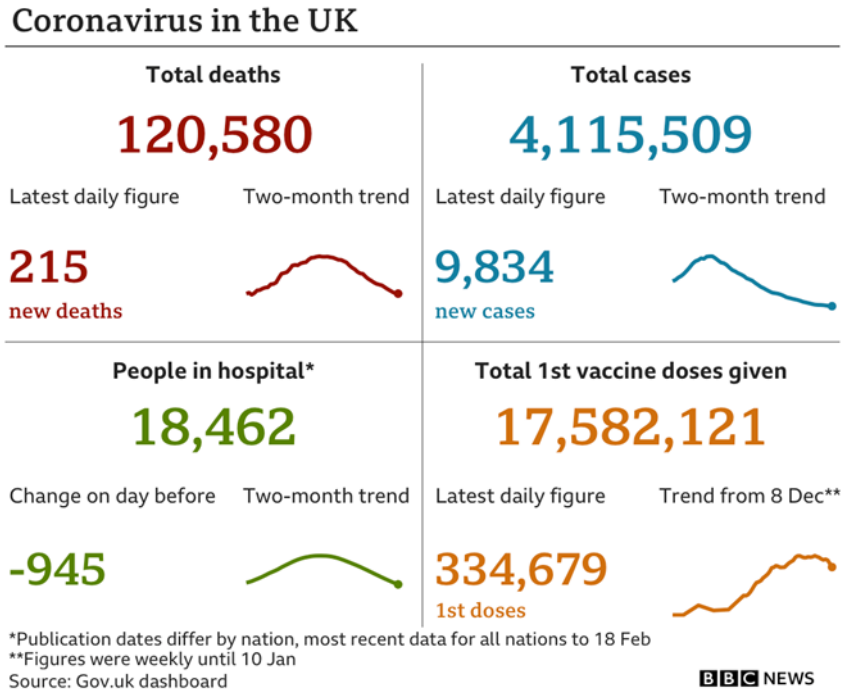
Interpreting Data in Times of Covid

By Ian Tonks[1]

The global pandemic of 2020 has turned everyone into data experts. It is a common occurrence on the television, in supermarket checkout queues, and when out walking the dog, to hear people discussing the latest daily coronavirus updates. Overnight everyone has acquired the knowledge and skills of data analysis. Figure 1 illustrates a typical daily coronavirus update taken from the BBC news website showing number of reported cases of coronavirus, hospital admissions, deaths due to coronavirus and since December 2020 the number of vaccinations.[2]

Figure 1: coronavirus updates



However, all social data such as illustrated in Figure 1 is subject to caveats and the pandemic obsession with data and the numbers contained within carry lessons which are applicable to other business and social settings.[3] There is a temptation in the presentation of this data to interpret a causal model from numbers of cases via hospital admissions to numbers of deaths. Indeed, to interpret data, or to ask what is the data for, requires a model? What is the underlying model behind the data in Figure 1? Where has this data come from and how has it been collected? What triggers a count of a covid case: is it based on a suitable test or self-reported symptoms? Are there common reporting standards across countries to enable an assessment of policy responses? In general, the more data points the better, but has reporting changed over time or has the underlying data generating process changed, possibly due to new variants of the virus? The remainder of this article examines these issues in the interpretation of covid data, provides some answers to the questions posed and draws lessons for the general interpretation of social and business data..

[1] School of Accounting and Finance, University of Bristol: I.Tonks@bristol.ac.uk
[2] This in turn is taken from the official UK government website: https://coronavirus.data.gov.uk/
[3] De Vaus, (2002). *Analyzing social science data: 50 key problems in data analysis* (Sage, 2002).

