Improving the statistics and analysis of coronavirus by avoiding bias in testing and incorporating causal explanations for the data

Norman Fenton*[1]*, Magda Osman, Martin Neil, Scott McLachlan

29 March 2020

The current COVID-19 strategic testing strategies - implemented to inform policy making - focus primarily on people already hospitalized with significant symptoms or on people most at risk. This seems to make sense for short-term medical reasons, but such testing is highly biased with sub-optimal consequences. Without understanding the causal explanations for the resulting data from such testing we end up with highly misleading conclusions about infection and death rates. Starting with an analogy of testing sweets for contamination, this short paper illuminates the need for random testing combined with causal models.

Suppose a factory makes flavoured sweets all the same shape and size. Each sweet can be one of various flavours, including strawberry, and each sweet is wrapped in a coloured foil wrapper. Unfortunately, a noxious contaminant has penetrated the factory and contaminated some of the strawberry sweets in a batch of 1000. Although the strawberry sweets are most often wrapped in a red foil wrapper, the colour of the foil wrapper is not a reliable indicator of the flavour of sweet contained therein (i.e. a strawberry sweet may sometimes be wrapped in a different colour foil wrapper).

We want to find out the following about this batch of sweets:

- What proportion is strawberry flavoured?
- What proportion of the strawberry flavoured sweets is contaminated?

It is too expensive to unwrap and test every sweet so we must do some sampling. But, instead of randomly selecting sweets, we decide to select only those wrapped in a red foil wrapper on the assumption that strawberry flavoured sweets are usually wrapped in red foil. There are 100 sweets wrapped in red foil and we discover that 50 of these are strawberry flavoured, and of those we find that, when tested, 5 are contaminated (see Figure 1).

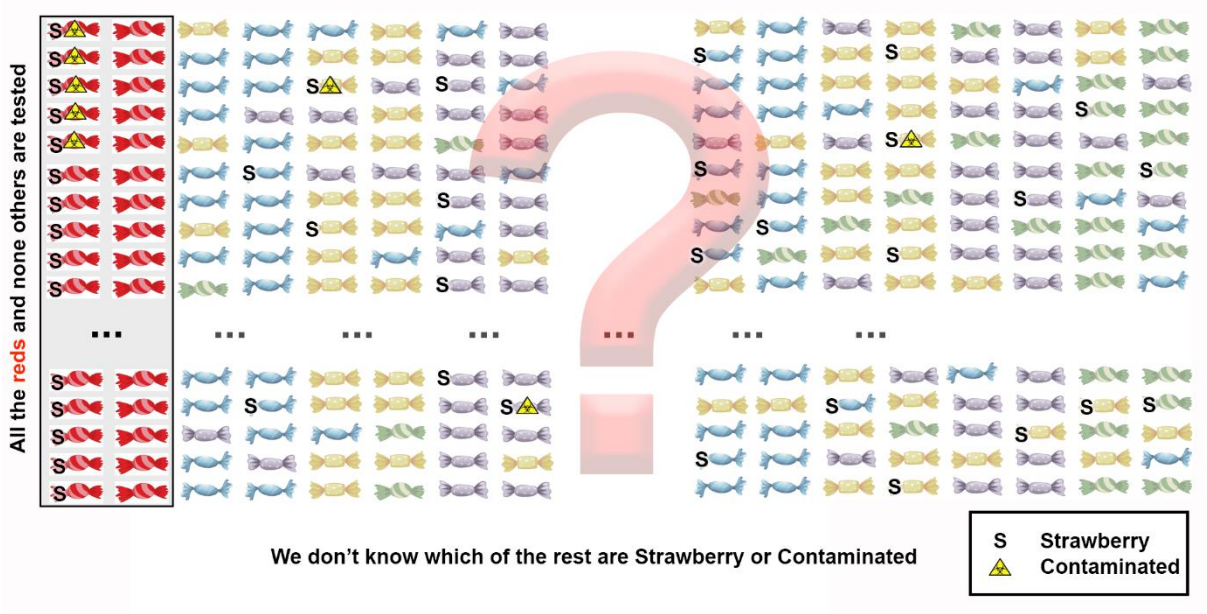

Figure 1 Testing sweets for contamination

[1] Corresponding author: n.fenton@qmul.ac.uk, Queen Mary University of London and Agena Ltd.

Are the following fair conclusions about the sweets in the batch?

