2023 INTERNATIONAL CONGRESS OF ACTUARIES
28 MAY – 1 JUNE 2023 · SYDNEY

Actuaries Institute Australia

International Actuarial Association
Association Actuarielle Internationale

The Good, the Bad and the Ugly of Data Democratisation - Lessons from a Data-Rich Pandemic

Prepared by Chris Dolman & Hugh Miller

Presented to the
2023 International Congress of Actuaries
28 May – 1 June 2023

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Abstract

The ongoing pandemic has featured unprecedented openness with data by countries and regions all over the world. Detailed and rich data is available to anyone who wishes to access it, allowing them to attempt to understand the pandemic and to analyse, comment and debate the policy response with data as a foundation.

‘Democratisation’ of data within organisations is also a growing trend. Similarly to what we have seen during the pandemic, this refers to the increasing openness and availability of data (in this case within an organisation), and the ability of employees to use that data in their daily working lives. Government departments have similarly been releasing more data on a regular basis - within and across departments, and sometimes publicly.

The wide availability of data during Covid-19 is often praised (including by the authors), yet it has not always led to positive outcomes. Whilst many good things have been enabled by the openness of data, there have been notable missteps, some highly public.

In this paper, we examine some of the effects that readily available data had on the pandemic response and the public’s general reaction and understanding to it. We include well known examples and our own personal experiences. Our aim is to illustrate examples of the good, the bad and the ugly of ‘democratisation’ of data, using the pandemic as a motivating example, to give guidance to organisations on maximising the ‘good’, and minimising the ‘bad’ and ‘ugly’ in their democratisation efforts.

1 The democratisation of data

1.1 Introduction

Democratisation of data has been a growing item of discussion in recent years. Generally, it refers to mechanisms whereby data is made more readily available within organisations - either private companies or within government[1]. In some cases, however, the use of the term bleeds into the broader topic of 'open data' - that is the external publication of data for public consumption[2].

In this paper, we aim to learn about the former by examining the latter. By examining how data was published publicly and then used by the general population during the pandemic, we can observe both good and bad features. This should provide lessons to us as we democratise data internally within our organisations or government departments - how can we create processes to enable the good, while protecting against the bad?

Our analysis of the use of data in the pandemic is limited by our paper's scope - it should not be considered to be complete, nor balanced. Much has been written on the topic of pandemic data usage already, and we do not attempt to add to this. Rather, we aim to cherry pick this already rich analysis so as to point out particular instances of both good, bad and ugly data usage which we feel demonstrate general principles which a corporate or government department's data democratisation programme should be aware of.

1.2 Trends in business

The typical enterprise data warehouse of the mid-late 20th Century was held under tight control, often under the IT department. Specialist staff would extract some of the data from core IT systems, transform and then store it in a form allowing downstream use within particular forms of analysis and reporting. Usually the data would be specified by a particular business request, and be constrained by budgets and resources - hence it would tend to include only a subset of the organisation's data. Access to this set of transformed data was granted only to suitably trained experts, who were entrusted with using the data appropriately. This created a scarcity in many organisations - novel questions of senior management which could (in theory) be answered by interrogation of the organisation's raw data might go unanswered. In some cases this was due to lack of specialist analytical resources, in others because the relevant data was simply not to be found in the data warehouse.

---

[1] The literature in this area is vast. For the interested reader, Lefebvre, Legner & Fadler (2021) provide a useful literature review and taxonomy of common themes, and highlight several case studies from a range of industries. For government the literature appears to have more history, for example Layne & Lee (2001)'s proposed model of e-government relies heavily on what we might now call democratised data.

[2] For a recent example, see Lane (2021)

The ‘data democratisation’ trend in contemporary organisations has its root in the ‘big data’ trend of the early-mid 2010s. The general idea was to address both of the bottlenecks described above - load all the data into the new data warehouse (rebranded as a ‘data lake’, to be architected less rigidly than a traditional warehouse, and allowing greater inclusion of unstructured data types), and provide access to more, sometimes all, staff - at least for non-sensitive data types. Not only would this allow senior management to more easily tackle those novel questions described above, but staff could discover new things by analysing more data - or so the theory went.

While in many organisations these ‘data lake’ projects floundered initially[3] - to be then popularised as ‘data swamps’ - many organisations have now got past those initial challenges. The general goals have usually remained and may now be achievable. By allowing more people to have access to more of the organisation’s data, as well as tools with which to analyse it, it is still usually believed that greater and faster insights can follow.

