. other health conditions that received inadequate care during this period, and similarly can undercount the number of COVID-19 caused deaths as lockdowns reduce mobility and economic activity that might typically lead to deaths, e.g. car accidents.

Confirmed case and death counts mechanically depend on testing, but records of tests administered T_{ct} are even worse. In the U.S., much of what we know about trends in testing patterns come from journalistic efforts like the Covid-Tracking Project (Lipton et al. 2020). They encountered all of the regular problems plus additional ones specific to ambiguity to what kind of test count is being reported (testing encounters, number of people tested, number of swabs tested, etc). The type of test performed (and its false positive and false negative rate) is almost never included as metadata. Nor are the rules about how tests are being rationed and distributed being recorded systematically.

The general failure to track COVID-19 spread directly has led to a proliferation of innovative attempts to use other signals such as web searches, searches of medical databases, social media posts, fevers reported by home thermometer, and traditional flu symptom survailence networks (Kogan et al. 2020). While promising, proxy measures require ground truthing and regular calibration using something like regularly timed serological surveys on smaller geographic samples of the population. It is precisely the lack of such capabilities that are motivating the search for alternatives in the first place.

Finally, non-pharmaceutical interventions are tracked by several academic and

nonprofit teams (Hale et al. 2020; Cheng et al. 2020). These interventions are intended to limit human mobility which is more directly measured by cell phone data which are being provided by companies like Google, Apple, and SafeGraph.

Inference

The workhorse theoretical model for infectious disease spread is the Susceptible, Exposed, Infectious, and Removed (SEIR) compartmental model (Brauer and Castillo-Chavez 2012). The intuition behind the SEIR model is that there are mechanical relationships, such as previous infections or deaths removing candidates from infection, the timing between exposure to the next possible transmission, and the degree to which immunity may exist in the population, which induce nonlinearities in disease spread. Disease spreads slowly at first, accelerates, and then burns out if left to its own devices. SEIR should be considered the theoretical floor for analysis, and an entire menagerie of extensions account for demography, testing, mobility, social networks, etc.

The necessity of directly including testing in models of disease spread can't be understated. Per capita cases are so temporarily correlated with per capita testing rates they are more of a proxy of testing availability than infections (Kaashoek and Santillana 2020). Spatially, per capita testing rates correlate with urbanity and a wide range of co-morbids (Souch and Cossman 2020). How many tests are given and to who varies systematically in response to conditions on the ground with both periods of rationing and blitzes.

Measuring the effect of interventions is difficult because they are assigned endogenously in response to both local conditions and national signals. Similarly, populations responded to both government orders and local conditions, often reducing their activity prior to being ordered to and also increasing their activity prior to being officially allowed to. Governments, the public, and the disease are all responding simultaneously to each other in often nonlinear and unobserved ways. Statistical instruments that cause government interventions but do not directly cause testing rates or rate of spread except through the government intervention are few and far between. Further, interventions are often implemented simultaneously or in a rolling cumulative pattern directly in response to changes in cases and testing results, making isolating the effect of any one treatment exceptionally difficult.

Even if we had an exogenous intervention, its treatment effect on the rate of spread is still unlikely to be identified since almost any intervention will affect both cases and testing. Estimating an effect on just spread requires imposing additional assumptions, e.g. sharp constraints on some parameters and informative priors on the relationship between the number of tests and the number of cases (Kubinec and Carvalho 2020).

Interpretation

One promising development is rigorous forecast evaluations (Reich et al. 2020). Notoriously, many early simple growth models fit to the takeoff period of infections performed well right up until the curve broke and then failed entirely. A parade of predicted peaks in cases since continue the tradition, with groups celebrating success on uninteresting short-term autocorrelations while ignoring failures on actually interesting shifts in trends. All we can do is develop a very long memory of predictions and constantly hold models accountable for their long run out-of-sample performance on unseen future data.

Other trends in initial COVID-19 work and reporting are less promising. Especially concerning is observational work that presents correlations as evidence of causation.

Without identification, correlations on short highly autocorrelated time series are as likely to be misleading as informative. The SEIR model expects a nonlinear and highly autocorrelated pattern of an increasing infection rate that then levels off independent of any interventions. An unscrupulous, or naive, analyst can easily find interventions that increased (or decreased) spread solely by where those interventions land in the natural disease cycle, completely independent of the intervention's actual effect.

