Models, data and statistics - why is it so difficult?

By Susanne Ditlevsen and Torbjörn Lundh

For a mathematical modeller these corona times have been extraordinary. Never has there been so much public interest in scientific research, where mathematical models and their predictions play a prominent role. Scientists are constantly being interviewed in public media, and researchers that have been working their entire lives outside the spotlight has suddenly become well-known media darlings. The public is allowed to peek into the messy engine room of scientific development; hypotheses, theories, analyses are presented before there has been any time to peer-review, and they are being criticised, updated and improved upon in public, as more data are arriving and more knowledge is obtained. This is usually happening behind the scenes, until a consensus has been reached. Now we lay the rails while the train is running, and this might be confusing and seem like researchers do not know what they are doing if you are not used to the scientific process.

What is a good model? And what is it good for?

An essential tool used by epidemiologists to describe the development of an infectious disease in a population and evaluate the effectiveness of various countermeasures is the use of mathematical models of various kinds. Models have thus become very important when expert knowledge is communicated to decision-makers and policies are formulated and justified.

How this should be done is not entirely clear, and many questions have been discussed during the corona crisis. Should we rely on the experts' knowledge or on forecasts from models? But what if different models provide varying forecasts? Which model should be used and how reliable are they really?

There are two extreme positions. One can argue that complex models should function as a direct basis for decision-making, so that predictions are "objective" and not based on gut feelings. One could also argue that the subject matter experts' knowledge should dominate, because they will have the best intuition and experiences with such problems. However, we believe that those two positions should not be opposed, but rather go hand in hand and complement each other. Let's look at some characteristics of scientific models that are often taken for granted by researchers but rarely discussed in the media [1].

Mathematical models, consisting of equations describing various variables, is just one type of models out of all types of scientific models. When a medical doctor describes how the virus enters a human cell through attachment to the ACE_2 receptor, then it is also a model, albeit a verbal or conceptual one. To fully describe how the virus interacts with human cells would require quantum mechanical explanations, which would be accurate, but not very informative. The different scales at which these model variants operate make it possible for researchers to isolate and zoom in on certain phenomena of interest.

A famous quote usually attributed to the statistician George Box says that "all models are wrong, but some are useful". The point is that models are tools for specific purposes, and these purposes are often of a practical nature. The usefulness of a model does not only depend on how well it describes the real world. In fact, that they work at all is exactly due to the fact that they are simplifications of a complicated reality. For example, the simplest models for how a disease is spreading simply describes the number of infected, ignoring all geographical and socioeconomical information. It will provide a rough estimate, but will be less sensitive to misspecifications of the many unknown parameters that invariably are needed in a more complex and specialised model, and might therefore, in a world of uncertainty, pro-

vide more robust estimates than more realistic models.

Simple or complex models?

We can compare a scientific model with a map. A 1:1 map that completely agrees with the landscape it intends to describe is useless as a map. Whether a given map is good or not depends on the problem it is trying to describe. If you need to find the quickest way to cross a city by car you need a different map than if you are a tourist searching for beautiful spots in the city centre or a good restaurant, or if you need to find a specific office in a large office building.

There is always (or should be!) some specific purpose or goal a model is trying to reach. Different researchers might have different goals, and thus, there will also be a large variety of models. In the case of the corona pandemic, there are very complex and detailed models with the purpose of understanding the underlying mechanisms, as well as more statistical models with the main focus of predicting number of infected or needs for healthcare measures such as hospital beds. This could be done by looking at the development of the pandemic in different countries, but without any or only very mild assumptions on the properties of the virus or the dynamics of the spread. These descriptions are not contradictory, on the contrary, they are parallel descriptions from different perspectives that could complement each other.

The point is that the models should not be looked upon as the truth, since they are always simplifying and idealising. Furthermore, more complex models are not automatically better. The models and their predictions should therefore be seen as supporting tools for political decisions, when results from different models are combined together with the empirical expert knowledge.

So how should decision makers relate to models and their predictions in the middle of the current crisis? There is no simple answer. Which models are needed for a given situation depends on various factors, but we believe that the best result is achieved when various models are used in parallel and the predictions, robustness and "understandability" are evaluated together with subject matter experts. Complex models may possibly explain more than simple models, but it is a risky business. Especially if uncertain model assumptions and simplifications are hidden or forgotten, or even being used for propaganda. Simpler models might provide a better overview, be statistically more robust because of the limited access to detailed data to validate the more complex models, but they might seem too arbitrary for non-experts.

Many data models that have been presented around the world to predict the spread of covid-19 have been very complex - despite the lack of validated data. These models are complex in the sense that they have many unknown parameters that are difficult to estimate or measure. Furthermore, the parameter values used are most often not the result of training and validation on a large and representative data base but are instead set manually, often without clearly justified support from empirical studies. Thus, one should be careful when interpreting the outputs of these models, especially with non-linear models where small parameter variations can cause large fluctuations in the model predictions.

Let's look at an example of a parameter appearing in many of the recent modelling attempts of the effect of societal measures to contain the epidemic. What is the effect of closing schools on the spread of the disease? School closure, of course, eliminates the risk of spreading infection in schools. However, it might increase the risk in the family. The total effect then depends on how you model the spread of infection in each environment. Assuming that pupils infect little or not at all in schools, the net effect is that the spread of infection can increase after a school closure. However, assuming the same spread of infection as in the family and leisure time, the spread of infection decreases significantly. Thus, depending on the model assumptions, the effect of a school closure can either be an increase or a decrease in the reproductive rate (that represents how many individuals an infected individual on average transmits the in-

fection to). It can also affect how the infection is spread between different age groups, such as the elderly. Is it the parents who look after the children, or grandma and grandpa?

