The story of the R number

How an obscure epidemiological figure took over our lives

Part 2: Modelling

How do you build a complex epidemiological model in record time with little or no reliable data? In the second instalment of his six-part series, **Gavin Freeguard** describes how different modelling groups in the UK used different data sources and assumptions to try to understand Covid-19 infection rates, and how this diversity proved to be a strength rather than a weakness

lakshmiprasad S/iStockPhoto.com

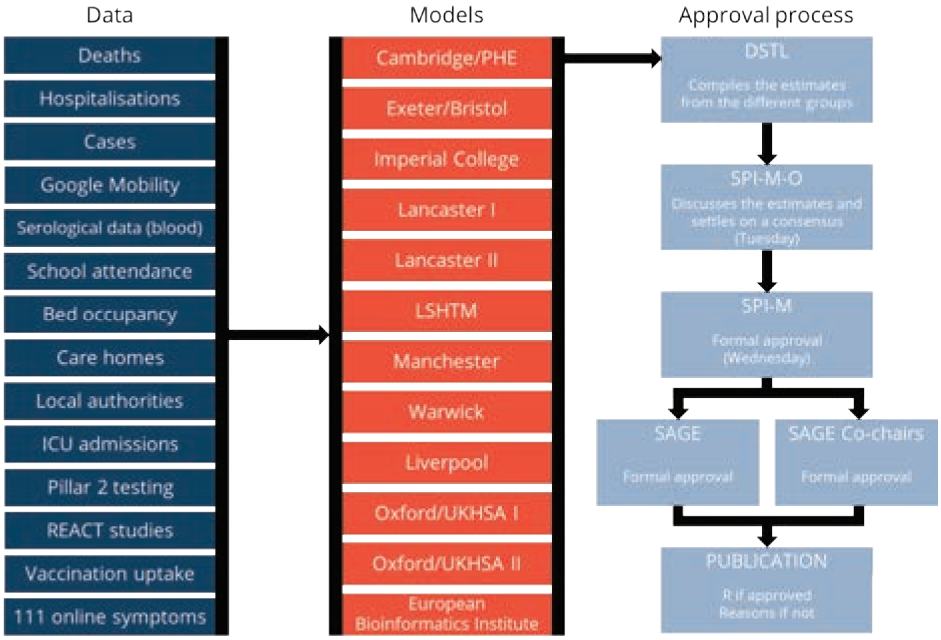

Figure 1: The data modelling teams and approval process for publishing the Covid-19 R number

> The model by itself has no significance; it proves nothing and explains nothing, except perhaps the mind of the man [*sic*] who made it. Its value lies in its potential use as a tool for understanding the patterns in which the disease occurs.
>
> (George Macdonald[1])

Since R cannot be measured directly, it can only be modelled. The guide to epidemiological modelling produced by the UK Health Security Agency (UKHSA) defines a model as "a simplified representation of reality". Epidemiological models aim to understand how a disease might spread, how it might affect the population and the public services designed to treat them, and, in some cases, the projected impact of different interventions (such as lockdowns). Data is entered into a model, which is set up to make certain assumptions – about the population and their behaviour, about the disease. The results are uncertain, due to varying data quality, varying assumptions, and the impossibility of replicating reality, let alone predicting the future (some models attempt to estimate R in the past or in the present, while others attempt to forecast what R will be under different future circumstances). Governments and scientists therefore try to use a number of different models, employing different data and making different assumptions, to understand what may be happening now and what may happen in the future.

The different R modelling groups used a huge range of data (see Figure 1). There was data on deaths and hospitalisations from, and cases of, Covid. For much of the pandemic, this came from Public Health England (PHE), now part of the UKHSA. Some modelling groups broke these down by geography – for example, cases at local authority level – and other characteristics, such as hospitalisations by age range. Others used data on swab tests, vaccination uptake and even serological data – testing samples of leftover blood from the National Health Service (NHS) Blood and Transplant Service to understand the level of Covid in the population. There was data from care homes, on intensive care admissions, and reports of symptoms to the NHS's 111 online service.

