AUTHORS:
Farai Nyabadza[1,2]
Alex Broadbent[3]
Charis Harley[2,4]
Abejide Ade-Ibijola[2,5]
Ebrahim Momoniat[1,2]

AFFILIATIONS:
[1]Department of Mathematics and Applied Mathematics, University of Johannesburg, Johannesburg, South Africa
[2]Data Science Across Disciplines Research Group, Institute for the Future of Knowledge, University of Johannesburg, Johannesburg, South Africa
[3]Institute for the Future of Knowledge, University of Johannesburg, Johannesburg, South Africa
[4]Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg, South Africa
[5]Department of Applied Information Systems, College of Business and Economics, University of Johannesburg, Johannesburg, South Africa

CORRESPONDENCE TO:
Ebrahim Momoniat

EMAIL:
emomoniat@uj.ac.za

HOW TO CITE:
Nyabadza F, Broadbent A, Harley C, Ade-Ibijola A, Momoniat E. Models and muddles in the COVID-19 pandemic. S Afr J Sci. 2021;117(9/10), Art. #9506. https://doi.org/10.17159/sajs.2021/9506

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
mathematical models, epidemiology, infectious diseases, public health, COVID-19

PUBLISHED:
29 September 2021

Models and muddles in the COVID-19 pandemic

> *It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*
>
> Sir Arthur Conan Doyle

The COVID-19 crisis is an opportunity for scientists to showcase their skill and the impact that good science can have on society. However, not all scientists have risen to the occasion with the sense of responsibility and accountability that their work deserves. Scientists worldwide have shown, and continue to show, great enthusiasm regarding the use of specific scientific tools, mainly modelling and predictive analytics, to estimate how the virus spreads and behaves and to assess interventions against counterfactual scenarios. In this Commentary, we question whether the application of these tools has always been appropriately managed by discussing the underlying elements of modelling which need to be understood and evaluated for results to be meaningful and credible.

A mathematical model must capture the principles that dictate the dynamics of what is being modelled: assumptions, constraints and relevant natural laws, for example. These principles serve as the 'rules' for understanding the results obtained and provide the context within which the model has meaning. Within this context, the model goes beyond being a mere collection of mathematical operations and represents – albeit in idealised or imperfect form – some feature of the actual world. Here we argue that these rules have often been ignored when engaging with the results obtained from mathematical models used for predictive purposes in the COVID-19 pandemic (including policy purposes) and from data-driven models designed via machine-learning methods.

To make our case, we first provide some context of the origins of disease modelling and then offer a 'current day' frame of reference which illustrates why caution is needed when employing models for prediction.

Infectious disease modelling

Infectious disease modelling is one small part of infectious disease epidemiology, which is a small part of epidemiology. How, then, did modelling come to dominate, not only the prediction of the spread of COVID-19, but also policy decisions with consequences reaching far beyond the death toll of the disease itself?

It is helpful to understand some of the conceptual evolution of epidemiology from its foundations, which were laid during the era of industrialisation in Europe.[1] Cities grew, bringing people into close proximity, many of them malnourished, with poor or non-existent sanitation and hygiene practices. Infectious disease flourished in these newly swollen human ecosystems, and epidemics were a regular occurrence. At the same time, information on disease incidence and deaths became readily available in concentrated form for the first time. Enterprising medical thinkers realised they might infer the causes of outbreaks from this information. Thus two new ways of thinking were born together, with epidemics lending their name to one epidemiology and the state that collected information lending its name to statistics.

It was epidemiology that taught us the health significance of personal hygiene. Epidemiology helped us uncover the 'germ theory' of disease[2], which ultimately turned out to account for infectious diseases and is central to contemporary Western medicine development.

Germ theory and hygiene theory were at odds during this period, with 'hygienists' seeing the germ theory as an attempt to let the authorities off the hook of ensuring more humane living conditions for the labouring classes.[3] The germ theory 'won' in the theoretical sense. The 'miasma' theory of disease was ultimately discredited: diseases are not caused by bad smells, and vaccination proved very effective. However, the recommendations of the hygienists were effective in another sense, because personal hygiene and good sanitation are necessary for people to dwell together in city conditions even with the technology of vaccination.

