Title: COVID-19 infection and death rates: the need to incorporate causal explanations for the data and avoid bias in testing

Norman E Fenton[1], Martin Neil[1], Magda Osman[2], Scott McLachlan[1,3].

[1]*Risk and Information Management, Queen Mary University of London, London, United Kingdom and Agena Ltd, Cambridge, UK*

[2]*School of Biological and Chemical Sciences, Queen Mary University of London, United Kingdom*

[3]*Health informatics and Knowledge Engineering Research Group (HiKER)*

{n.fenton, m.neil, m.osman, s.mclachlan}@qmul.ac.uk

Corresponding author: *n.fenton@qmul.ac.uk*

Abstract

COVID-19 testing strategies are primarily driven by medical need - focusing on people already hospitalized with significant symptoms or on people most at risk. However, such testing is highly biased because it fails to identify the extent to which COVID-19 is present in people with mild or no symptoms. If we wish to understand the true rate of COVID-19 infection and death, we need to take full account of the causal explanations for the resulting data to avoid highly misleading conclusions about infection and death rates. We describe how causal (Bayesian network) models can provide such explanations and the need to combine these with more random testing in order to achieve reliable data and predictions for the both policy makers and the public.

Misleading death rates

Suppose we wanted to estimate how many car owners there are in the UK and how many of those own a Ford Fiesta, but we only have sampled data on those people who visited Ford Car Showrooms in the last year. If 9% of the showroom visitors owned a Fiesta then, because of this selection bias in the sampled data, this would certainly overestimate the proportion of Ford Fiesta owners in the country. Estimating death rates for people with COVID-19 is currently undertaken largely along the same lines.

Take the UK as an example, here all testing of COVID-19 is performed on people already hospitalized with COVID-19 symptoms. At the time of writing[1], there were - according to the official NHS reporting figures - 33,722 confirmed COVID-19 cases (analogous to car owners visiting a showroom) of whom 2,921 have died (Ford Fiesta owners who visited a showroom). Concluding that the death rate from COVID-19 is on average 9% (2,921 out of 33,722) ignores the many people with COVID-19 who are not hospitalized and have not died (analogous to car owners who did not visit a Ford showroom and who do not own a Ford Fiesta). It is therefore equivalent to making the mistake of concluding that 9% of all car owners own a Ford Fiesta.

There are many prominent examples of this sort of erroneous conclusion. The Oxford COVID-19 Evidence Service (Oke & Heneghan, 2020) have undertaken a thorough statistical analysis. They acknowledge potential selection bias, but for them 'uncertainty' takes the form of confidence bounds around the (potentially highly misleading) proportion of deaths among confirmed COVID-19 patients. They also note various factors that can result in wide national differences, such as different demographic factors and differences in the way deaths are reported. The latter clearly may be a critical factor in explaining why a country like the UK's

[1] https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd

9% (mean) ‘death rate’ is so high compared to Germany’s 0.74%. We know that in the UK currently everybody who dies ***with*** COVID-19 is recorded as a COVID-19 death, even if the disease was not the actual cause of death; but it is also possible that people may die from the virus without actually having been diagnosed with COVID-19 (Henriques, 2020).

The fact that all the uncertainty around the death rates remains l anchored around one statistic, reported deaths per confirmed cases, is awkward. It fails to incorporate explicit causal explanations that might enable us to make more meaningful inferences from the available data, including data on virus testing. To do this we need explicit causal/graphical models (Pearl & Mackenzie, 2018).

The need for a causal model

Figure 1 is an example of a causal model, represented by a graph (called a Bayesian network), that might be applicable to any given country and its population. It shows that the COVID-19 death rate is as much a function of sampling methods, testing and reporting, as it is determined by the underlying rate of infection in a vulnerable population.