First let us address the modelling of a pandemic for which there is a long history in the discipline of epidemiology.[4] Data without an underlying model is just a set of random numbers which may or may not be useful. A model potentially identifies the causation between inputs and outputs. The simplest mathematical model of an infectious disease is the SI model. In this model there are just two states: susceptible, S(t); and infected, I(t), with a very simple flow chart that people are either susceptible or become infected. Let *S* be number of susceptible people in a population of size *n*, *X* be number of infected individuals, and β is number of people a contact randomly meets per unit of time. An infected person will be in contact with β*S/n* susceptible people, and the rate of new infections will be β*SX/n.* Then the differential equation for rate of change of X is

$$\frac{dX}{dt} = \beta \frac{SX}{n}$$

If we divide both side by n: $\frac{dX}{dt}\frac{1}{n} = \beta \frac{S}{n}\frac{X}{n}$ and writing in fractions of susceptible/infected persons, and noting that every person is either susceptible or infected, so $s = (1 - x)$ we may write

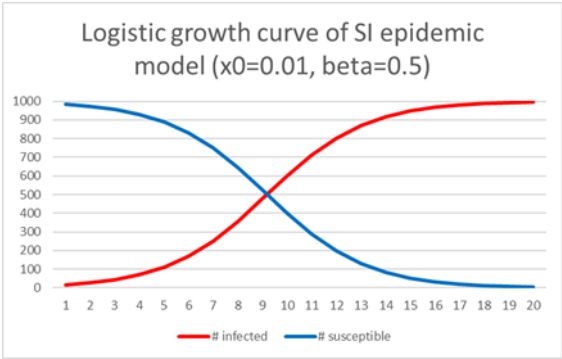
$$\frac{dx}{dt} = \beta(1 - x)x$$

which is a logistic growth equation with solution:

$$x(t) = \frac{x_0 e^{\beta t}}{1 - x_0 + x_0 e^{\beta t}}$$

where x_0 is the fraction of the population that is initially infected. A parameterisation of this model is represented in Figure 2 showing the daily share of infected and susceptible persons in a population of 1000 persons with an infection rate β = 0.5 (1 person every two days) starting with $x_0 = 0.01$.

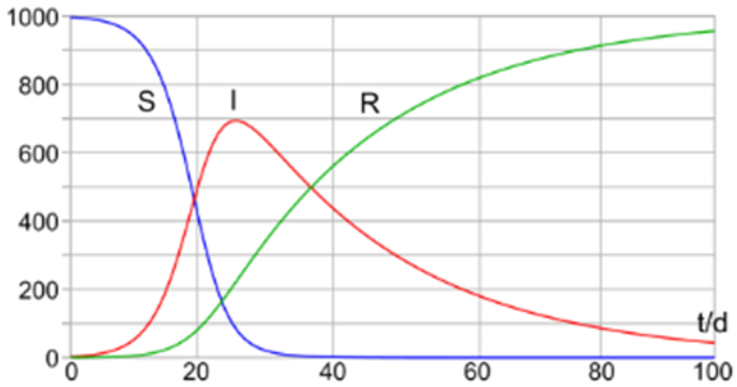
Figure 2: Logistic diffusion model



In this very simple model the shape of the curves are determined by the rate of infection, with a smaller value of β flattening the spread of the disease. A slightly more realistic model is the three-state SIR model, with the inclusion of an additional recovery/immune state. We will finesse the mathematics of this and other ever-more realistic models, and simply note the set of diffusion curves for the SIR model illustrated in Figure 3.

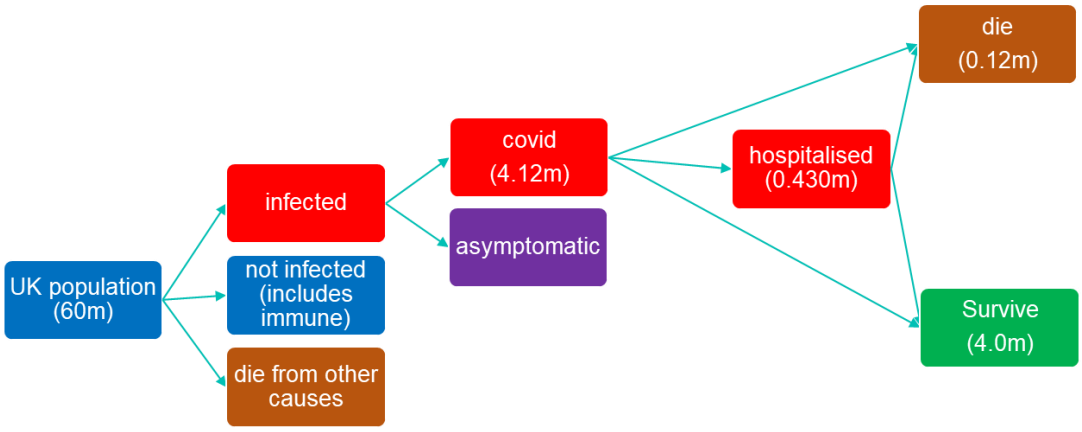
[4] Kermack, W. O., and McKendrick, A. G. (1927). "A Contribution to the Mathematical Theory of Epidemics". *Proceedings of the Royal Society A*. **115** (772): 700–721.