We note that in the remainder of this paper we will use the terms ‘data warehouse’ and ‘data lake’ interchangeably and loosely, without strict definition of either, essentially to refer to a centralised repository of data within an organisation (of whatever form or architecture).

1.3 Trends in government

In government, too, there has been a push for more open data provided in a timely fashion. This covers both more data being made available publicly, as well as broader sharing within government. The growth of dashboarding tools, such as Tableau and Microsoft Power BI, have also led to a push to have more data available in these formats, making information accessible to less data-savvy users.

On the public side, many governments are creating open data platforms where departments can regularly upload statistics that may be of interest to users. In Australia, data.gov.au and similar state-level websites perform this role. Some of this comes from a desire to improve transparency and accountability across government. Another factor is a belief that the public is the ultimate owner of government data, and so is entitled to access it if it’s provided in a responsible and safe way.

Governments also routinely set targets in priority areas - such as Closing the Gap[4] for outcomes for Aboriginal and Torres Strait Islander peoples, or a set of headline indicators to measure and drive policy[5]. Such approaches rely on high-quality indicators being collected, so that tracked improvements are both robust and meaningful.

---

[3] For example, see https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven or https://visual.ly/community/Infographics/technology/cios-big-data

[4] https://www.closingthegap.gov.au/

[5] For example in NSW: https://www.nsw.gov.au/premiers-priorities

For internal purposes, one trend is the increasing awareness of the importance of cross-agency links. Often people in need accessing a particular service will also be seeking other types of support - for example a person accessing income support might also be on a social housing waitlist or be receiving family support services. Better visibility of this across government enables better person-centred design. Data sharing of aggregate, unit record, and even linked data, improves the ability to make evidence-based policy decisions.

2 Open data and the pandemic

From the very beginning of the pandemic, governments across the globe recognised that data could be of assistance in the response. Importantly, it was generally also recognised that this was not just something of use to particular government departments, to be used in secret - publication of the data could serve a variety of additional purposes. Firstly it helped to satisfy the public's demand for information - understandable during a time of crisis. Additionally, publication of data could enable improvements in the pandemic response outside of the obvious areas of government departmental policy, allowing academics, journalists and broader civil society to use and analyse the data as they saw fit, to help the world deal with the crisis.

This openness of data created lots of good, but poor outcomes also ensued. What follows is a selection of examples of both good and bad aspects of this openness of data, which we feel are of direct relevance to a data democratisation programme.

3 The good

3.1 Quality of the best

A clear advantage of open publication of data is that it enables more research and analysis to occur, given a particular problem to solve. Under this model, the amount and speed of analysis is not limited by an exclusive tender process or a restrictive data access process, but by the time and enthusiasm of researchers. While this will no doubt lead to some poor analysis which can create problems (a topic to which we will return), it can also lead to impressive results.

One area that saw rapid development in the early stages of the pandemic was academic research on transmission and other COVID-19-related models. Traditionally this would have been restricted to a small number of research institutes commissioned by government to undertake modelling, with access restricted to the remainder. While this commissioned work still occurred[6], there was a significant amount of other activity that both contributed to the overall

[6] In Australia examples include The Burnet Institute https://www.burnet.edu.au/news/1506_initial_modelling_projections_for_second_epidemic_wave_in_sydney_nsw , and the Doherty Institute https://www.doherty.edu.au/our-work/institute-themes/viral-infectious-diseases/covid-19/covid-19-modelling/modelling . However, recent news stories suggest that not all

understanding of the pandemic and may have influenced commissioned modelling and overall policy. Examples included:

- People making use of natural experiments to understand the impacts of different policy responses such as mask mandates and lockdowns[7]
- Advanced models to understand transmission and latent case numbers[8]
- Independent reporting and analysis of mortality trends soon after data release, for example the regular bulletins of the Actuaries Institute's COVID-19 Working Group on excess deaths[9]

Most of this research was facilitated by the availability of data across the world. In turn, much of this research was published in an open way so that others could similarly benefit.

While the generation of a lot of insight enables the creation of good analysis, there will also be examples of poor analysis which must then be managed - as we discuss below.

3.2 Speed of innovation

Democratisation of data can also be an implicit feature of modern online services, which by their very nature make some information public. This can lead to some spectacular outcomes.

