The story of the R number

How an obscure epidemiological figure took over our lives

Part 5: So what did we learn?

The Times "waved goodbye" to R on 8 March 2021. "You won't be missed", wrote science editor, Tom Whipple. He quoted John Edmunds of the London School of Hygiene & Tropical Medicine: "The era of R is coming to an end ... come the autumn, with luck, all adults will be vaccinated and Covid's ferocity will have been blunted. And then R can at last return to where it is happiest: mathematical obscurity."

Even the advent of the Omicron variant in late 2021 failed to drive renewed interest in R. It still featured in the scientific consensus statements. It was still ritually reported as breaking news every Friday, like an artefact from an ancient ceremony that nobody recalls the significance of any more. Perhaps it was less a lack of interest, more a sense of routine – both in the population trying to return to theirs, and R settling into its own of regular scientific reporting.

What can we learn from the era of R?

First, for all its epidemiological importance, R should not have been a breakout star, but part of an ensemble cast of data helping us understand the pandemic. As the Royal Society notes, for all its limitations and uncertainties, a central estimate "is certainly a much better place to be in than just making a guess through verbal argument as opposed to detailed analysis where the assumptions are clearly laid out for all to see".[1] But as George Macdonald, the father of R, wrote in 1960, "the model by itself has no significance" and its only use is as a tool. R could not bear the full weight of

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.

unrealistic expectations placed upon it. As *The Times* put it in January 2021, "No other leading nation hinged policies directly on an epidemiological statistic".

Or at least, appeared to. For all the weight government put on R in its public and parliamentary pronouncements, it is unclear exactly how it was used. We need greater transparency on how exactly data informs decisions. Transparency was an asset to scientists through the pandemic – GitHub repositories, summaries, academic papers, even coding and calculation packages helping inform and communicate the pandemic response, even if there could have been greater openness, especially at the start. Politicians, through the select committee system, were instrumental in making the case for transparency. But it was never entirely clear what role R played in government's five alert levels, five tests, five indicators, four tiers, three tiers, three steps and much else besides in our attempts to live with and lift lockdown.

There was clearly some tension between the worlds of science and politics. According to the Royal Society, "Communicating uncertainty to policy makers is always difficult since understandably they seek clear and unambiguous advice". Fliss Bennee notes that "the real crux for us is that politics is the language of certainty, policy is the language of certainty even where it offers options". Government scientists and, at times, even the media were better at conveying uncertainty than most politicians. But there are challenges with scientific, as well as political, incentives: the journal *Philosophical Transactions of the Royal Society* considered academic publishing mechanisms "incompatible" with a rapidly evolving situation, and rapid policy research "may not be substantial enough to be published as a standalone manuscript", an important academic currency.[1]

More and broader scientific expertise at the heart of government might help. Jeremy Farrar says the "approach and language" scientists bring "has to be part of the ministry, and has to be accessible … they're talking, they're educating and informing, across the ministry, breaking down that sense of them and us, scientist and non-scientist" and that "science changes when data changes … [we need to] get people more comfortable with questioning and challenging it". Fliss Bennee felt her job was "to educate people about the value of having more specialists, and to have people who can understand the specialism but speak the language of policy and communications – more science communicators at the coal face". Policy owners need to become more comfortable with the "language of caution and certainty" and "need to accept, and publicly accept, that they cannot simplify something if they don't understand it".

A final key lesson is the need to think more carefully and practically about where the data behind R and other vital epidemiological numbers comes from – and be willing to invest in it. As one select committee report put it, "For a country with a world-class expertise in data analysis, to face the biggest health crisis in a hundred years with virtually no data to analyse was an almost unimaginable setback" (tinyurl.com/4kvytcpk). Meaghan Kall's "impression was that historically we have underinvested in IT and data flows in the NHS and just generally for public health in England." See boxout "'Frankenstein data sets'": gathering the data behind R" for more detail on the challenges of collecting Covid data.

She argues that a national health system, in contrast to fragmented ones elsewhere (like the USA), should be able "to look at things at the national level, reduce inequalities, standardise the way data is collected. It's an amazing opportunity." But it's not the reality. Rosalind Eggo says that "in the UK we have a very strong idea of collective health, the NHS is a good thing … [but] I don't think the connection has been made between providing your data and allowing it to be used as part of that collective improvement, the collective benefit of the NHS". The public are conspicuously missing from any discussions about how their data could be used. Even if the recovery from our starting position was "phenomenal", says Jeremy Farrar, "we just need to get there quicker. We can't afford to go through 7–8 months before data becomes good if you face something worse than Covid – we'd be wiped out by then."