Another concern is the pursuit of statistically distinguishable correlations over actually attempting to explain variation in COVID-19 outcomes themselves. Papers that can show a particular political party or demographic group is 'worse' on some COVID-19 dimension receive much attention. Such results lack strong explanatory power or clear policy recommendations, and so while great for making headlines, they do little to help us end the current pandemic.

Perhaps our greatest concern is the desire to setup straw man null hypotheses and then presenting the inability to reject them as positive evidence for medical and safety decisions, e.g. arguing that social distancing might not be required because a model was unable to statistically distinguish a large uptick in cases following a specific mass-meeting. In the best of circumstances, absence of evidence is not evidence of absence. Our underfit, undertheorized, and underperforming observational models are not the best of circumstances, and they are not sufficiently sensitive to evaluate more than macro-level general trends.

Conclusion

This necessarily brief review omitted positive developments in studying COVID-19 outside of macro-observational settings. There has been remarkable progress in areas of diagnosis, clinical treatment, and phylogenetic tracking. Data science has contributed to the rapid collaboration, development, and dissemination of research in a way not seen in prior disease outbreaks. We also neglected topics like tracing, and the accompanying contributions from the tech field such as monitoring through mobile apps and social media. Further, our review is overly U.S.-centric, with other countries like South Korea monitoring the disease so effectively they succeeded at containment without having to resort to difficult mitigation.

Any policy prescription toward COVID-19 should be viewed with a healthy respect for how little we actually know about the history of this pandemic. Practioners working on these questions and with these data will be deeply familiar with many of these concerns, but some may be especially subtle or less prominantly discussed within one's main field of study. At a minimum, there is research being produced today which ignores many of these known methodological problems and subsequently generates confusion for novice consumers of analysis. We hope that this partial enumeration of challenges in COVID-19 measurement, inference, and interpretation is compelling \ul{justification for intellectual caution among consumers and producers of COVID-19 knowledge alike.

Acknowledgments

Our thanks to the Center for Peace and Security Studies and its members, and to the Office of Naval Research [N00014-19-1-2491] and the Charles Koch Foundation for financial support. Thank you to all who provided feedback on the early draft.

Author contributions: Conceptualization, R.W.D., T.L.S., and E.G.; Investigation, R.W.D.; Writing - Original Draft, R.W.D.; Writing - Review & Editing, R.W.D. and T.L.S.; Funding - E.G.

References

Alamo, Teodoro, Daniel G. Reina, Martina Mammarella, and Alberto Abella. 2020. "Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic." *Electronics* 9 (5). Multidisciplinary Digital Publishing Institute: 827. https://doi.org/10.3390/electronics9050827.

Brauer, Fred, and Carlos Castillo-Chavez. 2012. *Mathematical Models in Population Biology and Epidemiology*. Vol. 2. Springer.

Cheng, Cindy, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. "COVID-19 Government Response Event Dataset (CoronaNet V.1.0)." *Nature Human Behaviour* 4 (7). Nature Publishing Group: 756–68. https://doi.org/10.1038/s41562-020-0909-7.

Group, COVID-19 India Org Data Operations. 2020. "Dataset for Tracking COVID-19 Spread in India." Accessed on yyyy-mm-dd from https://api.covid19india.org/.

Hale, Thomas, Anna Petherick, Toby Phillips, and Samuel Webster. 2020. "Variation in Government Responses to COVID-19." *Blavatnik School of Government Working Paper* 31.

Kaashoek, Justin, and Mauricio Santillana. 2020. "COVID-19 Positive Cases, Evidence on the Time Evolution of the Epidemic or an Indicator of Local Testing Capabilities? A Case Study in the United States." SSRN Scholarly Paper ID 3574849. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3574849.

Kogan, Nicole E., Leonardo Clemente, Parker Liautaud, Justin Kaashoek, Nicholas B. Link, Andre T. Nguyen, Fred S. Lu, et al. 2020. "An Early Warning Approach to Monitor COVID-19 Activity with Multiple Digital Traces in Near Real-Time." *arXiv:2007.00756 [Q-Bio, Stat]*, July. http://arxiv.org/abs/2007.00756.

Kubinec, Robert, and Luiz Carvalho. 2020. "A Retrospective Bayesian Model for Measuring Covariate Effects on Observed COVID-19 Test and Case Counts," April. SocArXiv. https://doi.org/10.31235/osf.io/jp4wk.