One example of this in the pandemic was the Covid Queue website - conceived and built over a weekend by a single software engineer from Sydney, initially to help his close friends and associates. This simple webpage aggregated the many disparate booking systems of government departments and health providers, so that people could find the next available vaccine appointment easily (this being an extremely common pain point, as the authors can personally attest). Perhaps owing to the sheer level of public frustration at the vaccine booking process at that point in time, this (extremely hacky) website quickly went viral and led to national news coverage within days.[10] While some of the news coverage was critical of government, we suggest instead that this demonstrates a useful feature of openness - people have a natural incentive to solve problems affecting them, and can do so rapidly if information is available allowing them to do so. It is likely this will be quicker than relying on a centralised agency to find problems and solve them, with their naturally limited resources and potentially conflicting priorities. It may also be of higher quality, particularly in cases where incentives are naturally aligned, as was the case for Covid Queue.

---

commissioned research was published: https://www.abc.net.au/news/2023-01-07/covid-data-government-secrecy/101827548

[7] For example, Islam et al. (2020) https://www.bmj.com/content/370/bmj.m2743

[8] For example, Gostic et al. (2020) https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008409 for Bayesian modelling of case numbers, or Kerr et al. (2021) for agent based models of COVID-19 dynamics

[9] For example, https://www.actuaries.digital/2022/12/07/covid-19-mortality-working-group-excess-mortality-continues-in-august-2022/

[10] For example: https://www.abc.net.au/triplej/programs/hack/new-site-dings-users-when-pfizer-jabs-are-available/13481814, https://7news.com.au/sunrise/on-the-show/genius-website-helping-australians-book-covid-vaccine-appointments-c-3608528

In the USA, the COVID-19 Tracking Project was an early volunteer effort to collate and understand COVID testing and case data. It similarly improved upon what was available through public channels.[11]

3.3 Fact-checking

With regularly published data from many countries in the world, it became extremely easy to fact check official spokespeople. Journalists did not have to accept commonly used, vague platitudes ("things are improving", "we've reached the peak", "no need to panic" etc), instead they could look at the data themselves and make their own minds up, including challenging the official narrative. Independent fact checkers such-as FactCheck.org could work far more effectively due to the openness of data[12].

In addition to fact-checking official sources, a significant amount of fact-checking was undertaken on media and online sources, including social media. Examples of claims that were quickly fact-checked included:

- The origin of COVID-19
- The role and effectiveness of interventions such as lockdowns, mask-wearing and social distancing
- The effectiveness and safety of different treatments after a COVID-19 diagnosis
- Vaccine safety and effectiveness

Evidence suggests that fact-checking reduces misconceptions, even if the effect may be temporary (Carey et al., 2022).

In all times, but especially in times of crisis, it is important for our discourse to be based on facts, in order that policy discussion can reflect reality.

3.4 Establishing a common basis and language

In many cases, having government-provided data imposed default definitions and a common basis for understanding and analysing how the pandemic was tracking. Given there were a plethora of possible definitions for COVID-19 cases, hospitalisations and deaths[13], having a formal definition in each jurisdiction allowed experts to to establish a common frame of reference for discussing trends and responses.

---

[11] https://www.theatlantic.com/science/archive/2021/03/americas-coronavirus-catastrophe-began-with-data/618287/

[12] See https://www.factcheck.org/a-guide-to-our-coronavirus-coverage/

[13] For examples of discussions around defining COVID deaths, see https://ukhsa.blog.gov.uk/2020/08/12/behind-the-headlines-counting-covid-19-deaths/ and https://www.aamc.org/news-insights/how-are-covid-19-deaths-counted-it-s-complicated

Transparency around definitions in turn enabled better international comparisons. These were often collated by news outlets and academic institutions.[14] Different policies across countries and their timing relative to COVID waves could then be more sensibly understood.

There were still many issues with such comparisons and interpretations - no countries had identical definitions and data collection processes. But an open-data approach greatly improved the situation, and reduced the chances that researchers would reach incorrect conclusions due to unknown biases in data practices.

3.5 Transparency and improved practice

Regular publishing of statistics meant that governments were more likely to be honest in their assessment of the pandemic. Transparency kept governments honest, but potentially also reduced pressure on governments from needing to find positive spins in the delivery of data.

Demand for good data also led to improved practices in many places. Perhaps the best-in-class was the introduction of the Coronavirus (COVID-19) Infection Survey in the UK, run by the Office of National Statistics. Since mid-2020 the Office took a random sample of the whole population and tested for COVID-19, rather than relying on people electing to get tested. This provided an unbiased estimate of cases that could be tracked over time[15]. This filled a significant gap in understanding about the extent of transmission (including asymptomatic transmission and reinfections), the accumulation of antibodies (both from vaccination and infection) throughout the population and insights into long COVID.