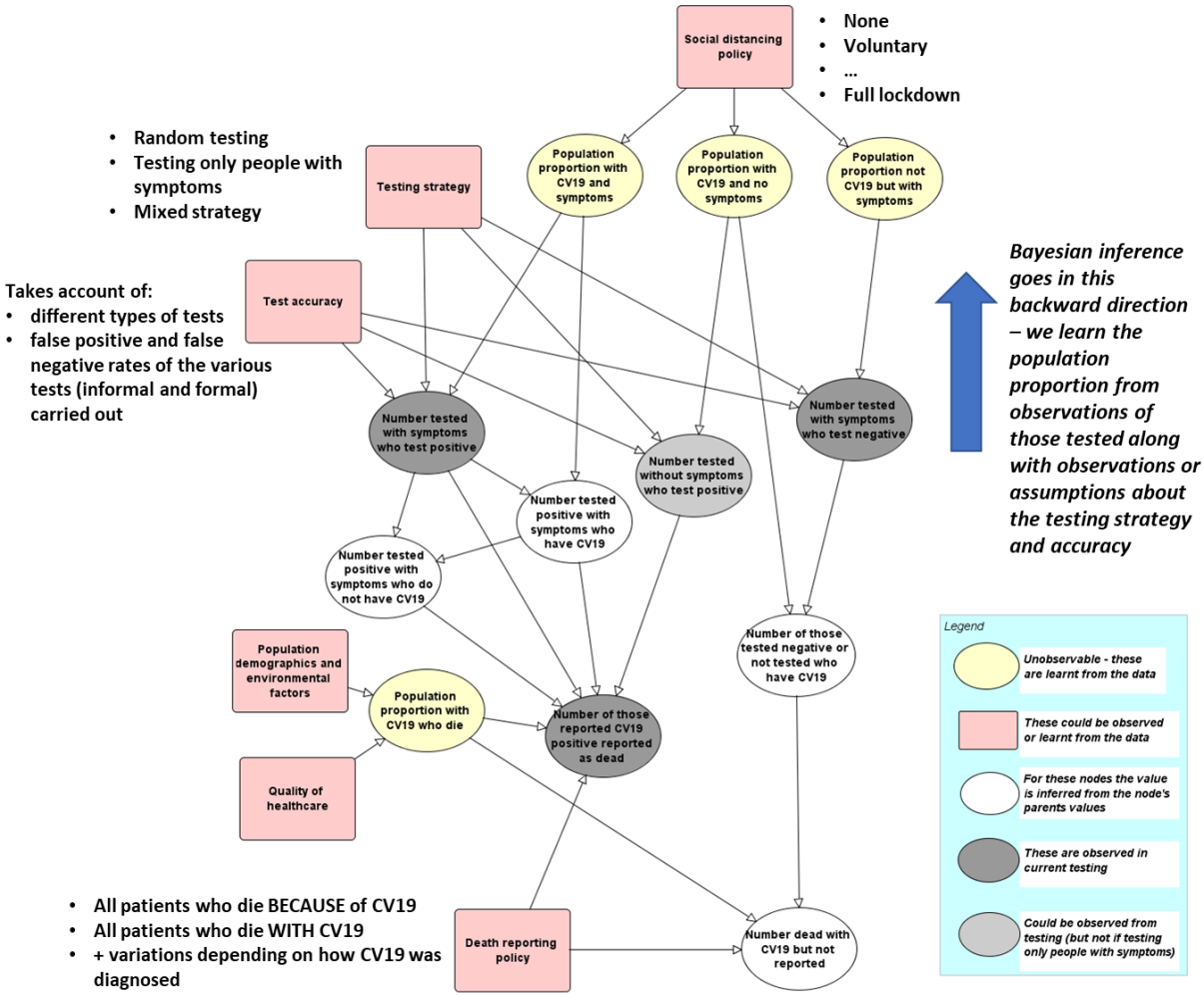

Figure 1 Causal Bayesian network model for learning population COVID-19 infected and death rates. This is a single time slice. It is can be updated daily and so enables us also to learn the true rate of spread of the infection. 'With symptoms' means serious symptoms.

The links between the variables in the model (nodes in graph) show how they are dependent on each other. For example, the "population proportion with COVID-19 who die" is dependent on "population demographics and environmental factors" as well as "Quality of healthcare" (which might cover factors such as intensive care unit capacity etc). Note that the model carefully distinguishes between whether or not a person really has COVID-19 and whether the person is ***classified*** as having COVID-19 (i.e. the model takes account of the potential for false positive and false negative test results). So, for example, the variable "Number of reported COVID-19 positive reported as died" (which is the official figure of 'COVID-19 deaths

reported) is dependent on the actual “Population proportion with COVID-19 who die”, the “Death Reporting Policy” as well as several other variables including “Number tested positive with symptoms who do not have COVID-19”.

