Better modelling of infectious diseases: lessons from covid-19 in China

More timely, accurate, and relevant data and methodological innovation could exploit the full power of modelling, argue **Feng Chen and colleagues**

Since Daniel Bernoulli studied smallpox inoculation from a mathematical perspective in 1760, mathematical models have proved invaluable to understanding and helping control infectious disease epidemics.[1] By simplifying real world phenomena to limited numbers of settings, transmission dynamics modelling uses mathematical models to describe, analyse, and predict infectious disease transmission dynamics and to produce tractable solutions in the face of quickly changing situations. Covid-19 has spread across the world since December 2019, causing millions of deaths and substantial economic losses.

Role of modelling

Since the start of the pandemic in China, transmission dynamics models have been at the forefront of understanding, predicting, preventing, and controlling the situation. Objectives include identification of epidemiological features to understand the disease, prediction of trends in disease, evaluation of control measures to inform decision making, and exploration of uncertainty.

Identification of epidemiological features

At the beginning of the outbreak, when almost nothing was known of the novel pathogen, the models explored the virus's crucial epidemiological features, such as the incubation period—the period between exposure to the pathogen and the appearance of the first symptoms—and the basic reproductive number (R_0)—the average number of secondary infections generated by the first infectious individual in a population of susceptible individuals. These key parameters helped advance our understanding of the features of the disease that we have not yet fully understood and realise the severity of the situation.[2,3]

KEY MESSAGES

- Mathematical modelling can help us understand and control infectious disease outbreaks, including the covid-19 pandemic
- Accuracy of prediction is limited by insufficient, inaccessible, or inaccurate data
- Greater information sharing and methodological innovation to deal with uncertainty are needed to improve accuracy
- Nevertheless, transmission modelling is a powerful tool for early warning and short term predictions

Short term prediction

As more data became available, these models could be fitted by the actual data and steadily refined to improve prediction of future trends, such as infection numbers and hospitalisation needs.[4] Models proved useful in predicting short term trends, on the scale of days or weeks, which was one of the major tasks of the Covid-19 Prevention and Control Expert Committee in February 2020 organised by the Chinese Preventive Medicine Association. These predictions allowed response teams to allocate healthcare resources efficiently and optimise containment strategies.

Evaluation of control measures

Because successful public health measures will change the course of an epidemic within days, by comparing the observed and predicted infection trends these models helped to quantitatively assess the effectiveness of the prevention and control measures. For example, the models made a great contribution to the Wuhan shutdown and national emergency response to delay the spread of the epidemic and averted high numbers of cases in China.[5] The models reflected the implementation of clinical diagnostic criteria and universal symptom survey to epidemic control in Wuhan.[6]

Exploration of uncertainty

Facing fast changing situations, the models are naturally designed to explore uncertainties by sensitivity analysis incorporating different parameters. For instance, based on current understanding of virus control strategies and vaccine effectiveness, the mathematical models warned that completely lifting the non-pharmaceutical interventions, even with highly effective vaccines, would lead to a substantial increase in covid-19 transmission.[7] Modelling results provide rapid feedback to inform future decision making.

Limitations

However, despite their extensive use in this pandemic, mathematical models have several important limitations, as we discuss below.

We cannot model what we do not understand

Key information required for modelling includes duration of incubation period, transmission route and transmissibility of the pathogen, and difference in transmissibility of cases during the incubation period and symptomatic period, which can be obtained from real data, previous experience, or expert opinions. The genome of the novel virus was sequenced on 2 January 2020, and shared with the global community nine days later.

Although the virus, named SARS-CoV-2 by the International Classification on Taxonomy of Viruses, was soon identified as belonging to the beta-coronavirus family, crucial epidemiological characteristics remained largely unclear. In the absence of data, scientists had to rely on similar respiratory infections, such as severe acute respiratory syndrome and Middle East respiratory syndrome to inform model design.[3] Many of these models ignored the virus's incubation period—the gap between infection and development of symptoms [8]—or underestimated its length as two to three days, as with severe acute respiratory syndrome. Other models assumed that transmission during the incubation period was zero or was equal to transmission during the symptomatic stage; both assumptions proved false.[3,9]

While several early epidemiological studies did provide key information that improved model accuracy,[10,11] our limited understanding of the new virus resulted in models with inappropriate structures and unverified parameters, which produced inherently flawed predictions.