Figure 3: Diagram of SIR model (initial values S(0)=997; I(0)=3; R(0)=0), and infection rate β=0.4; and recovery rate γ=0.04



There is an animated graph available on Wikipedia illustrating the "flattening of the curve" in Figure 3 (the red line I, showing the number of infected persons) as the infection rate declines.[5] The infection rate is a parameter within the model that the government can potentially control and is the basis of social distancing and lockdown policy responses: reduce the rate of infection to hold up the spread of the pandemic. Policy responses of handwashing, social distancing and lockdowns aims to reduce β and flatten the red curve. Slowing down the rate of infection allows health infrastructure capacity, such as hospitals, to better cope with pandemic and ensure the provision of other health services. This example illustrates the benefits of imposing a model on the data, since such modelling identifies parameters that can be affected by policies. Similarly, a vaccine is a policy response that reduces the number of susceptible people in the population. In Figure 4 we outline the flow chart in a more realistic model, to give a flavour for how more complicated settings may be allowed for, although we do not attempt to solve this model.

Figure 4: Possible flow chart for a pandemic



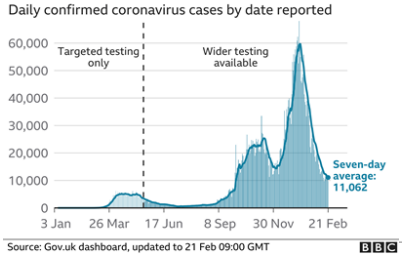
[5] https://commons.wikimedia.org/wiki/File:SIR_model_anim.gif

We may fill in some of these boxes using the reported data to get estimates of the prevalence of the virus within the population.

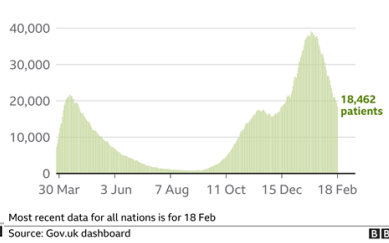
The data illustrated in Figure 1 also includes the most recent two-month time trend. In fact, data on these variables go back to the start of the pandemic in the UK from March 2020, and we reproduce these three diagrams in Figure 5. Which are also commonly reported.

Figure 5: Time trends for covid cases, hospital admissions and related deaths since March 2020

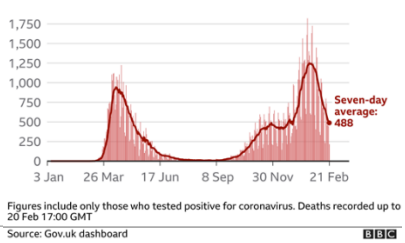
Panel A



Panel B



Panel C



Each diagram clearly identifies the first and second/third waves of the pandemic throughout 2020 and its effects, with the lull during the summer. From these pictures it is tempting to infer that the pandemic was relatively mild in its first wave and only became more widespread in the latter part of 2020, and the rate of hospitalisation and death rates conditional on cases were higher in the first phase, so that the virus is less virulent over time. However, this inference is based on the assumption that the reporting of testing is accurate and unbiased, and that the underlying virus and rate of transmission has remained the same throughout the year.