Then there was data about the behaviour of the population and the possible spread of the disease. Several modelling groups used Google Mobility data – "created with aggregated, anonymised sets of data" from users who have a Google account on their mobile device and "have turned on the Location History setting". Some used school attendance data in England, a data set the Department for Education published for the first time in April 2020 (bit.ly/42M24R6), based on a form for schools to fill out daily. There were Real-time Assessment of Community Transmission (REACT) studies, coordinated by Imperial College and market research company Ipsos MORI, which involved hundreds of thousands of swab samples. All of these data sets require their own supporting infrastructures, whether survey teams, laboratory capacity, modelling methods or computational power. The same is true of other data sets key to government's understanding of Covid-19, such as the Office for National Statistics' weekly Covid-19 infection survey (started in April and May 2020, aiming to sample swabs from 180,000 people each fortnight and blood from 150,000 people a month) or the CoMix survey of social contacts (which recorded "101,350 observations from 19,914 participants who reported 466,710 contacts" between March 2020 and March 2021; bit.ly/42QLjEk). As the pandemic went on, UKHSA even started to

Gavin Freeguard is a freelance consultant specialising in data, an associate at the Institute for Government, policy associate at Connected by Data and special adviser at the Open Data Institute. He was originally commissioned by Understanding Patient Data to develop these articles.

use novel sources like waste water to track disease spread (bit.ly/42LhKUX).

Right at the start of the pandemic, authorities in the UK lacked most of this data. The modelling groups had to look elsewhere. Imperial's Neil Ferguson told Parliament, "we relied mostly on extrapolating data from China. We had very little data on what was going on in the UK." "There wasn't any data flow … A lot of the early part was just getting the data," says the University of Exeter's Rob Challen. "Getting the data and then understanding the biases and how to improve them is better than not getting the data at all from our perspective." At that point, the best source of case, admission and death data was a repository on the GitHub platform run by a member of the public, who was scraping information from various coronavirus websites. Some groups turned to the Our World in Data website, based at the University of Oxford, and other crowdsourced data. Challen, a clinician turned maths PhD student working on research repositories for clinical data, was "Covidified" in March 2020 after a colleague messaged on collaboration platform Slack. It read something like: "hi everyone, general question: trying to trace down access to SUS [Secondary Uses Service, a data set on hospital admissions, appointments and attendances from NHS England] for ages, hitting a brick wall."

The lack of data on Covid tests was a particular challenge – tests ceased, apart from in hospitals, on 12 March 2020. "You couldn't rely on the case data," says Imperial's Samir Bhatt, "so the only way to understand the pandemic was via deaths." Rosalind Eggo of the London School of Hygiene & Tropical Medicine (LSHTM) says, "if you're only seeing the tip of the iceberg – hospitalisations, deaths – you don't know how broad that iceberg is". The Prime Minister's former adviser, Dominic Cummings, highlighted this in his evidence to parliament: "Once you are looking at ICU numbers as your leading indicator, you know that you are in a world of trouble." The first data on cases only started coming through to the modelling groups 6 days before testing was halted, according to parliamentary evidence from Patrick Vallance (bit.ly/3T8HQha; PHE, the Department of Health and Social Care (DHSC) and the modellers had agreed data sharing protocols on 17 February).

The modelling groups and their approaches

Who were the modellers? DHSC first published a comprehensive methodology (bit.ly/3UKyFoF) for producing the R number in April 2021, which listed several groups, all working as part of the expert Scientific Pandemic Influenza Group on Modelling (SPI-M) advising the government. By early 2022 the list included the University of Cambridge and PHE, the Universities of Exeter and Bristol, Imperial College London, Lancaster University (with two different modelling approaches), LSHTM, the University of Manchester, the University of Warwick and more recent additions the University of Liverpool, the University of Oxford (with two approaches, developed with the UKHSA), and the European Bioinformatics Institute.

An article in *Philosophical Transactions of the Royal Society B* notes that "in normal times, it is common for complex models to be developed over six months or even several years".[2] During the pandemic, some were stood up in days. "This is just what we do, right?", says Rosalind Eggo. "If there's a new outbreak and we have capacity, we start working on it." It quickly became apparent that Covid-19 "was going to be a very, very big deal". Her colleague Adam Kucharski remembers quick "back of the envelope calculation[s] for R" in early January using whatever data from China they could get, "as pretty much every modelling group on the planet did". This suggested an R for Covid-19 of at least 2.