- 5% of the sweets are strawberry flavoured (on the basis that approximately 50 out of 1000 sweets are already known to be strawberry flavoured)
- 50% of the sweets are strawberry flavoured (on the basis that 50 out of 100 sampled sweets were strawberry flavoured)
- 10% of all strawberry flavoured sweets are contaminated (on the basis that 5 out of 50 strawberry flavoured sweets in our sample were contaminated)

The answer is, of course no; none of these would be fair conclusions because our sampling approach was selective and biased towards only sampling sweets wrapped in red foil. We selected 100 red foil wrapped sweets and therefore sampled none of the other 900 sweets wrapped in foil wrappers not coloured red. Given this, any number of these, between 0 and 900, might be strawberry flavoured and hence any number up to 900 might be contaminated. All we can reasonably conclude is that **about 50% of sweets with red foil wrapping are strawberry flavoured and that 10% of these strawberry flavoured sweets were contaminated**. Suppose, for example, that 600 of the 900 unsampled sweets in our batch (i.e. those that are not wrapped in red foil) were strawberry flavoured, but that NONE of these strawberry sweets was contaminated. Then the true proportion of strawberry sweets that are contaminated is 5 out of 650, i.e. less than 1%.

Now, let's think of our sweet factory as an analogy for what is going on with CV19 (Coronavirus, Covid-19). Think of the sweets in the batch as the population in a country and in our analogy the strawberry flavoured sweets represent the people with CV19 and those contaminated are the people for whom CV19 results in death. So, the proportion of contaminated strawberry flavoured sweets represents the death rate from CV19 and the sweets with red foil represent people with significant CV19 symptoms. In many countries the policy for testing for CV19 is essentially equivalent to the selective and biased sampling process applied in our factory to the batch of sweets, since almost all testing of CV19 is performed on people already hospitalized with CV19 symptoms (which in our analogy are the sweets wrapped with red foil). To date, this has been the testing strategy in the UK where, at the time of writing, there are 19,522 CV19 cases (analogous to strawberry sweets wrapped in red foil) of whom 1,228 have died (strawberry sweets wrapped in red foil that are contaminated). Concluding that the death rate from CV19 is 6.3% (1228 out of 19,522) is therefore equivalent to making the mistake of concluding that 10% of all strawberry flavoured sweets contain a contaminant.

Yet such conclusions are widespread in published reports and analyses, such as the analysis conducted by the Oxford CV19 Evidence Service (Oke & Heneghan, 2020). In this analysis the UK death rate of 6.3% is exceptionally high compared to Germany where it is 0.74%, or New Zealand where it was 0% until today (Guardian, 2020). In sharp contrast Italy, with a death rate of 10.6%, is significantly higher still. Although the Oxford analysis acknowledges a degree of uncertainty in the form of confidence bounds, it completely ignores the different sampling policies applied in each country. More critically still, the Oxford study ignores the unknown and perhaps very high number of people infected with CV19 that are asymptomatic or have already recovered from the virus.

The Oxford study does state that other factors affect the death rate, namely:

- Selection bias can mean those with severe disease are being preferentially tested.
- There may be delays between symptom onset and deaths which can lead to underestimation of the rate.

- There may be factors that account for increased death rates such as coinfection, poorer healthcare, patient demographics (i.e., older patients might be more prevalent in countries such as Italy).
- There may be increased rates of smoking or serious comorbidities in those who have died.
- Differences in how deaths are attributed to coronavirus: dying with the disease (association) is not the same as dying from the disease (causation).

There are many others of course, including the rate of infection and how this is affected by social distancing and other prevention policies. But the Oxford team - and other analysts - fail to incorporate explicit causal explanations, and models, that might enable us to make more meaningful inferences from the available data, including data on virus testing. Figure 2, is an example of a causal model (Pearl & Mackenzie, 2018) for a given country and its population that shows that the CV19 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. Therefore, different countries may appear to have different death rates, but only because they have applied different sampling 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 process by which the data is generated we can better account for these differences between countries and more accurately learn the underlying true population infection and death rates from the observed data.