How good are complex models compared to simple models at predicting reality? That depends to a large extend what we mean by "reality". If "reality" is the data that we have access to, this is often limited to time series of infected and deaths with no further information on details like who the subject was infected by, which, if any, symptoms there were, what contact patterns the subject had adopted, how many that person further infected etc. For such rough data, a simple model such as the soon 100-year-old SIR model [2] can reasonably well recreate the time series from most countries. However, this does not necessarily imply that it can predict future numbers well.

For a complex model to have predictive ability, it is required that its unknown parameter values are selected on the basis of reliable and sufficiently informative training data. In order to reduce the risk of over-adaptation to training data, separate validation data are also required against which the model can be evaluated before it is put into use.

In general, but especially in a situation where adequate training and validation data are not available, the simplest model describing available data is preferable. This principle is the well known Occam's razor, or the rule of parsimony. In addition, simple models are generally more transparent in terms of how parameter choices relate to outcomes, i.e. more understandable and can thus be better tools for thoughts and discussions. However, the need for data-supported parameter selection and validation remains, even for simple models.

To sum up: Models with higher complexity than what training and validation data can support should be used sparingly as a basis for decision-making.

What is statistics?

Mathematical models of biological phenomena only become really interesting when we can test the models against data. Statistics is a tool for translating what we can observe into knowledge about the world. Many things we want to know about the world cannot be observed directly. For example: How long does it take for a person being infected with the corona virus to develop symptoms? How does this vary from person to person? How many don't develop any symptoms, but nevertheless infect others? You can then collect data and use statistical tools to interpret data and get an estimate for the answer to the questions you have asked.

Not all questions are easy to answer. Here are some examples of things we can estimate from data that go from easy to harder.
Easy: How long does the virus stay alive on different surfaces and under given conditions? This can be tested in a laboratory under controlled conditions.
Medium: How many have been infected or are possibly immune? This requires far more testing, but if we have a test for antibodies it can in principle be done.
Difficult: How long time passes from getting the infection before symptoms appear, and when does the person become infectious? It requires that we can identify the exact time a person has been infected, when the symptoms appear, and in what time interval the person has been infectious. This has to be done for many people, since there is probably a great deal of variability from person to person, and this is also important to understand.

We need to collect data, to gain knowledge, and not base our inferences on guesswork and gut feelings. Frequently, our ideas about the world is coloured by our own most recent personal experience, or a quick look at some statistical table with no thorough analysis. However, it is tricky to translate data to knowledge about the world! For example, take the data on number of infected or deaths due to the corona virus in different countries that we are all googling in these times. These numbers cannot be directly compared, because countries calculate numbers differently, and even more importantly, have different strategies for testing for corona. So even though we have nice tables that looks very "ob-

jective", we can not use it without a deeper analysis.

Parameter sensitivity

As the amount of parameters and assumptions in a model grows, so does the requirement to validate these assumptions. Parameter values are most appropriately validated against data; and assumptions and results should be tested through the usual scientific review. When mathematical models are used to make socially important decisions, this requirement is even more important.

One should be aware about the uncertainty in the statistical estimates. The more data, the less uncertainty on estimates. Therefore, estimates are constantly updated because more and more data is being collected. This has to be done carefully, because if the data collection process is not accounted for in the models and data is analysed incorrectly, we get a biased result, that is, a systematically wrong result. If you then collect more and more data, you will get increasingly more accurately estimates of something wrong. This is particularly important in larger, non-linear complex models, where small perturbations in parameters can make huge effects on the model's output.

High quality data – to learn about the disease or to combat the disease?

Good estimates of model parameters are needed to make useful predictions – and for that we need statistics. And to do statistics, we need data. Not only that. We need good quality data. But what does good quality data mean in this context?

In many countries there are a lot of discussions about the best test strategy. Testing requires resources, and therefore decisions on whom to test have to be taken. However, there is almost exclusively focus on one aspect of it, namely, what is the best strategy to get as many as possible through the crisis by containing the infection and limit the number of deaths. This is a very important issue, but there is another issue that is only rarely discussed. That is the statistical issue: What is the best test strategy to gain the most insight into the disease? The golden standard is randomised trials, which is the best way to ensure reliable estimates of parameters of interest. In addition to testing individuals who have shown symptoms, we should in parallel test a randomly selected sample of the population, whether they have symptoms or not, while registering important background variables, such as gender, where they live, age, activity level, state of health, possible symptoms, etc. This is the only way we can obtain solid estimates for the disease-specific parameters. Preferably, everyone in the sample should be tested several times. It would provide a data material to much more reliably estimate how long it takes from infection to symptoms, how many have no symptoms but are still infected, the herd immunity level, the reproduction rate, etc. If the tested are selected not randomly, but because they have symptoms, or have specific functions like health care workers, this will not be a random sample and estimates will be biased. It is conceivable that they are more exposed to viruses, and therefore the immunity may develop differently. Moreover, the age distribution is probably different than in the general population. Furthermore, it cannot inform us about those infected without or with mild symptoms. So we have to do randomised trials in more than only a few countries.

We sum up this story by quoting Rutherford D. Rogers: "We are drowning in information and starving for knowledge".

References

[1] P. Gerlee and T. Lundh Scientific models: red atoms, white lies and black boxes in a yellow book, *Springer*, 2016.

[2] F. Brauer. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2, 113–127, 2017.