In fact, these are rather heroic assumptions, and relate to stationarity and selectivity issues that plague much reported social data. In an analysis of a time-series of data it is important that the underlying process generating the data is stationary. That is, whatever is generating the underlying data, in this case the spread of the virus, needs to be the same at the start through to the end of the data series to allow comparisons to be made over time. In fact, it was well-documented in December 2020 that the virus had mutated and had become more infectious, which then means that comparisons over time become debateable. Further, there are at least two types of selectivity issues in the reporting of the testing data. In the initial phase of the pandemic in March 2020 the UK did not have the testing capacity to keep up with the spread of the virus, and therefore testing was limited to at-risk persons with severe symptoms. So people with mild symptoms or asymptomatic persons were not tested. This targeted testing was cheap and quick, but likely to under-report the true number of positive cases. The data in Panel A of Figure 5 will have incurred selection effects with the results being a poor indicator of the prevalence of virus in the population (rate of infection). Further these initial tests on the basis that they were new may have been of lower quality and less reliable, with testing quality improving over time. A further selection effect in the data is that over time behavioural responses of persons will have changed: the knowledge that covid primarily affects the elderly, may have resulted in this group of the population locking themselves away to protect against being infected, and hence the reported infection rates in later periods only reflects the infection rates of the sub-set of the population that are not taking precautions because they are not at-risk or are flouting the regulations.

There are at least three reasons why comparisons of the data over time are problematic: 1) Stationarity of the data generating process; 2) Changing reporting methods; and 3) Changes in behavioural responses.

The UK government improved its data collection as the pandemic has spread with an evolving four pillar testing strategy initially outlined in April 2020: 1) lab-lased swab tests (patients & frontline health staff); 2) swab tests with a range of partners with public and private laboratories (including lab-based and lateral flow tests, from October 2020); 3) antibody tests; and 4) surveillance surveys carried out by organisation such as the Office for National Statistics. Up until July 2020 individuals who tested positive under pillar 1 and pillar 2 were double counted; since November 2020 Public health England collates positive test results by geographical area based on the location given at the point of testing and this is the basis of the data in Panel A of Figure 5. The fourth pillar of the government's strategy includes a weekly survey administered by the ONS of a cross-section sample of the population. Initially in May 2020 this was a random sample of 21,000 individuals from 10,000 households (invited from previous ONS surveys) but increased to a random sample of 150,000 people from October 2020 to March 2021. These tests involve nose and throat self-administered swab tests for the sample of surveyed household members whether displaying symptoms or not. A subsequent blood sample of survey participants who have tested positive is taken to assess antibodies (pillar three of the strategy).

Such surveys although more accurate of the prevalence of the virus in the population are more expensive and slower to report. Although these surveys overcome the selection effects from only testing people with symptoms or some other criterion, the selection effects induced by the behavioural responses will still be present.

The hospital admissions data in Panel B of Figure 5 is the number of covid-19 patients in hospital on the reporting date, although the definitions of what constitutes a covid-19 patient varies across the four UK nations, and what constitutes a hospital has changed over time.

Reporting issues are also of concern when we make comparisons across countries as there is no international standard for the testing of covid cases, hospitalisations or the definitions of death resulting from covid. In the UK the definition of deaths due to covid reported in Panel C of Figure 5 has also changed over time. Since August 2020 this mortality data reports deaths that occurred within 28 days of a positive covid test; although earlier definitions for England included deaths that had occurred after any previous covid positive test irrespective of time. An alternative definition would be where the death certificate mentions covid as possible or contributory cause of death (certified by medical practitioner).

Another metric for the mortality effects of covid is the number of excess deaths: measured as the actual number above the usual number of deaths for the time of year. As shown in Figure 6, this excess-deaths measure has been positive for most of 2020 indicating that a spillover effect of the pandemic has been on deaths due to other causes, perhaps because the health-care system has been unable to cope with regular medical emergencies; or that covid has induced a behavioural response from at-risk individuals who have been unwilling to expose themselves to risks of covid.