Models are less powerful if data are inaccessible

If there is one thing more important than vaccines in this pandemic, it is data. Data are urgently needed not only on daily confirmed cases but also on transmission dynamics, population migration, individual symptoms, hospital admissions, treatment records, and contact tracing. Longitudinal data are particularly necessary for better understanding of the impact of covid-19 on population health and on health systems. While global response teams have exabytes of data, much of the data are kept in private databases. Most of the mathematical models developed so far are based largely on daily laboratory confirmed case numbers, with only a few incorporating population movement or migration data to reveal how the virus spread across the world.[12 13] These exceptions show the importance of data availability in understanding and combating the pandemic.

For us to exploit the full power of modelling, more relevant and accurate data must be made accessible. Individual databases should be combined, becoming greater than the sum of their parts and offering novel insights. A global, open minded sharing of data, combined with big data approaches and high speed network technologies, will help us to find better solutions to the covid-19 pandemic and to future epidemiological events, while protecting personal privacy and social security.

Accurate data are essential to truly understand the pandemic

In late January 2020, in the early stage of the epidemic in China, polymerase chain reaction (PCR) testing capacity was insufficient throughout the country; the situation was even more serious in Wuhan, the centre of the pandemic in China. This insufficiency resulted in a considerable delay between symptom onset and laboratory confirmation of infection. For infections before 22 January, average delays were 17.9 days in Wuhan, 15.8 days in the cities of Hubei province excluding Wuhan, and 12.7 days in China excluding Hubei (fig 1). These delays, though greatly reduced as testing capacity expanded, still averaged about three to seven days in February. Similar delays were seen in Germany.

According to a systematic review,[14] the 33 models of the covid-19 pandemic in China, with few exceptions, relied on officially released numbers of laboratory confirmed cases rather than on numbers of symptomatic individuals.[15] These models should therefore be interpreted with caution. One study incorporating dates of both symptom onset and laboratory confirmation, obtained from the National Notifiable Disease Report System database, estimated a peak effective reproductive number (R_t, or the mean number of people infected by a single infectious individual in an infection period) of about 3.54, considerably higher than the value (R_0=2.38) from the study relying on laboratory confirmed case numbers.[15] In addition, another study using the same data estimated the epidemic's turning point (that is, the point at which the daily emerging case rate began to decelerate) as 31 January, about nine days earlier than models incorporating laboratory confirmed case numbers.[4]

Insufficient testing capacity during the pandemic's early stages also resulted in a considerable number of unconfirmed infections—cases with typical symptoms or radiological evidence but without a positive PCR test or without PCR testing. We previously reported that unconfirmed infections accounted for about 40% of all cases in Wuhan in January 2020.[6] A model incorporating individual level data with symptom onset information and accounting for presymptomatic infectiousness estimated that 87% of all infections in Wuhan from 1 January to 8 March were unconfirmed, potentially including asymptomatic and mildly symptomatic individuals.[15]

While this estimate may seem high, it is supported by recent infection rate data. The latest large scale seroprevalence study in Wuhan found a 6.92% positive rate for pan-immunoglobulins against SARS-CoV-2 in the population,[16] equivalent to about 0.73 million infections, of which 94% were unconfirmed during the pandemic in Wuhan. This number is far beyond the public's imagination and all existing model predictions, though there is a possibility of cross reaction with antibodies of other coronaviruses, which may result in an overestimated infection rate.

Models must keep up with a rapidly changing situation

Accurate prediction of the pandemic using models is a seemingly impossible task. Time varying control measures, continually updated treatment protocols, increasing public health consciousness, and vaccination all affect the pandemic's trajectory. Without incorporating fast changing parameters, models would result in less accurate predictions. Two major factors may have shaped the pandemic trend: vaccination and viral mutation. Vaccination promises to end the covid-19 pandemic while allowing restoration of social activities. As of 13 September 2021, China had over 969 million people fully vaccinated against SARS-CoV-2, and about 2.3 billion people worldwide were fully vaccinated,[17] reducing the number of infections, critically ill cases, and deaths. However, a broad list of real world factors could impact on vaccine effectiveness, including immune response heterogeneity, financing shortfalls, regional inequalities, logistical challenges, difficulties in expanding manufacturing capacity, and viral mutations. Future modelling studies should account for these factors to provide more reliable results to inform decision making.

