Journal pre-proof

DOI: 10.1016/j.patter.2020.100022

COVID-19 is a Data Science Issue

COVID-19 is far more than just a data science issue, it's a massive public health problem that has resulted in many deaths, and is throwing a harsh light onto how we structure our society when it comes to important things like the availability and affordability of healthcare, worker's rights, and even freedom of movement.

But as a data scientist, I do think it's important to look at the situation through a data science perspective. We've all seen curves on Twitter —exponential, flattened, and otherwise—plotted in Excel, and been reassured by them or scared by them or wondered if we could trust them. That's a data science question, and there are many others like that I want to address here, in the hopes that what I write will inspire others to think about the data, and feel more empowered about what to do in this situation.

Data collection and interpretation

Data collection when it comes to infectious diseases is difficult at the best of times. The rise of Big Data has provided clinicians and researchers with the systems and ability to store and work with large amounts of data, but in public health critical surveillance systems remain primarily based on manually collected and coded data, which are slow to collect and difficult to disseminate. Traditional health surveillance systems are notorious for severe time lags and lack of spatial resolution, and our current situation has demonstrated clearly that systems that are robust, local, and timely are thus critically needed (Bansal, et al, 2016).

Data scientists working in public health can learn from their colleagues in other domains where real-time data acquisition and analysis of high resolution data is common. For a disease like coronavirus, where the majority of infections are mild, and are therefore self-treated, relying on hospital and general practitioner records to estimate the spread can be misleading in the early stages of the disease progression. Reporting tends to focus on morbidity and mortality, and it's easier to count the people who have actually presented themselves at health facilities for testing or care.

A key fact for us all to remember is that, for the majority of countries, we're not actually counting how many people *have* the virus – instead were counting the *reports* of how many people have the virus, and, like all metrics, those numbers vary according to how they're measured. An increase in the number of tests being carried out will result in an increase in the number of infections detected.

As of the date of writing, only Iceland has done systematic sampling of a broad enough sample of the population, including those that have not shown symptoms, to allow us to get a handle on how many people are asymptomatic [https://www.government.is/news/article/2020/03/15/Large-scale-testing-of-general-population-in-Iceland-underway/]. This shows not only the general prevalence of the virus in the general population, but further testing will give us an understanding of how the virus spreads and how well the containment techniques (social distancing, etc.) are working.

Leaving aside any conspiracy theories about governmental coverups, the simple fact is that testing for coronavirus is expensive. This means that the numbers collected in a given country will be influenced by not only how widespread the virus actually is, but also the financial ability

for the local health care facilities to give the test to everyone who presents with concerns that they're infected.

This is a classic data sampling problem, and something that data scientists can explain in ways that can ease the concerns of the general public about the increasing number of cases, while at the same time working with health professionals to better understand the spread and distribution of cases.

Collecting accurate data and understanding the limitations of the data that's already been collected, is an essential part of understanding the situation. Without good data, policymakers can't make good decisions. Data scientists can help with this.

Data modelling and prediction

Once we have data, the questions then become: what will happen next? How will the virus spread? What will happen to the spread if certain non-pharmaceutical interventions are put in place? How effective is social distancing in comparison with country-wide quarantine? What are the longer-term impacts of (for example) shutting schools for the next month, or longer? How can we track the spread of virus through our understanding of social networks and human behaviour? Is the risk of catching the virus as high if you watch the football match in the pub as compared with the stadium?

Data scientists, in collaboration with other researchers, are uniquely placed to be able to find answers to these questions. Work has already been published on this topic (Kucharski et al, 2020), but as the situation develops over time and different administrations have different responses, there will be divergence between what's modelled and what occurs. The simplistic assumption of an exponential growth curve at the beginning of the pandemic, though appealing for its drama (https://www.statnews.com/2020/03/10/simple-math-alarming-answers-covid-19/), will rapidly diverge with the reality of the situation, which is why continual monitoring is required.

More sophisticated modelling, such as done by the Imperial College COVID-19 Response Team [https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf], gives a more nuanced approach and has been very influential on government policy (in the UK), though it sacrifices the double checking process of peer review in favour of getting the information out quickly (although post publication peer review has occurred https://necsi.edu/review-of-ferguson-et-al-impact-of-non-pharmaceutical-interventions) . Other articles, much shared on social media [https://medium.com/@tomaspueyo/coronavirus-the-hammer-and-the-dance-be9337092b56] use data modelling to support an argument for specific public health policies.

It is a truth universally acknowledged that all models are wrong, but some are useful. Data science is needed to not only develop the models, but also to determine in which ways they're wrong, and which ways they're useful, because the results of these models will inform, along with data, the decisions that are made to combat the spread of this pandemic.