At this point, SPI-M was more like a network of modelling groups, convened from time to time. Biostatistician Paul Birrell, of the Medical Research Council Biostatistics Unit (MRC-BSU) at Cambridge, mentions his "semi-dormant SPI-M involvement – every quarter, maybe" before everything erupted in early 2020. Like many of the groups, Birrell was not starting from scratch. He was drafted in to help with the swine flu pandemic in 2009 – PHE's predecessor, the Health Protection Agency (HPA), had funded a project to develop pandemic modelling capability, but the swine flu pandemic came in the project's infancy, leading to the HPA asking the MRC-BSU for assistance. Over the next decade, PHE and a team at the MRC-BSU developed a modelling framework, getting bits of funding to progress it, meaning they were as ready as anyone in early 2020.

At that point, the best source of data was a repository on the GitHub platform run by a member of the public

Imperial's Samir Bhatt had also cut his teeth on pandemic flu. There were existing tools they could pick up, including a probabilistic programming platform called Stan. Imperial had lots of computational capacity (Scottish chief medical officer Gregor Smith would tell Parliament that the Scottish government model, based on code from one of Imperial's models, involved a supercomputer taking "about 56 hours to do its calculations"; bit.ly/48pNPCM). It had access to data and "very talented people processing and feeding in data" – but a lot still had to be built from scratch: "Every disease is slightly different, you need to do something bespoke."

When it comes to "building" a model, "some of it is thinking, some of it is actual coding", according to Rosalind Eggo. Most of the thinking had been done in advance, preparing for pandemic flu – "you kind of know your early targets" – but a lot of the data infrastructure was new. Data streams came from "different, unexpected sources. The data pipelines, all of that data engineering, was a big priority for us – we had a dedicated team for months who just processed, cleaned, prepared data for downstream analysis."

The LSHTM team wanted to avoid "reinventing the wheel – nobody has time for it", so they looked at the tools they had and worked out if they "made sense biologically and statistically" for Covid. Having the software tools "ready and robust and easy to use" is "an underappreciated part of preparedness", one that "is really critical for next time" (though pulling old code may still take some time, and will be more difficult for someone who didn't create it). Using standardised tools and programming languages, including R – which, Eggo notes, "is confusing for this conversation!" – means you can quickly "assess, calculate and understand the uncertainty you have in your estimates". She thinks testing these tools as part of any preparedness exercise is essential. Like some of the other groups, LSHTM publishes its package for calculating R (github.com/cmmid): "if you make your

There could be large variability in the estimates from individual groups. Outlying groups "might have to defend or explain"

package and code easy to use, and it's robust, people will use it."

Explaining the terminology

The different variants of models listed on gov.uk include a bewildering array of mathematical jargon. Some groups use a "deterministic age-structured compartmental model". A "compartmental" model means simply that individuals are divided into different compartments or groups, which may be further subdivided. Birrell explains that the Cambridge/PHE compartmental model divided the population into regions, each having its own epidemic model, which was assumed not to interact with any other region. The people within each region can be grouped – for example by age and by infection and vaccination status (susceptible, exposed, infectious or recovered, harking back to Kermack and McKendrick[3] nearly a century before). These models can give a more granular view of the epidemic, identifying trends and answering questions specific to different regions and groups.

The "deterministic" bit means that the equations don't have any randomness in them – the epidemic curve is fixed, the output determined simply by the data put into it and the conditions set. "There's obviously randomness in the world", says Birrell, adding that the model assumes that in large populations this will average out. "Stochastic" models, by contrast, build randomness in – SPI-M Operational sub-group (SPI-M-O) documents explain that "the same input data, conditions and parameter values may lead to different outputs each time. Stochastic models are generally run multiple times and an average of outputs taken."

Then there are "renewal equation" models, like LSHTM's. "Infections that are happening now are the result of infections that were happening a few days ago, that are the result of infections happening a few days earlier, essentially cycling from one generation to the next," explains Kucharski. These equations allow you to "take one step back in time".

The LSHTM model is one of several different types of model "fitted" to other data. Eggo explains: "You have your model that generates an output, such as hospitalisations per day. You have your data – your observations of hospitalisations per day. 'Fitting' means you define a distance between the observation and the model output. Then you change the parameters of the model to decrease the distance between the model and the data. We call a model 'fitted' when we have decreased the distance between the model outputs and the data as much as possible."