Figure 6

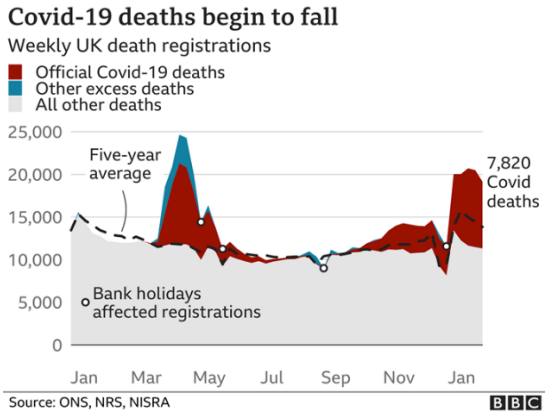
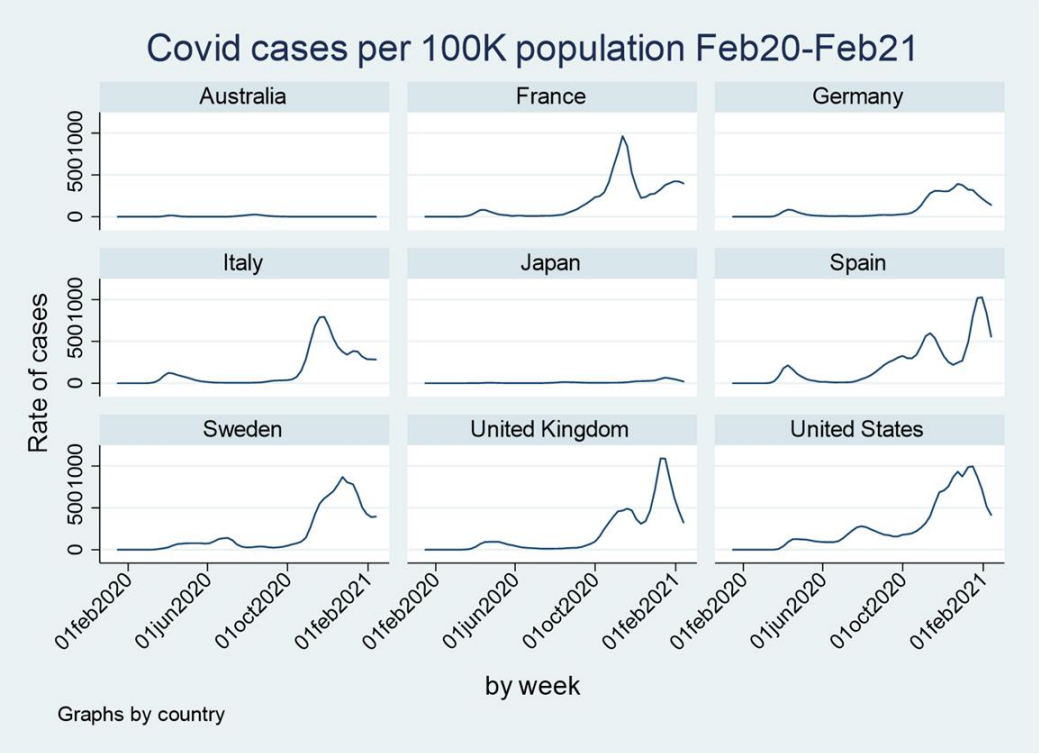


Figure 7 shows that the UK has apparently fared badly in international comparisons of covid cases and ultimately death rates, but this could be due to a multitude of factors including demographic differences of people most at risk, density of populations, medical infrastructures, responses to the pandemic as well as different reporting standards. The data in Figure 7 is sourced from European Centre for Disease Prevention & Control[6], and their website contains a warning about making international comparisons because of cross-country differences in collecting and defining the data.

Figure 7



Finally, a feature of the pandemic in the UK has been the "reliance on experts", primarily through the government's SAGE Committee (Scientific Advisory Group for Emergencies). This reliance grates with the downplaying of scientific data by politicians in other situations. For example, during the

[6] https://www.ecdc.europa.eu/en/covid-19/data-collection

Brexit campaign Michael Gove responded to a letter to the Daily Telegraph by a group of economists warning about the economic consequences of Brexit, by saying “people in this country have had enough of experts” (3 June 2016). Donald Trump criticised the US Federal Reserve for being “wrong so often” and suggested 2021 would be “one of our best ever years” (11 June 2020). The advantage for politicians of expert scientific panels is that difficult political decisions can be blamed on these experts, and politicians with concerns about re-election can not always be relied upon to make the best decisions. In financial markets many of the developed world’s central banks are independent of political meddling: the US Federal Reserve (since 1913), Germany’s Bundesbank (since 1945), The Bank of England (since 1998) and the European Central Bank since it was formed in 1998.

In summary, the takeaway from these data issues is that any social data comes with a health warning that there are reporting and selection issues that change over time, and any interpretation of the data needs to be minded that comparisons may not be like-with-like.

Ian Tonks
University of Bristol
26th February 2021