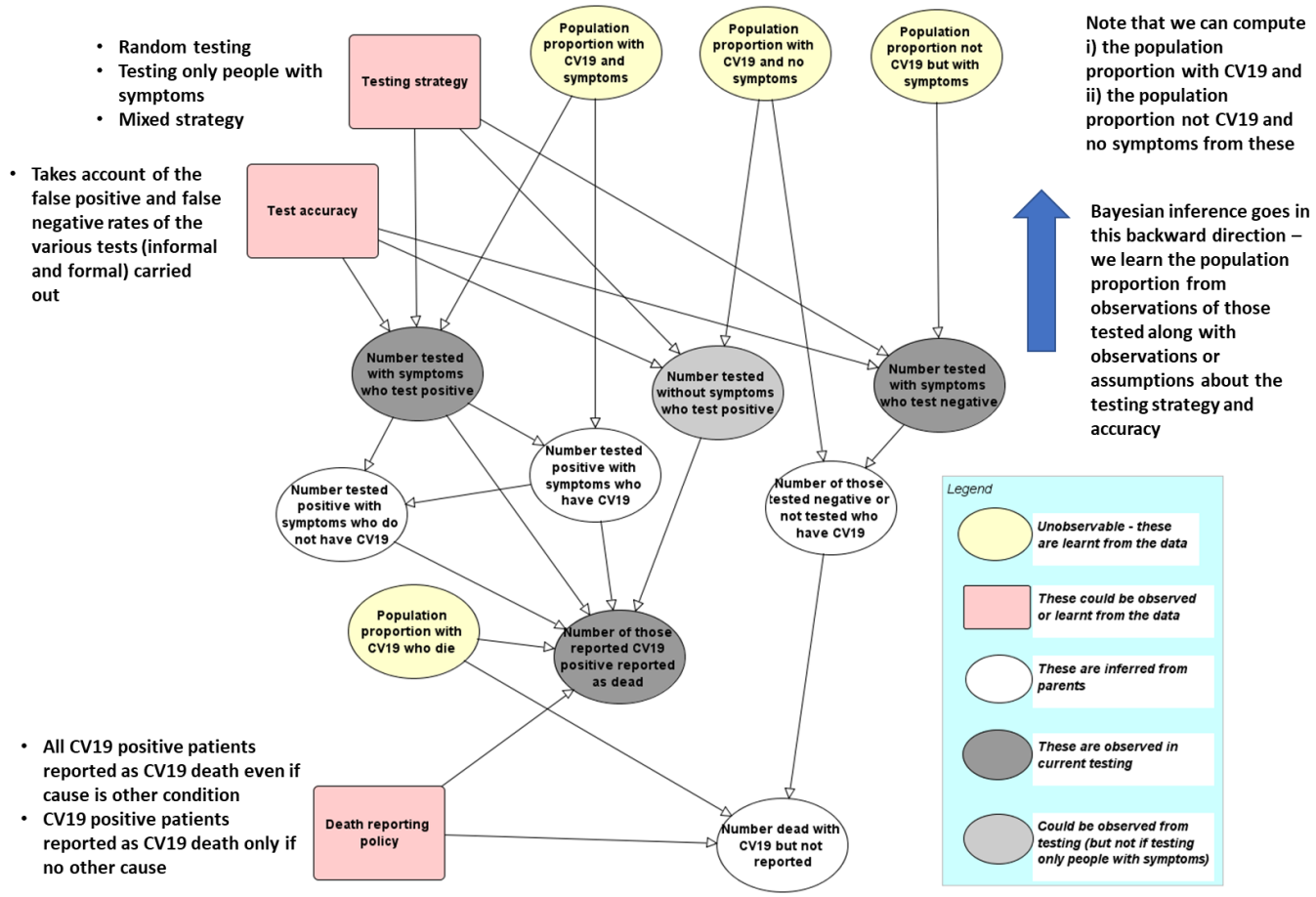

Figure 2 Causal Bayesian network model for learning population CV19 infected and death rates. This is a single time slice. It is can be updated daily and so enables us also to learn the true rate of spread of the infection. 'With symptoms' means serious symptoms.

Even more importantly, in the absence of extensive random testing applied throughout the population we will learn almost nothing about the number of people with CV19 who are asymptomatic or have already recovered. Mass testing is likely expensive, and we know there can be long waiting times for test results but without it we are effectively blind and any attempt to quantify these is pure speculation. Likewise, there are also many known complications in substantiating the accuracy of the tests (false positive, and false negative rates), but it remains the best way to find out what the underlying infection and death rate is. Hence, the most effective strategy to avoid selection bias and reduce the distortions in reported statistics is to carry out random testing (Lourenco et al., 2020; Ortiz-Ospina & Hassell, 2020; Pengilley, 2020). But, to better determine the prevalence, severity, and ultimately societal impact of CV19 the results of such testing needs to be combined with a causal model (Fenton & Neil, 2018) like the one in Figure 2.

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/

Guardian. (2020). New Zealand: first coronavirus death is woman in 70s. Retrieved from https://www.theguardian.com/world/2020/mar/29/new-zealand-first-coronavirus-death-is-woman-in-70s

Lourenco, J., Paton, R., Ghafari, M., Kraemer, M., Thompson, C., Simmonds, P., … Gupta, S. (2020). Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. *MedRxiv*, 2020.03.24.20042291. https://doi.org/10.1101/2020.03.24.20042291

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

Ortiz-Ospina, E., & Hassell, J. (2020). How many tests for COVID-19 are being performed around the world? Retrieved from Our World in Data website: https://ourworldindata.org/covid-testing

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

Pengilley, V. (2020). Coronavirus experts call for random blood testing to slow the spread. Retrieved from ABC News website: https://www.abc.net.au/news/2020-03-17/coronavirus-experts-call-for-random-blood-testing/12060666