Viral mutation is facilitated by the pathogen's large scale spread. More virulent and transmissible variants, such as the delta variant, which is 40% to 60% more contagious than previous variants,[18] alter vaccine effectiveness and global infection dynamics greatly. How the models account for such viral mutation possibilities, especially before it is widespread, to better predict the future is challenging.

In addition, the dominant viral transmission route has changed from potential animal-to-human transmission at the beginning of the outbreak to broad human-to-human transmission. On 30 November 2020, routine PCR testing among the staff at an aquatic product company in Jiaozhou, Qingdao, China, identified one asymptomatic positive case; further contact tracing identified an additional infection among the co-workers. After scientists from the Centre for Disease Control and Prevention comprehensively studied the tracing investigation, gene sequencing, and video records, the case was recognised as a potential contaminated cold chain product-to-human transmission.[19] Going forward, scientists must continually incorporate new information into their models, while also ensuring that this information is reliable.

Perspective

In 1976, the British statistician George Box famously stated that all models are wrong, but some are useful. Over four decades later, referring specifically to covid-19 modelling, Siegenfeld and colleagues noted that understanding what models cannot predict is sometimes more important than understanding what they can predict.[20]

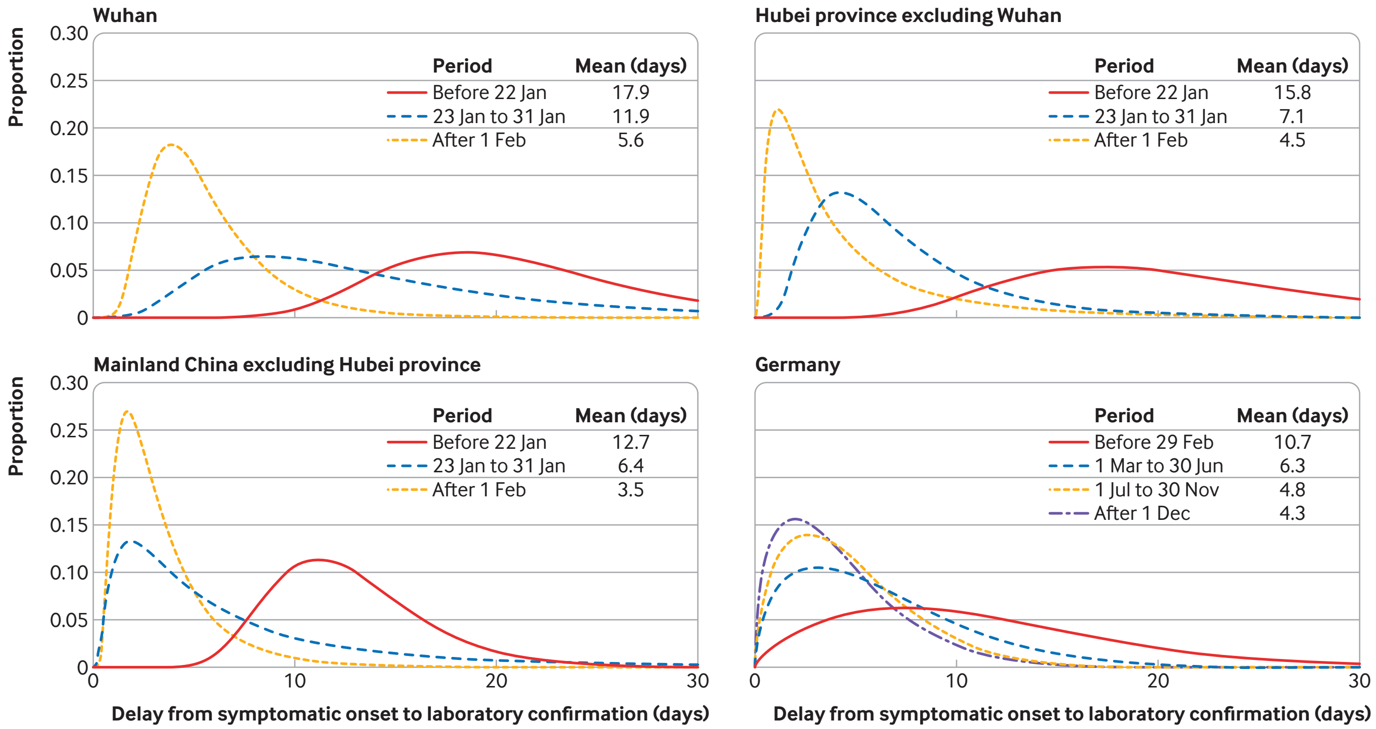

Fig 1 | Distribution of the delays between symptom onset and laboratory confirmation of viral infection. Distributions were fitted in (A) Wuhan, (B) Hubei province excluding Wuhan, (C) mainland China excluding Hubei province, and (D) Germany, using the statistical simulation incorporating lognormal distribution and daily frequency data (extracted from[10 11 15] and https://global.health)

Given the highly dynamic nature of disease outbreaks, even models that account for rapidly changing parameters cannot predict future numbers with total accuracy. Meanwhile, long term predictions of mathematical models have mostly proved badly wrong. In fact, if effective interventions are in place, changing the epidemic's dynamics, these predictions are bound to be wrong; they will be correct only in the absence of such interventions.

Despite these limitations, mathematical modelling is one of our most powerful tools for detecting, understanding, and combating infectious disease outbreaks. As stated above, a model is only as reliable as the data underlying it. Increased amounts of data, through improved case identification methodology and expanded information sharing, is crucial for models to effectively recognise and mitigate future public health emergencies. On the other hand, model optimisation and methodological innovation are urgently needed to deal with the imperfect data to give early warning of major public health emergencies. Importantly, mathematical modelling should be one of the most valuable tools to reflect great uncertainties or warn of the worst situation. An appreciation of the shortcomings of models not only clarifies what they can't do but helps anticipate what they can do.