4 "The Bad and the Ugly"

4.1 Poor data literacy

While we both regularly observe and comment on poor data literacy,[16] we were still impressed by the sheer creativity of some of the data-related mistakes we observed during the pandemic. At times such mistakes were innocuous and easily corrected, but on occasion they led to confusion and - in the worst cases - misinformation.

---

[14] For example, the Johns Hopkins coronavirus research centre, https://coronavirus.jhu.edu/map.html, and the Our World in Data Coronavirus pandemic hub, https://ourworldindata.org/coronavirus, were early, high-quality, efforts to enable international comparisons.

[15] https://www.covidsurvey.ons.gov.uk/ . A recent perspective on the role of the survey is available at https://theconversation.com/the-ons-has-published-its-final-covid-infection-survey-heres-why-its-been-such-a-valuable-resource-202589

[16] We have both previously published Actuaries Digital articles on the topic, interestingly in both cases on sports commentary. See: https://www.actuaries.digital/2022/02/02/the-reaction-to-nadals-impossible-victory-cognitive-biases-and-data-literacy/ and https://www.actuaries.digital/2021/08/03/olympic-gymnastics-underperformers-or-statistical-quirk/

A very famous example arose from what appears to be a lack of awareness of Simpson's paradox (Simpson 1951) - something which frequently befuddles the uninitiated[17]. In late 2021, the UK Office of National Statistics published raw data showing higher COVID deaths in the vaccinated population (of people aged 10-59). This led many commentators to conclude that vaccinations were increasing the risk of COVID death, and unfortunately helped to drive a number of conspiracy theories. However, the risk of COVID death increases dramatically towards the upper end of that age range, and the vaccinated population at that time were (and still generally are) concentrated in that population. Simply: older people die more often from COVID (all else equal), and are also more likely to be vaccinated. Had the data been published using finer age bandings (which ultimately did occur), the reverse conclusion would've been clear - vaccinations materially reduced the risk of COVID death, across the age spectrum.

There was a concerted effort to refute the false claims made by many commentators, including by the IFoA's COVID-19 Actuaries Response Group[18]. Unfortunately, once the error became viral and made its way into popular consciousness, including the conspiracy theory space, it became more difficult to refute (and has been repeated regularly ever since, across a range of datasets and countries). No doubt this led to at least some reduction in vaccine takeup.

4.2 Misunderstood definitions

In January 2022 social media posts[19] were circulating saying that only 17,000 deaths in the UK were due to COVID, or an even lower figure of 6,000, based on new death certificate data that had been analysed. This was an order of magnitude lower than official estimates (at the time, about 140,000), which had the risk of confusing and upsetting many people.

What was particularly interesting about this was that the numbers were technically 'right' in the sense that they were some of the numbers released by the Office of National Statistics[20]:

- The 17,000 figure related to the number of deaths due to COVID-19 where the person had no other (usually pre-existing) health conditions. For instance, a person with high blood pressure who died of COVID-19 would have been excluded. The difference between 17,000 and the higher death count reflects the very high proportion of the population with pre-existing health conditions.
- The 6,000 figure was the number of people who died of COVID with no other cause of death listed on the birth certificate. This definition is clearly too narrow, since for example if a person died of pneumonia caused by COVID-19, it would not be counted. Most deaths caused by COVID-19 listed other conditions in this way.

---

[17] See also the discussion at https://www.ntu.ac.uk/about-us/news/news-articles/2021/11/in-the-wrong-hands,-vaccination-statistics-can-prove-deadly-simpsons-paradox-shows-why

[18] https://covidactuaries.org/2021/11/22/simpsons-paradox-and-vaccines/

[19] https://www.bbc.com/news/60145237

[20] https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/deathsfromcovid19withnootherunderlyingcauses?s=08

The broader lesson is clear - unless we understand the way numbers have been defined, and how that compares to the thing we wish to measure, there is the potential to use data poorly. This is particularly problematic when misaligned data appears to confirm pre-existing (but incorrect) views about the world.

This issue extends to other contexts. Often in government there is a disconnect between what we can measure and what we can see. For example, we can measure the number of people presenting to homelessness services, but this is only a weak proxy for the number of people who are actually homeless - many people sleeping rough, couch-surfing or living in unsuitable dwellings will not present to services.