Fast forward to the mid-20th century, and the advent of antibiotics and other innovations such as ventilation put many infectious diseases even further on the back foot, even though viruses remained stubbornly resistant to direct medical solution. In the 1940s and 1950s, epidemiological attention shifted to another 'epidemic': that of lung cancer, which had rocketed from a virtually unknown condition during the previous two decades. Why?

Through a remarkable methodological development, Sir Austin Bradford Hill and others perceived the significance of an apparently tiny difference between the odds of smoking among people living with lung cancer and among people living with other cancers, which translated into a remarkably large risk ratio.[1] They were able to anticipate and check for an extensive range of plausible confounders. With others, such as Jerome Cornfield[4], they were able to corral evidence from other domains for a causal connection between smoking and lung cancer and against the leading rival hypothesis: the 'constitutional hypothesis' that some gene caused both.

In this episode, the modern discipline of epidemiology was born. Two characteristics are relevant to the present narrative. Firstly, this was a discipline of campaigning. The real 'win' for these epidemiologists was not the scientific case that smoking caused lung cancer, but the recognition in 1964 in the US Surgeon General's report that smoking causes lung cancer.[5] This led to a series of regulations and public health advice – changes that were fought by tobacco companies and continue to be fought in many Asian countries today.[6]

James Lind had sought to convince the naval authorities to provide limes to sailors against scurvy, eventually succeeding (hence the term 'limeys' for British sailors). Less successfully, the problematic and abrasive Ignaas Semmelweis had sought to instigate hand washing in the General Hospital in Vienna. In London, John Snow persuaded authorities to remove the water pump handle in Broad Street, which was drawing water from the polluted River Thames to create an epicentre of cholera in the district.[1]

But with the smoking and lung cancer episode, the political engagement of epidemiology was affirmed for chronic diseases. Contemporary epidemiologists continue not only to seek the scientific truth about, for example, sugar and all-cause mortality but also to campaign for sugar taxes.

The second feature of the Bradford Hill story is its informality, immortalised in nine 'viewpoints' for assessing causality.[7] Bradford Hill urged epidemiologists to consider causal hypotheses from various perspectives. How strong is the association? Is a causal link biologically plausible? Does evidence from basic sciences support it? Is there a dose-response relationship? And so forth. These were subsequently interpreted as a checklist and remain in use today. But that was never the intention. They were not meant as *sine qua non* for causal inference but as guides to the ultimate question: Is there any hypothesis that better explains the evidence than that of cause and effect?

While epidemiology subsequently developed in mathematical complexity, this relatively informal, subjective approach did not. Causal inference remains a stubborn philosophical problem, and so the inability to define and proceduralise it is not surprising. But it is something of a challenge when peers seek to assess each other's work. It is also hard to teach and something of an embarrassment for those who prefer to think of epidemiology as closer to the natural than the social sciences. Efforts to formalise causal inference now form a considerable part of the epidemiological methodological agenda, and are a growth area for the discipline.[8,9]

In infectious disease epidemiology, modelling provides a way to formalise the central question, which concerns predicting the course of a disease, even after the cause is known (as is more often the case now, where it was not a hundred years ago). Modelling makes use of new computing power and enables the consequences of assumptions to be worked out in detail. This highlights the nature and justification of the assumptions themselves, the sensitivity of predictions to those assumptions and inaccuracies in the data – all of which is beneficial. It also enables predictions that are much better supported than would otherwise be the case – provided the assumptions themselves are well supported and the data reliable.

Thus, contemporary epidemiology is influenced by two paradigm-shaping instincts: the sense that campaigning for public health policy is part of the epidemiological mandate and the desire for methodological progression within the science towards more formal approaches.

Models in the making

All models are wrong, but some are useful.

