Modelling the COVID-19 pandemic – how it can be done and why we should be cautious

Murad Banaji

03/04/2020

Mathematical modelling plays an increasingly important role in understanding the spread of disease. We have seen mathematicians respond with remarkable alacrity to the COVID-19 crisis releasing models which attempt to explain and predict the frightening numbers we see growing on our screens each day. Some of these models have already generated significant debate, criticism and even controversy. Models can potentially have an impact on public discourse, for example by suggesting the extent to which measures like lockdowns, or contact tracing and quarantining, are likely to be successful at slowing the spread of the disease. Not all of the models will stand the test of time as we gain better understanding of the disease. This note attempts to provide a rough idea for the non-mathematician of what people modelling epidemics do, and why, in the case of COVID-19, conclusions based on modelling need to be treated with caution.

I myself am a mathematician with an interest in disease modelling – but am not an epidemiologist or a public health expert.

Methodologies

How does one build a mathematical model of the spread of a communicable disease, namely a disease which can be transmitted from person to person either directly or indirectly? Well, we start with individuals who may be in different states, for example, susceptible to infection, infected but not infectious, infected and infectious, immune from infection, and perhaps more. These individuals can *interact* with each other either directly, or via other organisms or shared objects termed "vectors" of the disease, and in so doing the individuals may change state. We'll call interactions of the kind which might transmit the disease "contacts". Sometimes contact is indirect - in the case of malaria, a mosquito (the vector) may bite an infected person and then a susceptible person, and thus transmit the malaria parasite to the susceptible person causing them to become infected. In the case of COVID-19, the vector might be a shared table sneezed on by an infected person, and then touched by a susceptible person – we'll still think of this as the infected and susceptible person interacting *via* this table. Aside from interactions which may spread disease, we are interested in processes such as recovery, vaccination, or death, namely processes through which individuals change state or are removed from the picture without necessarily interacting with others. For example, a susceptible person who is vaccinated becomes an immune person, as does an infected person who recovers, assuming that recovery is accompanied by immunity.

Although conceptually we start by thinking about individuals, in many modelling approaches, individuals are not the basic units. Most models will ultimately run on computers, and having individuals as the units causes an explosion in model size and computing power needed. It may also be difficult to find the kind of data needed to construct the model, and hard to draw general conclusions from the model. To reduce this complexity, individuals in each category are often grouped into *populations*. Thus we might have a susceptible population, an infected population, a recovered population, etc. In this way of thinking, when a susceptible person becomes infected, they "move" from the susceptible population into the infected population; if an infected person recovers and becomes immune to the disease, they move from the infected to the immune population; and so forth. If an individual in some population dies, the size of that population decreases.

Underlying all epidemic models are the likelihoods of various occurrences, most importantly those which lead to disease transmission. Public health debates often focus on what *can* happen without

quantifying the risks of it actually occurring – for example many people may fear getting unwell by touching a lift button, without considering if it is likely. But to build reasonable models, modellers need access to probabilities, not just possibilities. When the units in the model are individuals rather than populations it is natural for probabilities to enter explicitly: each time a susceptible and an infected individual meet we may, roughly speaking, toss a (biased) coin to decide if disease transmission occurs. In population models though we often hide away the probabilities in "rates" of transmission. Typically, the rate at which people become infected will be a function of the total number of infected people and the total number of susceptible people. Often modellers will try to infer these rates from real-world data on the disease without worrying too much about the exact mechanisms behind them.

Once we've set up a model we "run" or "simulate" it. This generally involves "solving" equations approximately or exactly, more often than not with the aid of a computer, and then trying to get some useful information from the solutions. This might be a prediction of what happens next in different scenarios, or an explanation of some surprising data that we have observed. We might also want to use the model to help predict the effects of interventions such as social distancing, greater hygiene, contact tracing and quarantining, or lockdowns on the spread of disease. Model predictions can be compared to data which wasn't used in constructing the model to give a sense of how well the model is performing.

With these generalities in mind, let's come to why we need to be cautious when it comes to models of COVID-19 at this moment in time. In my view, this disease presents a modeller with exceptional challenges for many reasons.

Ignorance

The first problem is ignorance. There seems to be so much that we don't really know about COVID-19. Take the most basic question of fatality rates. There is huge variation between countries' so called "case fatality rates", namely the death rates amongst individuals who are *confirmed* to have the disease. Explanations for these variations are speculative and often unconvincing. Data collection and recording is of variable quality and the numbers need to be treated with extreme caution. Also varying widely are estimates of a more important number than the case fatality rate, namely the "infection fatality rate". This is the proportion of infected individuals, including those not picked up through testing, who will eventually die – the true fatality rate if you like. Since most infections are not picked up this number has to be estimated, and data scientists have come up with a variety of plausible values for COVID-19. But the spread in these estimates is large. And some have speculated that the actual value of this number varies considerably between different populations.

Apart from fatality rates, we don't yet know how most transmissions occur. We know that the disease can be transmitted in airborne droplets, and that the virus can remain viable for various lengths of time on different surfaces. There is some evidence that poorly ventilated indoor spaces are often a site of transmission. Without better understanding of transmission, it is hard to know what counts as "contact" between individuals – does singing in a choir next to an infected individual pose an infection risk? Can receiving money from an infected individual spread the disease? To what extent can asymptomatic or presymptomatic individuals transmit the disease? There is evidence that they can do so, but we don't know how significant such transmission routes are overall. The list of what we don't know goes on. There is not yet clarity on whether viral load or initial viral dose correlates with worse outcomes, whether recovered individuals can be reinfected, or what is the greatest length of time that someone can remain infected and infectious. As better data becomes available, some of these questions will be answered, but at this stage there is a lot of guesswork.