Definitional issues naturally extended to comparisons between countries, potentially tripping up unwary analysis. For example, the UK uses a definition of death within 28 days of a COVID-19 diagnosis, but Belgium has a much broader definition (e.g. it includes suspected COVID-19 related deaths in nursing homes). Other countries, by contrast, have much narrower definitions - for example China restricts counts to deaths from pneumonia or respiratory failure, potentially excluding deaths arising from exacerbated pre-existing conditions. Definitional transparency is therefore important in understanding the nuances of such comparisons.

4.3 Variable data definitions and quality

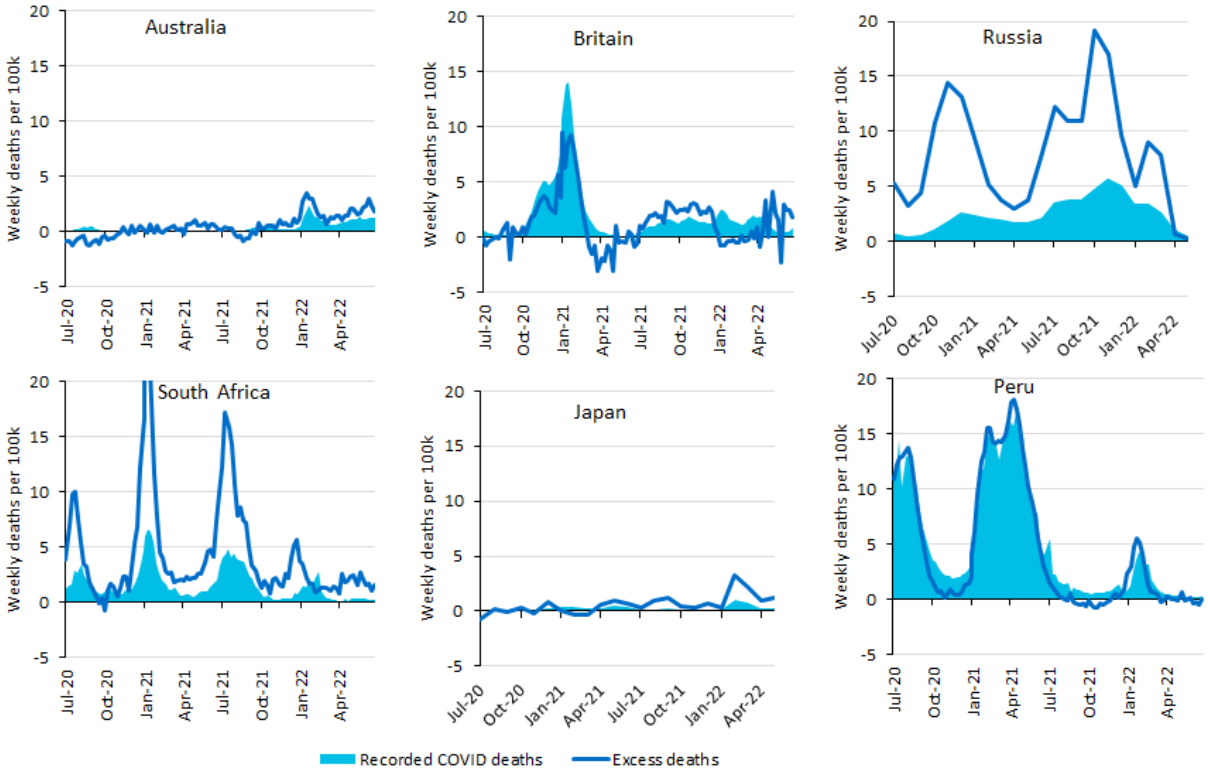
Differences in the method and quality of data collection have obvious implications for making comparisons or judgements on data.

One obvious area where data quality varied was in the degree of testing in a particular country or region. Reported case numbers based on testing will dramatically underestimate the true number of cases, if testing is difficult to access. For example, one of us was symptomatic in around May 2020, but could not access testing at that time as no known ‘close contact’ existed - mildly symptomatic potential cases such-as this would not have been captured in official data. Since the testing regimes varied significantly across the globe, country comparisons were fairly meaningless, though league tables were still produced[21] and the raw data was compared to see which countries were ‘faring better’.