George E. P. Box

During the COVID-19 pandemic, the tradition of campaigning and the associated sense of urgency may have contributed towards some unfortunate lapses in the use of models. Models are abstract representations of real phenomena, and are useful for making predictions. At best, a good model has two facets: accuracy and simplicity. The accuracy is vital in linking the model to reality, while simplicity is paramount for understanding. Despite their usefulness, models are always shrouded with limitations. We discuss some of these in the present section.

Two important considerations – assumptions and reliable data – were often not transparently communicated or verified, which naturally had an impact on the effectiveness of South Africa's response which mathematical models primarily influenced. A significant number of these models has been published globally since the beginning of the pandemic.[10,11] Given that there is no known effective pharmaceutical treatment for COVID-19 (at the time of writing), mathematical models have shaped policy with respect to non-pharmaceutical interventions, intending to limit transmission between persons and contaminated environments, and in so doing 'flatten the curve' of infected persons.

Lessons learnt from the SARS outbreak in 2003 and the MERS outbreak in 2002 provided a medical understanding of how coronaviruses affect the lower respiratory tract. Taiwan was one of the first countries to implement the non-pharmaceutical interventions learnt from SARS. Some of these strategies include the wearing of masks and contact tracing. Modelling helps us to quantify the impact of these preventative measures on the spread of the disease.

From a modelling perspective, the integrity of a mathematical model is in its assumptions, consanguineousness to the available data, and the power to predict the epidemic trends in the short or long term. In building mathematical models for the pandemic in South Africa, it is crucial to consider the following: heterogeneity in the population densities, economic realities, inconsistent policies and inharmonious enforcement of regulations. Model-building is thus not an abstract exercise but requires deep contextualised knowledge. In South Africa, appreciation of socio-economic dynamics is critical in the modelling process, such as overcrowding, a large informal sector, and high levels of poverty. Within this context, social distancing in South Africa, for instance, should be understood within the context of a heterogeneous distribution of populations and varied patterns of movement within and between cities and provinces. These social realities have implications for the pandemic's spread as densely populated areas, such as public transport hubs and spaces, are hotspots for transmission. The resurgence of the epidemic in many countries has seen the emergence of a more recent dynamic termed a 'superspreader event', threatening the fragile equilibrium South Africa has achieved. Furthermore, as prevention fatigue sets in, the relaxation of preventative efforts can be a source of disease recrudescence. Models that capture such scenarios are of interest from a policy formulation and disease management perspective.

There is always a trade-off between model complexity and its tractability – the more complex the model, the less tractable and vice-versa. Many of the recent models had a few noticeable challenges when it came to functioning as workable solutions. Firstly, the overestimation of predicted numbers led to panic amongst the public, given a poor understanding among most that, while models are useful tools, they should not be over-interpreted, especially when considering long-term projections. Being dynamic, the implementation of an intervention of any kind will immediately impact the progression and trajectory of the disease described by the model. Secondly, many models have been built to provide predictions for scenario planning without clearly explaining the underlying assumptions which inform these predictions. Lastly, models depend on assumptions, and the sensitivity to errors in these assumptions should be aligned to the social-economic dynamics of a given setting to create realistic outcomes.

Given the complexity of a functioning society with varied dynamics, models should ideally be interdisciplinary. The role of social, cultural and human behaviour and economic consequences of the pandemic and any possible interventions cannot be ignored when modelling a pandemic. Thus, while the role of mathematical models as tools for understanding the transmission dynamics of COVID-19 in South Africa cannot be underestimated, one thing needs to be kept in mind: all models are necessarily approximations of the real world, being simplifications of reality driven by the need to answer specific questions and in many cases one particular question. Models capture certain aspects of a phenomenon, under certain assumptions, while relying on relevant data sets where the quality, accuracy, specificity, availability and usability of the data are key to the usefulness of a model.

The big difference in 2020/2021 has been the impact of social media and greater transparency, which meant that modelling as a tool took a leading role in combatting COVID-19. Scientists involved in the modelling of MERS/SARS did not benefit from exposure to social

media. During previous disease outbreaks, only mathematicians had access to resources to do the modelling, and often the products of such engagement were kept within the academic community. The power of modern computers allows anyone with basic knowledge to develop models of the spread of COVID-19. The effective use of social media provides for wide dispersal of the forecasts and information learnt from such models.

