2020: A year of data we cannot trust

Contribution to World Statistics Day 2020, the motto of which is "Connecting the world with data we can trust"

Ondrej Vencalek, Czech Statistical Society

During 2020, the lives of people all over the planet have been affected by the threat of the coronavirus disease 2019 (COVID-19). It is not just the direct impact of the disease on the health of those who undergo it. I dare say that various restrictive measures directly or indirectly affect the lives of most of the inhabitants of this planet. All these claims can be supported by data. In this contribution I want to point out that even the simplest questions about COVID-19 have ambiguous answers because we do not understand what the available data mean.

Let us start with those affected the most – the infected ones. How many people have been actually infected? It's (seemingly) easy to find out: Just enter “COVID-19” into your favourite search engine and you will find the information that 36.1 million people have been infected worldwide so far (information retrieved on October 8, 2020). This number can be found, for example, at wikipedia (1–2) or at the Center for Systems Science and Engineering at Johns Hopkins University (3). In fact, 36.1 million is not the number of people infected by the coronavirus (more precisely severe acute respiratory syndrome coronavirus 2, known as SARS-CoV-2), but the reported number of cases of COVID-19 – an infectious disease caused by SARS-CoV-2. This difference is very important. Wikipedia spots a small warning: “The way of reporting and testing capacity varies from country to country, and the actual number of people infected is probably higher”. So, the answer to the (seemingly simple) question of how many people have been infected by SARS-CoV-2 is unknown. We only have data on the number of positively tested. However, these numbers depend not only on the presence of the virus in the population, but also on other factors, especially the intensity and strategy of testing (how many tests are done and who is being tested). If we stopped testing altogether, the number of positive cases would stop increasing, but the virus would hardly stop spreading. There seems to be a difference between a “positively tested person” and a “person suffering from COVID-19”, yet according to the WHO definition, anyone with laboratory confirmed COVID-19 infection, irrespective of clinical signs and symptoms, is considered to be a “confirmed COVID-19 case” (4). Many people who have the virus on their mucous membrane but exhibit almost no symptoms are also identified as COVID-19 cases. On the other hand, there certainly are people who have contracted the virus but have not been tested and therefore are not formally “sick” (as with other diseases, a person with a mild course may stay at home and may not be among the tested). Numerous seroprevalence studies, summarised in a meta-analysis by John Ioannidis published in July this year, suggest that the actual number of people infected by SARS-CoV-2 may be much higher than the number of those tested positive (5).

Let us now turn to those who have not yet been directly affected by the disease. Their lives are also affected by a number of restrictions imposed in an effort to slow down the epidemic. We would expect that decisions that affect the lives of virtually all citizens of the state are made after a thorough analysis of the epidemiological situation, i.e. that they are supported by data. Consider, for example, the Extraordinary Measure issued by the Ministry of Health of the Czech Republic on September 23, 2020 (6). In the justification of the restrictive measures, we read that “the goal is to maintain the case fatality rate, as it has been in the Czech Republic so far, in the range of about 2–3%, without increasing it to a global average of almost 7%, or even 10 and more percent, as is currently the case especially in France (here the case fatality rate is already almost 18%), ...”. The document does not state where these numbers came from. On the Johns Hopkins Coronavirus Resource Center website (7), countries can be

sorted by case fatality rates (CFR). It should be noted that CFR is not only a property of the virus describing its "ability to kill" but is also affected by the functionality of health systems – the ability to care for patients with this virus. This, to a large extent, again depends on the testing strategy. Countries with the highest CFR are Yemen (28.9%), Italy (10.8%), Mexico (10.4%), Ecuador (8.2%) and the United Kingdom (7.8%). The global average is 2.9%, for France it is 4.7%, for the Czech Republic 0.9%. So the numbers are very different. Which "data" can we then trust? Further, we should bear in mind that only such cases should be considered, in which we already know whether death or recovery has occurred, and that the number of all infected should be taken into account, not only the officially confirmed cases. These numbers are estimated in various seroprevalence studies. Then, however, the fatality rate comes out around 0.24% (5), which is about ten times smaller than the above-mentioned values. Which "data" can we trust now? Unfortunately, one may chose – either to claim that COVID-19 is a very serious disease, or to claim otherwise.

Asking two very simple questions – the number of infected people and the case-fatality rate of the infection – we have found several mutually exclusive answers. We are surrounded by data, but in order to obtain relevant information from them, we must understand how this data was obtained and agree on the definitions of notions they measure. We must understand the difference between reality (the actual number of infected) and what the data shows about reality (the number of positive tests). An initiative to increase "data literacy" has recently emerged (8) which highlights the issues I have pointed out in this contribution. It is no coincidence that the authors of this initiative ask "How much trust do we place in the data?"

Reference

1. Template:COVID-19 pandemic data - Wikipedia. [Online] [Cited 8 Oct 2020] https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data.

2. COVID-19 pandemic - Wikipedia. [Online] [Cited 8 Oct 2020] https://en.wikipedia.org/wiki/COVID-19_pandemic.

3. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). [Online] [Cited 8 Oct 2020] https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.

4. WHO COVID-19: Case Definitions. [Online] 7. 8 2020. [Cited 12 Oct 2020] https://apps.who.int/iris/rest/bitstreams/1296485/retrieve.

5. Ioannidis, John. The infection fatality rate of COVID-19 inferred from seroprevalence data. medRxiv. 2020.

6. Mimořádné opatření - omezení provozoven a provozů služeb s účinností od 24.9.2020 do 7.10.2020. (in czech) [Online] [Cited 8 Oct 2020] https://www.mzcr.cz/wp-content/uploads/2020/09/Mimo%C5%99%C3%A1dn%C3%A9-opat%C5%99en%C3%AD-%E2%80%93-omezen%C3%AD-provozoven-a-provoz%C5%AF-slu%C5%BEeb-s-%C3%BA%C4%8Dinnost%C3%AD-od-24.9.2020-do-7.10.2020.pdf.

7. Mortality Analyses - Johns Hopkins Coronavirus Resource Center. [Online] [Cited 8 Oct 2020] https://coronavirus.jhu.edu/data/mortality.

8. Data Literacy - Schweiz. [Online] [Cited 8 Oct 2020] https://en.data-literacy.ch/.