Excess death models have been a useful validation of the issue, identifying where large gaps exist which suggest an underreporting of COVID-19 deaths. For example, Russian reported deaths initially suggested a less severe pandemic experience compared to other nations such as Britain. But excess deaths showed a large gap and a relatively severe pandemic. The Figure below also shows that South Africa saw significant underreporting, whereas other countries like Peru were able to correctly quantify the magnitude of the mortality burden.

[21] For example, see https://ourworldindata.org/covid-cases

Figure 1: Comparison of excess deaths and recorded COVID-19 deaths per 100,000 people over 2020/21 and 2021/22, selected countries



Source: The Economist Excess death tracker, https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker

The figure also shows negative excess deaths for some countries in some periods (for example, Australia during most of 2020) - a phenomenon observed more broadly too. Some causes of death were definitely reduced - car accidents and flu related deaths were lower in periods of lockdowns. It is also worth noting issues of noisy mortality data - for example, if 2019 deaths were unusually high in a period this could appear as negative excess deaths in subsequent years.

Even within countries, variation in data collection and quality made simple interpretations hard. Many UK news organisations reported 'record numbers' of new cases in September 2020, when more than 6,000 cases were recorded[22]. This is despite the fact that daily case numbers were likely over 100,000 cases per day earlier in the pandemic, when the degree of testing was relatively low.

Jurisdictional differences within countries could prove problematic. While there was a lot of regular state-level reporting in the USA, many issues were raised by academics and news

[22] See, for example, https://news.sky.com/story/coronavirus-uk-reports-highest-ever-number-of-new-daily-cases-12080489

organisations around poor quality and inconsistent data.[23] Where responsibility is devolved to states, strict standards around reporting is necessary to minimise the pitfalls of inconsistent collection.

4.4 Information overload and distraction

The pandemic represented a time when the best available information changed rapidly. This creates a risk: unless people are routinely checking the information, they will have different understandings of the state of play.

One area that demonstrated this shifting sand was vaccine efficacy. Many people were keen to understand the how much their risk changed after receiving a vaccine, however in many ways it was an impossible question:

- Efficacy could be measured in different ways (chances of catching the virus, having significant symptoms, hospitalisation, death)
- Efficacy changed as new variants emerged
- The level of COVID-19 in the community, and the level of vaccination, altered the dynamics of the likelihood of being exposed to or contracting the virus
- Vaccine effectiveness waned over time.
- Longer term impacts of COVID-19 infection are still poorly understood, hence any effect of the vaccine on longer term outcomes is also not well known

Differing evidence meant that a common language around efficacy was challenging, leading to information overload and confusion.

A related issue is one of distraction in the face of noisy data. A lot of time and effort was put into publishing daily numbers including long daily press briefings in many countries. The daily numbers became a common topic of (virtual) water cooler conversation. However, many conversations we observed failed to take into account basic features like weekly seasonality (e.g. testing numbers, hence case numbers, typically dropped over a weekend). People also failed to recognise the noise in the daily data - we observed creative ‘explanations’ put forward for a daily increase or reduction, and we observed politicians questioned in such a manner (“why have the numbers gone up, today?”). By spending time and energy on such discussion, more fruitful topics are crowded out.

Additionally, most of the implemented policy responses had an impact over longer time scales. Even at the shorter end, with single cycle transmission and incubation periods spanning anything from a few days to a couple of weeks, policies to reduce the spread (lockdowns, mask mandates, etc) could not be expected to materially change the emerging data in a day or so. Other policies were much longer term - vaccine rollouts took months, and efficacy could realistically only be measured over a longer timescale of virus exposure. Hence daily data did little to help us understand the impact of policies. When Australian states finally elected to move

[23] For example, a summary of issues at November 2020: https://www.nytimes.com/2020/11/23/opinion/coronavirus-testing.html

to weekly reporting it was likely partially in recognition that daily reporting added to the overall noise without necessarily adding a huge amount of value.

4.5 Unfortunate Motivations

When data looks bad, it can sometimes motivate people to resolve the underlying issue at hand. However, it can also motivate people to seek to manipulate the data, to save face (broadly in line with Campbell's Law (Campbell, 1979) and the related Goodheart's Law (Goodheart, 1975)). For example, this might involve activity which changes the measurement itself but not the desired thing to be measured.