Lipton, Zach, Josh Ellington, smike, James Ouyang, Ken Riley, Joshua Ellinger, Jeff Hammerbacher, Olivier Lacan, Jason Crane, and space-buzzer. 2020. "The Covid-Tracking Project." Zenodo. https://doi.org/10.5281/zenodo.3981599.

Meyerowitz-Katz, Gideon, and Lea Merone. 2020. "A Systematic Review and Meta-Analysis of Published Research Data on COVID-19 Infection-Fatality Rates." *medRxiv*, May. Cold Spring Harbor Laboratory Press, 2020.05.03.20089854. https://doi.org/10.1101/2020.05.03.20089854.

Peeling, Rosanna W., Catherine J. Wedderburn, Patricia J. Garcia, Debrah Boeras, Noah Fongwen, John Nkengasong, Amadou Sall, Amilcar Tanuri, and David L. Heymann. 2020. "Serology Testing in the COVID-19 Pandemic Response." *The Lancet Infectious Diseases* 0 (0). Elsevier. https://doi.org/10.1016/S1473-3099(20)30517-X.

Pouwels, Koen B., Thomas House, Julie V. Robotham, Paul Birrell, Andrew B. Gelman, Nikola Bowers, Ian Boreham, et al. 2020. "Community Prevalence of SARS-CoV-2 in England: Results from the ONS Coronavirus Infection Survey Pilot." *medRxiv*, July. Cold Spring Harbor Laboratory Press, 2020.07.06.20147348. https://doi.org/10.1101/2020.07.06.20147348.

Reich, Nicholas G, Jarad Niemi, Katie House, Abdul Hannan, Estee Cramer, Steve Horstman, Shanghong Xie, et al. 2020. "Reichlab/Covid19-Forecast-Hub: Pre-Publication Snapshot." Zenodo. https://doi.org/10.5281/zenodo.3963372.

Souch, Jacob M., and Jeralynn S. Cossman. 2020. "A Commentary on Rural-Urban Disparities in COVID-19 Testing Rates Per 100,000 and Risk Factors." *The Journal of Rural Health*, June, jrh.12450. https://doi.org/10.1111/jrh.12450.

Sun, Albert, Tiff Fehr, Archie Tse, Rachel, and Wilson Andrews. 2020. "New York Times Coronavirus (Covid-19) Data in the United States." Zenodo. https://doi.org/10.5281/zenodo.3981451.

USAFacts. 2020. "US Coronavirus Cases and Deaths." Zenodo. https://doi.org/10.5281/zenodo.3981486.

Wang, Guannan, Zhiling Gu, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, Lei Gao, and Li Wang. 2020. "Comparing and Integrating US COVID-19 Daily Data from Multiple Sources: A County-Level Dataset with Local Characteristics." *arXiv:2006.01333 [Stat]*, June. http://arxiv.org/abs/2006.01333.

Weinberger, Daniel M., Jenny Chen, Ted Cohen, Forrest W. Crawford, Farzad Mostashari, Don Olson, Virginia E. Pitzer, et al. 2020. "Estimation of Excess Deaths Associated with the COVID-19 Pandemic in the United States, March to May 2020." *JAMA Internal Medicine*, July. https://doi.org/10.1001/jamainternmed.2020.3391.

Wolf, Ashley, Asaf Ary, and Hossein Firooz. 2020. "Yahoo Knowledge Graph COVID-19 Datasets." Zenodo. https://doi.org/10.5281/zenodo.3981432.

Yang, Tong, Kai Shen, Sixuan He, Enyu Li, Peter Sun, Pingying Chen, Lin Zuo, et al. 2020. "CovidNet: To Bring Data Transparency in the Era of COVID-19." *arXiv:2005.10948 [Cs, Q-Bio]*, July. http://arxiv.org/abs/2005.10948.

Zhang, Chiqun, Chaitanya Donthini, and Microsoft Open Source. 2020. "Bing-COVID-19-Data." Zenodo. https://doi.org/10.5281/zenodo.3978733.

Zohrab, J, Ryan Block, Cameron Chamberlain, Larry Davis, Minh Nguyeñˆ, Alastair Gifillan, Adam Hughes, BriceWolfgang, and andys1376. 2020. "COVID Atlas Li." Zenodo. https://doi.org/10.5281/zenodo.3981563.