I would urge all data scientists wishing to help in these modelling efforts to not just simply grab the data and plug it into their preferred analysis software. The numbers that result can be terrifying, especially without the domain specific knowledge that epidemiologists have to put it all into context. Instead I'd encourage joining in with the Kaggle COVID-19 Open Research Dataset Challenge (CORD-19) [https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge] where we can all work together as a team, and play to our respective strengths.

Data visualisation and communication

Infographics and data visualisations are a useful and helpful way of putting risks and raw numbers into perspective. Data scientists have the experience and understanding to be able to accurately and helpfully put this information into a context that is visually appealing and yet easy to understand. Data scientists can also create interactive and continually updated information sources which draw from the latest data, thereby ensuring that everyone is kept up to date with the latest numbers.

Making the right information easy and appealing to share is crucial in the current climate where most people get the bulk of their news from social media. Conversely, visualizations have the potential to incite fear and alarm just as much as they have the potential to inform (https://medium.com/nightingale/ten-considerations-before-you-create-another-chart-about-covid-19-27d3bd691be8). Please bear in mind that a substantial proportion of the general population are not mathematically inclined (up to and including having dyscalculia) and so therefore what seems obvious to you as a data scientist might not be to others.

The hashtag #FlattenTheCurve and its associated schematic graph/gif have been widespread on social media (https://www.fastcompany.com/90476143/the-story-behind-flatten-the-curve-the-defining-chart-of-the-coronavirus). It's an appealingly packaged message that is easy to share, quick to understand, and one that gives people ownership of the situation. If we take basic precautions, like washing our hands, we can help slow the growth rate so that it doesn't overwhelm our health services. This is a powerful and important message to send.

Other excellent example of simple, but effective visualisations can be found at https://informationisbeasutiful.net/visualizations/covid-19-coronavirus-infographic-datapack/ and also at John Hopkins University Coronavirus Resource Center https://coronavirus.jhu.edu/map.html. Interactive calculators such as the one by Gabriel Goh [http://gabgoh.github.io/COVID/index.html] are also useful, but can be alarming for members of the public with little understanding of data science or epidemiology, or little willingness to dive into the citations and underlying assumptions.

Similarly, the WHO's coronavirus situation reports are released on a daily basis, and are quick to read, with the main information summarised in a clear and accessible fashion. Standardising the production and reporting of these numbers is a data science task, as is producing and defining them (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports).

After it all ends

None of us can accurately and definitively predict the outcomes of this pandemic, the total infection rates or the final death toll, at this point in time. These are difficult times for us all, because of the worry and the uncertainty of the situation. We simply don't know what will happen, but with data science, we have a better chance of predicting it accurately than we ever have had in the past.

But even once the pandemic is over, there will still be data science work to do (Perakslis, 2020). The data collection may stop because there are no new cases, but that's the time to make sure that what was collected is stored and managed properly. Looking back at the situation with the benefit of hindsight can be painful, but it's the only way we can ensure that lessons learned during this time are fully understood. After it all ends is also the time to look at the systems and structures (medical, scientific, social) that worked, and the ones that didn't, with the view of improvement.

We will also need to take stock of some of the measures that were employed to deal with the situation. Yes, in times of emergency, tracking infected persons via their cell phones may well be the sensible thing to do to contain the transmission of the virus. But once the immediacy of that situation is over, we will need to ask ourselves if there is a way that could have been done that

didn't impinge on personal privacy, or the other rights we hold so dear. What are we, as a society, prepared to accept, in order to ensure our health and safety? And what can data science do to raise our societal health, while minimising the effects on our rights?

Remember the people behind the numbers

Data science deals with numbers, statistics, curves and distributions. We do this because it's easier to work with numbers on a population scale, and our tools work best when fed with large amounts of data. This is alright.

If I have one plea, it is that all of us, data scientists or not, remember that behind those numbers are human lives. Real people who are worried and afraid for themselves or for their loved ones. It's easy enough to calculate a mortality curve that shows that it is those who are 60+ years old, or suffer from pre-existing conditions, are the most likely to die. It's another thing to realise that those figures show that it's your aged parents or grandparents, or immuno-compromised friends and family, who are most at risk.

We live in challenging times, but with the right data, and the right science, we can make a difference.

Sarah Callaghan,

25 March 2020

References

Shweta Bansal, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, Cécile Viboud, Big Data for Infectious Disease Surveillance and Modeling, *The Journal of Infectious Diseases*, Volume 214, Issue suppl_4, December 2016, Pages S375–S379, https://doi.org/10.1093/infdis/jiw400

Kucharski, Adam JSun, Fiona et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study, The Lancet Infectious Diseases, Volume 0, Issue 0 Published:March 11, 2020DOI:https://doi.org/10.1016/S1473-3099(20)30144-4,

Perakslis, 2020, A Primer on Biodefense Data Science for Pandemic Preparedness, Patterns