There were some stunning examples of this on display during the pandemic, including by world leaders. Then US President Trump's now infamous: "If we stop testing right now, we'd have very few cases, if any"[24] highlights the natural motivation caused by high official case numbers - slow down testing to make the numbers look lower (of course, the actual case numbers are unchanged by the measurement basis imposed, we merely change our degree of ignorance of reality by changing our measurement basis). Whether such an order was ever made is perhaps debatable[25], but the motivations are clear, and unfortunate.

In the USA, another example related to the treatment of cases and deaths related to prisoners.[26] In some cases prisoners were granted release while hospitalised, meaning that deaths would not count as a coronavirus death in prison. In other cases deaths where COVID-19 was a contributing factor were not counted as coronavirus deaths. In both cases the impact is to reduce the number of reported deaths in prison. Similar issues related to the reporting of case numbers in prisons in various states.

The World Health Organisation has recently criticised China's narrow definition of COVID deaths, which only includes pneumonia or respiratory failure in hospital, which will understate the true death count.[27] While some rationale of the definition may be practical rather than an attempt to minimise the estimate of COVID deaths, it appears consistent with a government that is claiming a 'decisive victory' over COVID.[28]

---

[24] As reported by Weixel in The Hill: https://thehill.com/policy/healthcare/502819-trump-on-coronavirus-if-we-stop-testing-right-now-wed-have-very-few-cases/ , also widely covered elsewhere

[25] See coverage by CNN: https://edition.cnn.com/2020/06/22/politics/donald-trump-testing-slow-down-response/index.html

[26] See https://www.nytimes.com/2021/07/07/us/inmates-incarcerated-covid-deaths.html. https://uclacovidbehindbars.org/ and https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2021/10/27/some-states-are-cloaking-prison-covid-data, for example

[27] https://www.bbc.com/news/world-asia-china-64167052

[28] https://www.theguardian.com/world/2023/feb/17/china-victory-covid-deaths-virus

4.6 Incorrect insights and misinformation

The availability of data and the collective hunger to understand the pandemic led to a large amount of research - turning data into insights. About 20,000 scholarly articles about COVID-19 were shared in the first four months of the pandemic, of which a third were preprints - by definition unreviewed. These results were then available for other researchers and policy-makers, and were often adopted in rapid time.

Some of this speed is welcome. The UK RECOVERY trial found that dexamethasone reduced death rates by up to a third for critically ill patients. The drug was being used within a week of the findings preprint being published on medRxiv, well before the article was reviewed and published in the *New England Journal of Medicine*. The use of dexamethasone has probably saved over a million lives.

But not all stories of rapid insight generation were so positive:

- An April 2020 preprint uploaded to SSRN claimed observational results that suggested anti-parasitic drug ivermectin reduced mortality. The results were quickly disputed and the paper withdrawn by May that year[29], but not before several countries adopted ivermectin as an official treatment.[30] It also led to people taking the drug without medical advice, sometimes in dangerous quantities.
- Relatedly, other studies suggesting ivermectin as an effective treatment were shared but never published in peer-reviewed journals. Some of these were used in meta-analysis, even after their withdrawal, making it hard to 'unscramble the egg'.
- Unfortunately, even peer review is not a perfect protection. A French study supporting the use of hydroxychloroquine was accepted for publication less than a day after submission in March 2020.[31] The study was one reason for a surge in demand for a drug that ultimately proved ineffective as a treatment.
- Perhaps most shockingly, papers published by the *Lancet* and *New England Journal of Medicine* claiming hydroxychloroquine was ineffective were apparently based on a large detailed patient dataset collated by the Surgisphere Corporation in the USA. It turned out that the underlying dataset was suspect and some authors were provided with only summary data.[32]

These cases demonstrate that 'insight generators' carry a large responsibility to do high-quality work, and be upfront about potential limitations.

At the extreme end of the spectrum of bad insights is misinformation - people who will use results that are wrong, or very selective, to push an agenda, often ideological. In our view this

---

[29] https://www.mdpi.com/1999-4915/13/11/2154
[30] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050401/
[31] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102549/
[32] https://www.theguardian.com/world/2020/jun/04/covid-19-lancet-retracts-paper-that-halted-hydroxychloroquine-trials

goes beyond being merely bad - misinformation is plainly ugly. Issues of misinformation have been persistent across the pandemic:

- The origin of the virus
- The incidence and mortality of the virus (in both directions - we have seen claims of death counts being both unreasonably inflated and deflated via ‘creative’ analysis)
- The effectiveness of preventative interventions (e.g. wearing masks, specific medicines)
- The safety and effectiveness of vaccines.