Contributors and sources: FC, president of the China Statistical Theory and Methodology Committee and vice president of the International Biostatistics Society China Branch, conceived, designed, and supervised this analysis. YH, professor president of the China Biostatistics Society, helped interpret the second arguments in this analysis and critically reviewed the manuscript. QJ critically reviewed and helped interpret the arguments in this analysis. YZ helped interpret the arguments in this analysis and critically reviewed and revised the manuscript. FS helped interpret the arguments in this analysis and critically reviewed and revised the manuscript. YW performed analysis, reviewed the literature, interpreted results, and drafted the manuscript. YW, YZ, QJ, YH, and FC served as the members in the covid-19 prevention and control expert committee organised by the Chinese Preventive Medicine Association. All material to support the arguments in this analysis were from publications or publicly accessible resources. FC is the guarantor.

The study was partially supported by the National Natural Science Foundation of China (82041024 to Feng Chen) and the Bill & Melinda Gates Foundation (INV-006371).

Competing interests: We have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Commissioned; externally peer reviewed.

This article is part of a collection proposed by the Peking University Centre for Public Health and Epidemic Preparedness and Response. Open access fees were funded by individual institutions. *The BMJ* commissioned, peer reviewed, edited, and made the decision to publish. Li-Ming Li advised on commissioning for this collection. Jin-Ling Tang, Di Wang, and Kamran Abbasi were the lead editors for *The BMJ*.

Yongyue Wei, associate professor of biostatistics[1]

Feng Sha, associate professor of biostatistics[2]

Yang Zhao, professor of biostatistics[1]

Qingwu Jiang, professor of epidemiology[3]

Yuantao Hao, professor of epidemiology and biostatistics[4]

Feng Chen, professor of biostatistics[1]

[1]Department of Biostatistics, School of Public Health, Center of Global Health, China International Cooperation Center for Environment and Human Health, Nanjing Medical University, Nanjing 211166, China

[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[3]School of Public Health, Fudan University, Shanghai 200433, China

[4]Department of Medical Statistics and Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou 510080, China

Correspondence to: F Chen
fengchen@njmu.edu.cn

1 Dietz K, Heesterbeek JA. Daniel Bernoulli's epidemiological model revisited. *Math Biosci* 2002;180:1-21. doi:10.1016/S0025-5564(02)00122-0

2 Wang H, Wang Z, Dong Y. Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China. *Cell Discov* 2020;6:10. doi:10.1038/s41421-020-0148-0

3 Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* 2020;395:689-97. doi:10.1016/S0140-6736(20)30260-9

4 Wei YY, Lu ZZ, Du ZC. [Fitting and forecasting the trend of COVID-19 by SEIR($^{+CAQ}$) dynamic model]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2020;41:470-5.