Data-driven models

There have been ongoing concerns about the quality and availability of data relating to the current pandemic.[12] Yet, aside from being employed for the development of predictive models, these questionable data are underlying some of the most critical metrics that are being used to gauge our progress in fighting the novel coronavirus. Hugely debated results published in a preprint in April 2020 present two antibody tests conducted by universities in California. It was claimed that possibly 28 to 85 times more people had been exposed to COVID-19 than had been detected using the PCR method.[11] If so, a slight alteration in the relevant denominator used to calculate the mortality rate from COVID-19 would have shifted mortality from a figure of 2.5–3%, which public health officials had been working with, to between 0.12% and 0.20%. Given that seasonal influenza's mortality rate is about 0.1%, it is clear that such a change in the mortality rate describes an entirely different pandemic.[13]

Recent events seem to support the argument that we may be overestimating the infection fatality rate of COVID-19. In September 2020, it was indicated that upon taking into account asymptomatic cases, the infection fatality rate had shifted from between 2% and 3% to 1%.[14] More recent updates by the CDC and WHO indicate an infection mortality rate of between 0.65% and 0.5–1%, respectively, while the work by Ioannidis[15] indicates a median infection fatality ratio of 0.23% across 51 locations. The conclusion to be reached is that when there are errors in raw data or our data sets are limited and constantly changing, we need to be cautious about reaching conclusions regarding the nature of a phenomenon. Inappropriate or inaccurate statistics can generate an image that differs vastly from the reality they are trying to capture.

It is no secret that, similar to mathematical model development, data science relies heavily on assumptions made in the scientific process. One of these assumptions is that the data set used in the scientific process represents the studied entity/population. Hence, when data are skewed, limited, or contextually inappropriate, the results of the whole scientific process are most likely to be incorrect. The idea of 'wrong data' has been identified as artificial intelligence's (or AI's) biggest risk factor.[16]

There are concerns regarding the quality of data from all countries, often for similar reasons.[12] Instead of national health surveillance systems that can be relied upon to provide reasonably accurate data, there is a patchwork of voluntary data-gathering processes in place at most hospitals. Naturally, not all hospitals report the data, and the data are not consistent from hospital to hospital.[12] Furthermore, delays in obtaining data from hospitals and other health facilities lead to data that do not reflect the current situation on the ground.[17] When models employ such data for predictive purposes, they end up *predicting the past*. The use of contextually inappropriate data to design a model is also concerning. An example of this is the use of data from China to predict the spread of South Africa's epidemic. Where machine learning has been used with inappropriate data, the results that have been obtained are misleading. Yet the implementation and reporting of results obtained from machine-learning models are on the increase. Why is that?

Conclusion

Epidemiology has a long and fine tradition of engaging public policy to change it. The discipline also naturally seeks technical development of its methodology. In the case of COVID-19, these two instincts – campaigning and formalising – came together in an unfortunately unholy alliance. As has been remarked, infectious disease modelling is only one part of infectious disease epidemiology; but it is particularly striking because it appears to represent methodological progression. When the campaigning instinct kicks in, there is a danger of overreach. Policies are pushed that simply fail to consider all factors because the models do not consider all factors. Both epidemiology and public health can do better in the future by considering a more extensive range of health consequences beyond the deaths of people with the virus in their bloodstream, by being less belligerent about the importance of these consequences above other policy priorities. Humility is an epistemic virtue.

As current data on the coronavirus are not reliable in the sense that we are constantly adjusting for inaccuracies or obtaining new information, our aim should rather be to create opportunities for further research on the novel coronavirus in the future. As such, our goal should be to create data repositories, structure data cleaning processes, design data pipelines, and develop better tools to model past pandemics to gain a deeper understanding for further comparison. Learning from the past should be our strategy to become a society that reflects on past mistakes, assesses current inadequacies, and then moves forward with greater awareness and humility.