R may be an abstract number but it had a real impact on people's lives. It dictated our movements – even if it was not always clear exactly how – and dominated our conversations. The birth of the UK's Covid-19 R number is an anthology of different stories: of the strengths and weaknesses of the scientific method and the interactions of science and politics, of the industry and ingenuity of public servants, of the thirst of the public for information, of the trade-offs and nuances in distilling a hugely complex set of processes into one "simple" number. There are good stories – of world-leading mathematical modelling and of rapid improvement – but also bad: the fact that such improvement was necessary in the first place, and the consequences of that and some of the political decisions informed by R. R might be abstract; one number that is not is the more than 200,000 deaths from Covid-19 in the UK and the impact on those they leave behind.

George Macdonald looked at epidemic diseases through the ages – plague, scarlet fever, cholera – and concluded that while these examples "are from past history", "the future may be expected to mirror them and our only defence lies in an increasing knowledge of the factors which determine the geographical distribution of disease". Our response to Covid-19 provides us with plenty of knowledge to tackle future pandemics as long as we learn the right lessons. Macdonald also exhorted us to "stand back a bit" and view diseases as entities in themselves with their own life history. Doing so with the R number illustrates more than anything that R, as an entity, is what we choose to make of it: the challenges in creating and communicating it, in deriving it and deciding things with it, are the result of human decisions. ■

Glossary

- **CPNS – Covid-19 Patient Notification System, created by NHS England**
- **NHS – National Health Service, the UK's publicly funded health-care system**
- **PHE – Public Health England, executive government agency created in April 2013 to protect and improve the nation's health**
- **SGSS – Second Generation Surveillance System, an application to monitor notifiable diseases**

"Frankenstein data sets": gathering the data behind R

> They [deaths] were being fed to us from NHS England. Hospitals would just email – one email per person – "Mr John Smith has died at the Royal Free Hospital with coronavirus, here's his date of birth and his NHS number". We were pulling together "line lists" – a data set with one row per person – from these emails at the very start.
>
> Meaghan Kall, lead epidemiologist, Public Health England Covid-19 epidemiology cell

At the beginning of the gravest pandemic to hit the UK in a century, some of the most important data for understanding the virus was being pulled together in the most piecemeal way imaginable.

Meaghan Kall's experience was not unique. In Wales, Fliss Bennee "didn't have a direct feed … Every day I and colleagues phoned every single ICU [intensive care unit], 'who have you got today?', at midnight – 'how many cases do you have, how many real, how many suspected?' We just didn't know." Dominic Cummings, chief adviser to the Prime Minister at the start of the pandemic, told Parliament that "in all sorts of ways it [the data] didn't exist" – the "data system" in March 2020 "was me wheeling in [a] whiteboard … and Simon Stevens [NHS England chief executive] reading out, from scraps of paper, numbers from the ICUs."

The Cabinet Secretary, Simon Case, told an audience that the centre of government "started off with officials emailing Excel spreadsheets back and forth late at night, to be turned into PowerPoint slides for ministers the following morning" (tinyurl.com/bde83vn9). Marc Warner, chief executive of Faculty, an AI company working with the NHS, told the *Sunday Times* that the NHS system was "completely dysfunctional in a fast-moving crisis. Thousands of spreadsheets a day were bombarding NHS headquarters and then being manually integrated in Microsoft Excel, through copying and pasting" (tinyurl.com/5n8chhdh). Other sources paint a similar picture, of ad hoc data sharing via formats not designed for data – emails, phone calls, Microsoft Word documents – that required time and effort to make usable. The Royal Society criticised the "uneven data quality and slow access to information" that were "a major impediment to good epidemiological analysis of the state of the epidemic and predictions of future trends" (tinyurl.com/mrxzey5v). Sir Patrick Vallance told Parliament that early on, "it was difficult for SAGE to accurately assess the state and trajectory of the outbreak at that time due to the lack of data" (tinyurl.com/3tp58kr6).