Variability

Distinct from ignorance, a second difficulty for a modeller is variability. There are very real differences in the way the disease behaves in different individuals and environments. Focussing on individuals first, the time course and severity of the disease seems to differ immensely from individual to individual. At one end of this spectrum we have people unaware they have the disease, while at the other people die. COVID-19 seems to be heavily age-biased, but we don't yet understand why, and whether age is truly the variable of importance or there are other variables which correlate with age and lead to older people being worse affected on average. Although fatality rates clearly rise with age and comorbid conditions, it is possible that they vary with a whole host of other factors as yet unidentified.

One approach to variability is to try to understand it and incorporate it into models. If we are being very ambitious, apart from focussing on transmission of disease, we may also try to understand the interplay between human physiology and viral dynamics to find out how the disease plays out inside us. A good understanding of this will help explain why individuals respond so differently to infection. In the case of COVID-19 it is early days and there is a great deal of clinical and laboratory study which needs to happen before our understanding is sufficient to model this interplay – such work takes time. There are no quick fixes here.

Aside from biology, environments too vary immensely. Epidemics occur in a context – the physical environment couples with social, cultural and political factors to affect the way that they spread through populations. How tactile we are, notions of personal space, notions of hygiene, and so forth, are heavily dependent on culture and personality, and may correlate with gender and age. Contact structures between people show great variation which is necessarily simplified or sidestepped in models. Factors such as population density, public sanitation, and the existence of shared spaces can also vary dramatically between localities, altering indirect structures of contact and affecting the disease spread. There may be subtle complications affecting the disease dynamics – for example the nature of contact within groups more likely to be asymptomatic, such as children, may be very different from contact between groups more likely to be symptomatic, such as the elderly.

Not every kind of variability necessarily matters when building models, but variability can't just be brushed under the carpet – it can be important in itself. You may have heard that each COVID-19 sufferer on average infects two or more others in an uncontrolled situation. This number can be estimated from data on infections of the kind that most governments are releasing at the moment and actually seems to vary significantly from country to country. It is termed the basic reproduction number – or R0 (read "R nought") – of the infection and is very definitely *not* an innate feature of the disease. Rather, it arises from the disease, human behaviours, and the environment acting in concert. This is good news – if it were an absolute unchanging quantity then we could never effectively fight disease since most attempts to do so can be seen, basically, as attempts to lower this number to below 1.

Anyway, although R0 may get quoted a lot, it is a crude average figure, and its variation may be as important as the number itself. There is evidence that some coronavirus-infected individuals, because of their behaviours and perhaps their biology, do not cause many further infections, while others cause a large number of infections. If we ascribed R0 values to individuals, we would see a wide spread in these values. To see why this variability is important for the disease dynamics and control strategies consider the following hypothetical scenario where most individuals are low transmitters (let's call them "recluses"), while a few are very high transmitters (let's call them "socialites"). If a recluse is infected initially, it is quite likely that the infection will not cause an

outbreak. But the opposite is true if a socialite is infected. On average the disease will need to be introduced into a population several times to cause an outbreak. And even when the disease is widespread, measures such as education, contact tracing and quarantining could be quite successful in slowing its progress, as long as they are thorough enough to reach most of the socialites. If, on the other hand, there was no variability in R0, then assuming this value is greater than one any introduction of the disease into a population will generally cause an outbreak. And control measures which miss a few infected individuals are likely to be followed by new outbreaks. Population models generally cannot include variability in R0 values, and thus risk missing such effects – individual-based models can however take this variability into account.

Conclusions

So what can we conclude about the usefulness of attempts to model COVID-19 so far? Our lack of knowledge, plus the well-documented variability between individuals and contexts presents a huge hurdle to model-building. Any model will be forced to dramatically simplify reality and its conclusions will become suspect. Once we have somehow built a model we may take an empiricist position and focus on evaluating it by comparing its predictions with data. However, it is fair to say that the quality of data on COVID-19 is poor. Testing levels and data collection methodologies vary widely from country to country and transparency may be limited. Thus model testing also presents difficulties.

Bearing in mind ignorance, variability and the poor quality of much available data, our main conclusion is that we need to be cautious in using models to inform public health decisions about COVID-19. All the models constructed so far and in the foreseeable future will be based on assumptions which, if unpacked, many an epidemiologist or public health expert would question. Nevertheless, modelling at this stage is not useless. Rather than relying on quantitative predictions of models, we should focus on their explanatory power, power to stimulate debate, and the qualitative insights they give us. They may suggest hypotheses, or provide new testable theories about what is happening underneath the surface of available data. The biggest positive to come from some of the models that have made it into the public sphere is that they have sparked important debates about numbers, data, interventions. Some very basic models have served to educate about measures such as social distancing. Given the severe economic and human consequences of some mitigation and containment strategies like lockdowns, it cannot harm to ask models to contribute to the discussion on how likely they are to be successful, and to help evaluate alternative strategies. Models may also be able to highlight just how serious problems like inadequate hygiene in hospitals or a lack of protective gear for medical and support staff in hospitals could prove to be.