Eggo doesn't think any of the models are "better" than the others – it's a strength for the UK to have "lots of different assumptions going on, people fitting to different data sets, or fitting in different ways". Some models focus on producing estimates of the key epidemic numbers – Challen says the distinctive feature of the Exeter/Bristol model is its "simplicity and speed. It's sort of the minimum that you have to do to get an estimate of the reproduction number" which is useful for quick ad hoc analysis. Others are designed to answer more specific questions, and the impact of particular measures. The Royal Society argues that "The diversity of models employed by SPI-M to produce predictions is a strength not a weakness, especially with a novel infectious disease w[h]ere there are many unknowns", and considered SPI-M's approach "a very pragmatic one"[4]. Different models providing similar results can lead to greater confidence in them. Adam Kucharski's book[5] quotes epidemiologist Caroline Buckee quoting author Virginia Woolf: "Truth is only to be had by laying together many varieties of error."

Bringing the modelling results together

At the height of the pandemic, modelling groups would generate their results on Monday and submit them, via a spreadsheet, on Tuesday. These files would be aggregated by the Defence Science and Technology Laboratory (DSTL), an agency of the Ministry of Defence perhaps better known by the location of its headquarters, Porton Down. The DSTL had the necessary computational capacity to do so, using CrystalCast software from Riskaware, a Bristol-based software company specialising in modelling "incidents" and human, environmental and security challenges.

The DSTL performed a "meta-analysis" on all the submissions, accounting for the different modelling methodologies, data and uncertainties to produce a range for R. They applied a "reliability score", using the number of Covid-19 hospital admissions (fewer cases means a less reliable estimate), ranging from 0 (the result could indicate just a small, local, clustered outbreak) to 3 (the estimate likely applies to the whole region). Scores of 0 or 1 are accompanied by a specific caveat on publication by government: readers should take "particular care" in their interpretation of the estimate and it "should not be treated as robust enough to inform policy decisions alone".

SPI-M-O, including all the modellers, would meet over Zoom at 2.30 p.m. on a Tuesday equipped with a huge information pack. Rob Challen describes these meetings as "fairly non-confrontational". There could be large variability in the estimates from individual groups, given the different data sources and assumptions. Outlying groups "might have to defend or explain", according to Paul Birrell. Neil Ferguson told Parliament that while there were "always" differences, they will be "similar enough for there to be a fair degree of confidence that these would not be qualitative differences for policy conclusions" (bit.ly/3wngVWi). Turning points in the epidemic presented particular challenges: Kucharski remembers September 2020, "when some things started to tick up in the early data sets … you're hinging on one or two data points that are suddenly higher than they ought to be, there's lots of discussion about what's more or less reliable". Rosalind Eggo points to the emergence of the Alpha variant: "everything gets put back on the table, everybody has a really open mind, everyone has theories and hypotheses that they go off and test. 'Let's revisit everything almost that we thought before' … everyone is prepared to challenge not only each other, but themselves."

Once agreed, these consensus estimates and other data would go to SPI-M on Wednesday morning. After that, the estimates would go to the Scientific Advisory Group for Emergencies (SAGE) for approval (or, if no meeting were scheduled, to its co-chairs Patrick Vallance and Chris Whitty instead). SAGE member Jeremy Farrar says people "probably think it's 20–30 people who gather a few times a week, most weeks. Actually it's hundreds of people, more if you count the number of people they represent, that work through committees." If approved, the R would be sent to "a group of cross-government recipients before publication" and published on gov.uk on Friday. When the UK-wide R was still being reported, it would also be published on the PHE (later UKHSA) Coronavirus dashboard (searches for results by postcode brought up the relevant regional R for many months afterwards). The consensus statements would be published at some point. If the estimates were not approved, the SAGE co-chairs would advise on whether the estimate would be released, the reasons for non-approval, and next steps (including a statement on gov.uk that the R had not been approved).