Preparations for a pandemic

Pandemic preparedness plans touched on data, but not in much practical detail; where they made recommendations, they appear not to have been taken up. Virtually none of them mention R. The Public Health England (PHE) 2014 Pandemic Influenza Response Plan (tinyurl.com/575px86w) details data collection protocols for the FF100 – the "first few hundred" cases of a new disease – possible only "if a systematic approach" had been developed "in advance", including contact tracing. Primary objectives included estimating the secondary attack rate (other people in a household who fall ill) and serial interval (time taken between the primary and secondary cases showing symptoms); the estimation of the "basic reproductive number" was a secondary objective. FF100 protocols were enacted as Covid-19 hit, but "the data we anticipated was nothing like the data we got", says Cambridge's Paul Birrell. The nature of Covid-19 meant it was not plausible for people to go to the doctor, different sources of hospitalisation data were "difficult to interpret in a hurry", there were differing interpretations of what counted as "an admission" (people diagnosed once admitted? People admitted with Covid but not because of it? People who got Covid in hospital?) and it was "difficult to pick up a signal because of reporting delays". It can be "quite difficult to anticipate what data you're going to end up having".

A Scientific Pandemic Influenza Group on Modelling summary for pandemic flu in 2013 (updated in 2018) said planning should include facilitating "the early collection and sharing of data between nations", but did not specify data types or formats. The 2016 flu simulation, Exercise Cygnus – which argued the public was more likely to stomach difficult decisions if they were "made in an open, transparent and inclusive way" – recommended establishing a cross-government working group, to clarify the "process and timelines for providing and best presenting data on which responders will make strategic decisions". Participants "were unclear about how epidemiological information would be produced and disseminated". Cygnus was set up in week 7 of an epidemic and focused on the "treatment" and "escalation" phases (tinyurl.com/mrcjbb35); it skipped the preceding "detection" and "assessment" parts (and "recovery" at the end). Jeremy Farrar says such exercises risk asking "did we get the outcome we wanted?" at the expense of "greater curiosity about what it's telling you about the truth that lies underneath the bonnet". Many of these exercises were predicated on planning assumptions for pandemic scenarios quite different from Covid.

It was not just in the UK that all the plans and preparation exercises failed to deliver the data necessary to face a pandemic. Journalists at *The Atlantic* realised the US federal government was relying on their data (tinyurl.com/4rt5yes9): pandemic plans "stressed the importance of data-driven decision making" but "largely assumed that detailed and reliable data would simply … *exist*. They were less concerned with how those data would actually be made." Countries that did better – South Korea, Taiwan – had thought about data streams, legislation about data linkage and even suspending some privacy laws "miles before Covid hit", according to Adam Kucharski: "they had a good set up that was good to go". The UK, meanwhile, was having a debate on digital contact tracing in the middle of a pandemic.

Tracking a new disease

By the end of January – before those emails of individual deaths started hitting their inboxes – PHE was standing up its systems to monitor "the novel Wuhan virus". Meaghan Kall says it was an "all hands on deck situation" building the "epi cell" for coronavirus, a unit that produces data to help build the epidemiological picture of a disease (testing numbers, case numbers, hospitalisations, deaths). They knew they needed to "create

a more automated system with the ability to move data electronically and with minimal manual input".

A new disease meant some new data collection – such as a seroprevalence survey, detecting the disease in donated blood – but PHE also turned to existing work, routines and resources. Incident response protocols had been enacted for outbreaks of swine flu, Ebola, bird flu and other diseases. An existing lab system, the Second Generation Surveillance System (SGSS), had been designed to monitor notifiable diseases – those PHE was legally mandated to monitor, such as tuberculosis, cholera, salmonella and MRSA. Covid-19 became a notifiable disease under the Public Health (Control of Disease) Act 1984 at 6.15pm on 5 March 2020 (tinyurl.com/4bjvs4wn; tinyurl.com/4wpcfdpm). The SGSS meant that labs around England could test samples and diagnose cases with a ready-made reporting pipeline back to PHE, an improvement on the "emails and pieces of paper, scanning in PDFs, test results being emailed" of the early days of the pandemic, "which was completely unsustainable". PHE also used its existing "data lake", a repository of PHE and external data sets, like Hospital Episode Statistics (NHS England admissions, appointments and attendances) or death certificate data from the Office for National Statistics. "We could never have responded if we didn't have this previous specific investment."