5 Tian H, Liu Y, Li Y. An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in China. *Science* 2020;368:638-42. doi:10.1126/science.abb6105

6 Wei Y, Wei L, Jiang Y. Implementation of clinical diagnostic criteria and universal symptom survey contributed to lower magnitude and faster resolution of the covid-19 epidemic in Wuhan. *Engineering (Beijing)* 2020;6:1141-6. doi:10.1016/j.eng.2020.04.008

7 Yang J, Marziano V, Deng X. Despite vaccination, China needs non-pharmaceutical interventions to prevent widespread outbreaks of covid-19 in 2021. *Nat Hum Behav* 2021;5:1009-20. doi:10.1038/s41562-021-01155-z

8 Dehning J, Zierenberg J, Spitzner FP. Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science* 2020;369:eabb9789. doi:10.1126/science.abb9789

9 Kucharski AJ, Russell TW, Diamond CCentre for Mathematical Modelling of Infectious Diseases COVID-19 working group. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *Lancet Infect Dis* 2020;20:553-8. doi:10.1016/S1473-3099(20)30144-4

10 Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19). 2020. http://www.asean-china-center.org/english/2020-03/4756.html.

11 Special Expert Group for Control of the Epidemic of Novel Coronavirus Pneumonia of the Chinese Preventive Medicine Association. [An update on the epidemiological characteristics of novel coronavirus pneumonia (covid-19)]. *Zhonghua Liu Xing Bing Xue Za Zhi* 2020;41:139-44.

12 Peirlinck M, Linka K, Sahli Costabal F, Kuhl E. Outbreak dynamics of covid-19 in China and the United States. *Biomech Model Mechanobiol* 2020;19:2179-93. doi:10.1007/s10237-020-01332-5

13 Linka K, Peirlinck M, Sahli Costabal F, Kuhl E. Outbreak dynamics of covid-19 in Europe and the effect of travel restrictions. *Comput Methods Biomech Biomed Engin* 2020;23:710-7. doi:10.1080/10255842.2020.1759560

14 Guan J, Wei Y, Zhao Y, Chen F. Modeling the transmission dynamics of covid-19 epidemic: a systematic review. *J Biomed Res* 2020;34:422-30. doi:10.7555/JBR.34.20200119

15 Hao X, Cheng S, Wu D, Wu T, Lin X, Wang C. Reconstruction of the full transmission dynamics of covid-19 in Wuhan. *Nature* 2020;584:420-4. doi:10.1038/s41586-020-2554-8

16 He Z, Ren L, Yang J. Seroprevalence and humoral immune durability of anti-SARS-CoV-2 antibodies in Wuhan, China: a longitudinal, population-level, cross-sectional study. *Lancet* 2021;397:1075-84. doi:10.1016/S0140-6736(21)00238-5

17 Coronavirus Resource Center, John Hopkins University and Medicine. Summary of vaccination statistics. 13 Sep 2021. https://coronavirus.jhu.edu/vaccines/international.

18 Burki TK. Lifting of covid-19 restrictions in the UK and the delta variant. *Lancet Respir Med* 2021;9:e85. doi:10.1016/S2213-2600(21)00328-3

19 Qingdao Health Commission. The recent epidemic prevention and control situation of Qingdao. 2020. http://wsjkw.qingdao.gov.cn/n28356065/n32563056/n32563057/200928142349396662.html

20 Siegenfeld AF, Taleb NN, Bar-Yam Y. Opinion: What models can and cannot tell us about covid-19. *Proc Natl Acad Sci U S A* 2020;117:16092-5. doi:10.1073/pnas.2011542117

Cite this as: *BMJ* 2021;375:n2365
http://dx.doi.org/10.1136/bmj.n2365