Acknowledgements

E.M. acknowledges support from the National Research Foundation of South Africa (grant no. 103483).

Competing interests

We have no competing interests to declare.

References

1. Morabia A. History of epidemiologic methods and concepts. Basel: Birkhauser Verlag; 2004. https://doi.org/10.1007/978-3-0348-7603-2
2. Broadbent A. Why philosophy of epidemiology? Philosophy of Epidemiology. London: Palgrave Macmillan; 2013. https://doi.org/10.1057/9781137315601
3. Carter KC. The rise of causal concepts of disease: Case histories. London: Routledge; 2003. https://doi.org/10.4324/9781315237305
4. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. Int J Epidemiol. 2009;38:1175–1191. https://doi.org/10.1093/ije/dyp289
5. United States Surgeon General's Advisory Committee on Smoking. Smoking and health: Report of the Advisory Committee to the Surgeon General of the Public Health Service [webpage on the Internet]. c1964 [cited 2021 Aug 07]. Available from: https://profiles.nlm.nih.gov/101584932X202
6. Broadbent A, Hwang S-S. Tobacco and epidemiology in Korea: Old tricks, new answers? J Epidemiol Community Health. 2016;70:527–528. https://doi.org/10.1136/jech-2015-206567
7. Bradford Hill A. The environment and disease: Association or causation? Proc R Soc Med. 1965;58:259–300. https://doi.org/10.1177/003591576505800503
8. Rubin DB. Causal inference. In: Tierney R, Peterson P, Baker E, McGaw B, editors. International encyclopedia of education. 3rd ed. Amsterdam: Elsevier; 2010. p. 66–71. https://doi.org/10.1016/B978-0-08-044894-7.01313-0
9. Van der Weele TJ. Explanation in causal inference: Methods for mediation and interaction. Oxford: Oxford University Press; 2015.
10. Enserink M, Kupferschmidt K. Mathematics of life and death: How disease models shape national shutdowns and other pandemic policies. Science Insider. 2020 March 25; Health. https://doi.org/10.1126/science.abb8814
11. Chowdhury R, Heng K, Shawon MSR, Goh G, Okonofua D, Ochoa-Rosales C, et al. Dynamic interventions to control COVID-19 pandemic: A multivariate prediction modelling study comparing 16 worldwide countries. Eur J Epidemiol. 2020;35:389–399. https://doi.org/10.1007/s10654-020-00649-w
12. Irfan U. How common is Covid-19? What 2 controversial antibody studies can and can't tell us. Vox. 2020 April 24 [cited 2021 Aug 07]. Available from: https://www.vox.com/2020/4/24/21229415/coronavirus-antibody-testing-covid-19-california-surve
13. Woodie A. How the lack of good data is hampering the COVID-19 response. Datanami. 2020 April 27 [cited 2021 Aug 07]. Available from: https://www.datanami.com/2020/04/27/how-the-lack-of-good-data-is-hampering-the-covid-19-response/

14. Yong SJ. Clarifying the true fatality rate of Covid-19: Same as the flu? Medium. 2020 September 08 [cited 2021 Aug 07]. Available from: https://medium.com/microbial-instincts/clarifying-the-true-fatality-rate-of-covid-19-same-as-the-flu-8148e38b9ab5

15. Ioannidis JPA. Infection fatality rate of COVID-19 inferred from seroprevalence data. Bull WHO. 2020 October 14. https://doi.org/10.2471/BLT.20.265892

16. Korolov M. A.I.'s biggest risk factor: Data gone wrong. InsiderPro. 2018 February 13 [cited 2021 Aug 07]. Available from: https://www.idginsiderpro.com/article/3254693/ais-biggest-risk-factor-data-gone-wrong.html

17. Short A. Reporting delays and political pressure contribute to the CDC's incomplete picture of COVID-19. Business Insider. 2020 May 19 [cited 2021 Aug 07]. Available from: https://www.businessinsider.com/why-the-cdc-coronavirus-data-is-so-bad-2020-5?IR=T