Challenges remained, not least in recording deaths, a key data set for R modellers. Initially, PHE received details only of deaths in hospitals, via those individual emails (which they compiled into "line lists", a spreadsheet with one line for each patient), and via the Covid Patient Notification System (CPNS) built by NHS England. Modellers needed a more complete picture. ONS data on death registrations – the Births and Deaths Registration Act 1836 requires all deaths to be registered with the Local Registration Service and General Register Office in England and Wales – is the gold standard, but given delays in the registration process, took 10–14 days to come through. This was too long a lag. PHE looked at the data they had and merged their line lists with the CPNS, the SGSS data, and with the NHS Demographic Batch Service (which contains patient records). If there was enough personally identifiable information (PII) in the data, PHE adapted a system they had previously used for a very different purpose: to ensure anyone taking part in their research who had died would not receive any further correspondence. They ran their SGSS surveillance data through the NHS systems to find people flagged as having died overnight; combining that with the actively reported hospital deaths gave them a "backbone" of data on deaths. This led to the early realisation that the NHS had been undercounting deaths.

Linking data sets, such as surveillance lab data with hospital data, around PII requires "getting permissions and approvals … we may not need to know who that patient is, but we need that information to link it across systems". Kall says this "very important" process could take years, but was "expedited" during Covid. This work really matters – "issues around privacy cannot be forgone in our urgency to link data sets and create numbers" – but involves a lot of form filling: "due diligence and admin, justify what we're linking the data for, how it's being used, a data protection impact assessment against all of our linkages, risk assessments".

For all the work that goes into this data, it is still far from perfect. Kall notes how data structures are often "cobbled together" over time, evolving into "Frankenstein data sets … they start as one thing, evolve to do something else, get something added into them". Data can end up being used in a completely different way to that intended: something like the Health Episode Statistics data set was designed for administrative purposes, but was repurposed during the pandemic for health surveillance.

The first stop for all this data was PHE's own in-house modellers – "our first customers", who prompted the creation of the deaths data set. Sent the daily outputs in basic text or CSV (a spreadsheet file format), they would sanitise the data (e.g., removing PII) so it could be shared with others, and upload it daily to a server accessible to all of the other modelling groups. ■

Acknowledgement

Thanks to Understanding Patient Data (understandingpatientdata.org.uk) who first commissioned this text.

Reference

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

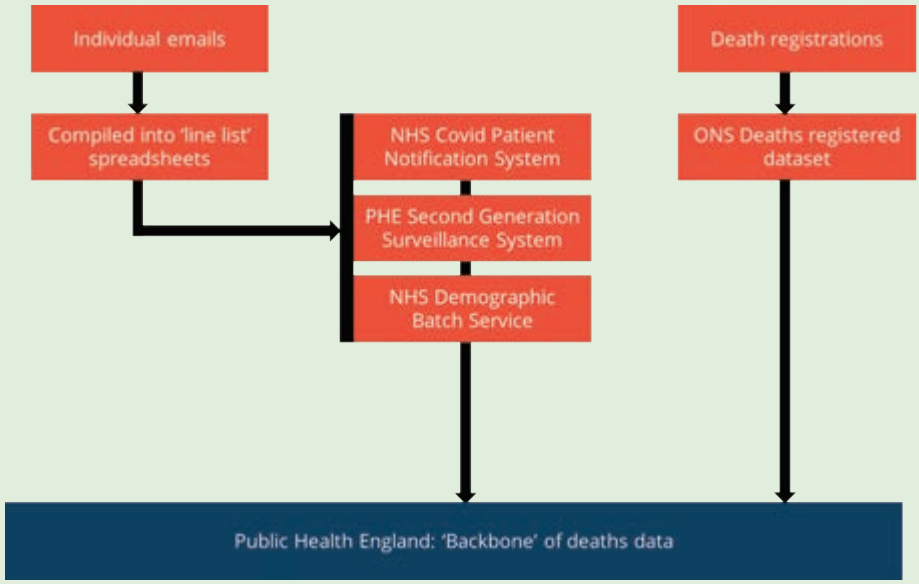

Figure 1: PHE's system for collecting data on Covid deaths, as of 5 March 2020