The strength of each dependency, as well as the uncertainty associated with these is captured using probabilities and statistical distributions. When observed data is entered into the model for specific variables that are subsequently observed all of the probabilities for, as yet, unknown variable are updated using an AI algorithm called Bayesian inference. The model in Figure 1 is more accurately called a (causal) Bayesian network (Fenton & Neil, 2018), because it explains the underlying process by which the observed data might be generated. We have developed such models for many similar problems and are currently gathering the data needed to determine this causal model.

Therefore, while clinical, demographic and environmental factors can lead to genuine differences in death rates shown across different countries, very large differences may be likely to be caused by the application of different sampling strategies and reporting policies (Binnicker, 2020; FindX, 2020) and not necessarily because they are managing the virus any better or that the virus has infected fewer or more people. With a causal model that explains the data generating process we might better account for these differences between countries and more accurately learn the underlying true population infection and death rates from the observed data.

The need for more random testing

In the absence of community-wide testing, only random testing, applied throughout the population, will help us learn the number of people with COVID-19 who are asymptomatic or have already recovered, and hence also estimate the underlying infection and death rates. It will also help us learn the accuracy of the testing undertaken (false positive, and false negative

rates). Random testing remains the most effective strategy to avoid selection bias and reduce the distortions in reported statistics, but it also needs to be combined with a causal model in order to better determine the prevalence, severity, and ultimately societal impact of COVID-19.

Currently it seems there are no state-wide protocols in place in any country for randomised community testing of citizens for COVID-19. Spain did attempt it, but that involved purchasing large volumes of rapid COVID-19 tests, and they soon discovered that some Chinese-sourced tests had poor validity and reliability delivering only 30% accuracy – resulting in high numbers of false positives. Countries like Norway have proposed introducing such tests, but there is uncertainty around how to legislatively compel citizens to test – and what might constitute an appropriate randomisation protocol. In Iceland, they have voluntary sampling which has covered 3% of the population, but this isn't random. Some countries with large scale testing, like South Korea, might get closer to being random.

The reason it is so hard to achieve random testing is that you have to account for several practical and psychological factors. How does one collect samples randomly? Gathering samples from volunteers may not be sufficient as it does not prevent self-selection bias.

During the H1N1 influenza pandemic of 2009–2010, there was a lot of anxiety about the disease that created "mass psychogenic illness" (Wheaton, Abramowitz, Berman, Fabricant, & Olatunji, 2012). This is when hypersensitivity to particular symptoms leads to healthy people self-diagnosing as having a virus – meaning they would be highly incentivised to get tested. This could, in part, further contribute to false positive rates if the sensitivity and specificity of the tests are not fully understood.

While self-selection bias is not going to be eliminated, it could be reduced by running field tests. This could involve asking the public to volunteer samples in locations where, even in a lockdown state, they might be expected to attend and also from those in self-imposed

isolation or quarantine. In any event, when statistics are communicated at press conferences or in the media, their limitations should be explained and any relevance to the individual or population should be properly delineated. It is this which we contend is lacking in the current crisis.

References

Binnicker, M. (2020). Emergence of a Novel Coronavirus Disease (COVID-19) and Importance of Diagnostic Testing: Why partnership between clinical laboratories, public health agencies and Industry is essential to control the outbreak. *Clinical Chemistry*. https://doi.org/https://doi.org/10.1093/clinchem/hvaa071

Fenton, N. E., & Neil, M. (2018). *Risk Assessment and Decision Analysis with Bayesian Networks* (2nd ed.). CRC Press, Boca Raton.

FindX. (2020). COVID-19 Diagnostics. Retrieved from https://www.finddx.org/covid-19/

Henriques, M. (2020). Coronavirus: Why death and mortality rates differ. Retrieved from BBC Future website: https://www.bbc.com/future/article/20200401-coronavirus-why-death-and-mortality-rates-differ

Oke, J., & Heneghan, C. (2020). Oxford COVID-19 Evidence Service. Retrieved from https://www.cebm.net/covid-19/global-covid-19-case-fatality-rates/

Pearl, J., & Mackenzie, D. (2018). *The book of why : the new science of cause and effect*. New York: Basic Books.

Wheaton, M. G., Abramowitz, J. S., Berman, N. C., Fabricant, L. E., & Olatunji, B. O. (2012). Psychological Predictors of Anxiety in Response to the H1N1 (Swine Flu) Pandemic. *Cognitive Therapy and Research*, *36*(3), 210–218. https://doi.org/10.1007/s10608-011-9353-3