In July 2021, responsibility for R moved from the SAGE apparatus to the Joint Biosecurity Centre (JBC), now part of the UKHSA. The JBC was created by the then prime minister, Boris Johnson, in May 2020 out of the chaos the centre of government experienced in the early days of the pandemic – with a data infrastructure unfit for purpose, something was needed to get a grip. The creation of the UKHSA, announced in August 2020, reflected perceived failings by PHE and an attempt to set up "a single command structure to advance" the country's pandemic response. This all might imply more political ownership of R, but nothing much changed for the modellers – "still the same workflow, still meet at a similar time, explaining to a slightly different audience", says Birrell – although the UKHSA's Epidemiology Modelling Review Group took on responsibility for scrutinising and agreeing the consensus statement, rather than SAGE, with approval from the chief executive of the UKHSA.

Reflecting on the process

Several modellers would have welcomed receiving data closer to that originally processed by PHE, including patients' personal characteristics. Eggo says, "If you get ten hospitalisations – critical for future forecasting – lots of people saying 'it's people with pre-existing conditions' – well, what do they have? We don't have that. It gives you a much better idea of how at risk the population is if you have that information. … This is the NHS – a joined-up organisation – if you know who the person is, you can check the GP [general practitioner, i.e. family doctor] record – except we couldn't." The sensitivity of the data raises information governance issues, but – according to Rob Challen – "the question is whether what we're doing for SPI-M, advising the government on what their pandemic response should be, is important enough that the value if we had been able to access detailed person level data, would have outweighed the risks of a researcher using that data maliciously". There were "lots of positive messages" from the National Data Guardian, an advisory body to the DHSC that ensures citizens' confidential information is safeguarded, but "those haven't really fed through to change behaviour". Some modellers mentioned OpenSAFELY, an initiative from the Oxford University DataLab (now the Bennett Institute for Applied Data Science) which provided frameworks for accessing data in a secure way.

R calculations were seldom a surprise because of the openness of the modelling groups

Sir Jeremy Farrar says that in his experience, the public – when asked – "have been incredibly willing to share information and data. They want to know what it's used for, who's using it, [that it's] transparent and [that there's] a social contract around that." Eggo understands patients might be reticent, particularly in light of General Practice Data for Planning and Research (a rushed attempt by the NHS during the pandemic to link together more patient data, which actually led the number of people opting out of sharing their data to double). "Researchers were not able to get insights into the pandemic as we, in the UK, with the NHS, should be able to. … You want to explain that privacy is so important … yes, we have your individual-level data, but we don't care about you as an individual in that data. … Everyone contributes to our understanding through their information just being there. I'm not sure people fully understand the good it can do." Samir Bhatt says, "The more that people give access to data and the more they can be assured of their own security, the better it is for science."

Others reflected on the importance of openness. Paul Birrell says, "Everyone involved became more transparent as the pandemic has gone on", and that openness and public interest "forced us to do better in how we communicate and what it is we communicate". R calculations were seldom a surprise because of the openness of the modelling groups, according to Kucharski. But academic incentives mean scientists need to publish in journals to do well, while "most policy people just want the answer", says Bhatt. Kucharski says that that "does sometimes create challenges in analysis … we want our stuff to be public, but you have to think about who's collecting data, the effort they're putting in – it's a lot of work, it's not good form for us to grab it and publish it". He also notes the "important distinction" to be made "between data that's not being shared and data that doesn't exist.

"Some people working in the wider tech field were quite naïve about how patchy epidemiological data is – complaining it wasn't being shared when actually it wasn't being collected." ■

References

1. The R. E. Dyer Lecture: Epidemiologic Models in Studies of Vector-Borne Diseases, 1960.

2. Brooks-Pollock, E., Danon, L., Jombart, T. and Pellis, L. (2021) Modelling that shaped the early COVID-19 pandemic response in the UK. Phil*osophical Transactions of the Royal Society B*, **376**(1829), 20210001.

3. Kermack, W. and McKendrick, A (1991) Contributions to the mathematical theory of epidemics – I. *Bulletin of Mathematical Biology*, **53**(1–2), 33–55. Reprint of the original 1927 paper.

4. Royal Society (2020) Reproduction number (*R*) and growth rate (*r*) of the COVID-19 epidemic in the UK: Methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. Preprint, 24 August. bit.ly/4bUIHJM

5. Kucharski, A. (2021) *The Rules of Contagion: Why Things Spread – and Why They Stop*. New York: Basic Books.

Next issue

In part three, we will take a closer look at how the data behind R was painstakingly gathered. Thanks to Understanding Patient Data (understandingpatientdata.org.uk) who first commissioned this text.