While much of the misinformation could be quickly debunked by a curious reader, misinformation did find some willing audiences that undermined policy responses to the pandemic.

The ultimate challenge in a world of excess insights and misinformation is to identify what is worthwhile, and give people the maximum opportunity to learn from the ‘good’ and not be unduly troubled by the ‘bad’, nor seduced by the ‘ugly’.

5 Implications For Your Organisation

As the pandemic showed us, releasing data for broad use can create lots of good, but bad and ugly uses of that data can also emerge. The challenge then, for organisations that ‘democratise’ their data, must be to encourage the good, weed out the bad, and prevent the ugly. Otherwise, they may run the risk of data-driven chaos! In this section, we suggest some high level practices for organisations to consider.

5.1 Validated Sources of Truth

A common first step in a democratisation strategy is to first create a centralised, ‘trusted’, data lake or warehouse, with the idea being that this can then be ‘opened up’ suitably to a wide range of people. Typically this also involves removing or masking certain forms of sensitive data. While in some organisations this centralisation process can be quite an onerous task, it is one which we consider to be valuable. Of particular relevance to this paper is that this centralised warehouse becomes the ‘source of truth’, from which analysis and discussions can then be based.

In building such a centralised repository, data processing decisions will be required in order that (for example) similar datasets from different IT systems can be aggregated or compared. By centralising this taxonomic work, organisations can avoid situations where similar sounding data are unfairly compared - as we saw in the pandemic when raw country level data was compared, despite having different underlying definitions and collection bases for things like case numbers.

Additionally, by centralising the initial data processing, organisations will avoid the natural variability in decisions which would arise from different data teams performing essentially the same task. This will help people to avoid arguments over the nature of the data itself, or

arguments over the results of analysis which might otherwise arise due to variation in data processing decisions.

5.2 Monitoring Data Uses

As we saw in the pandemic, allowing general access to data can result in a lot of good. People often understand their own problems very well, and if data can help them to solve those problems, making it readily available can be a good thing.

However, making data more generally available is likely to also cause issues unless data usage is carefully monitored. Clearly, some bad uses of data might emerge. This could range from innocent mistakes such as technically incompetent analysis causing poor decisions to be made, through to deliberate misuse of data for political or personal ends. Traditional data governance topics such-as preventing invasion of privacy may also be challenged as data becomes more available and widely used. To mitigate this, we suggest that a corporate data democratisation strategy should carefully consider a proactive monitoring scheme for data usage, in order that bad uses can be shut down before they cause harm. In the pandemic, this was largely organic; organisations can be much more deliberate about how this could occur. One example could be an offshoot of an existing risk or audit function, with general independence from any particular business function, acting as an independent arbiter or fact checker. To be effective, such a function should have substantial powers - for example to remove an offending report or analysis, issue a correction, or (perhaps for repeat offenders) removal of data access prior to some suitable training being undertaken.

Part of this monitoring function is cultural - ensuring the organisation is primed to prioritise good quality analysis and insights and deprecate poorer quality insights, even in the face of potential supporters of that poor analysis. It can be hard to curtail an incorrect narrative once it has been shared widely, so early intervention is needed.

5.3 Warnings

While a democratisation strategy necessarily involves a lot of openness and trust of users, this does not mean that no guidance should be given to them. Organisations should be expecting to include substantial metadata alongside the data within the centralised data warehouse described above - this might include things such as data lineage, constraints over usage, summaries of data quality, or classification of the data (for example in terms of privacy). Alongside this typical metadata, organisations should consider issuing warnings to users about common misuses of the data. While much of this could be constructed in advance (with a suitable imagination), a process should be developed such that warnings are continuously developed as issues are observed and corrected in practice. This may be an additional role for the audit team noted above.

5.4 Training and Data Literacy

Many people (in business, and life in general) have only a rudimentary knowledge of statistics. This can lead to issues, as we saw during the pandemic. While an independent team as discussed above can play a role in correcting any such errors, it is clearly better to reduce the chance of an issue occurring. Hence, organisations aiming to democratise data should offer significant training and development opportunities to their staff. In our observation, training in this area tends to be deficient, with focus tending to be on the mechanical operation of software tools (e.g. “How to build a dashboard in Tableau”). More emphasis needs to be given to core statistical concepts, common issues in data analysis (e.g. noise, outliers, missing data, etc), training on inference, and common fallacies to avoid (such as we saw in the pandemic repeatedly with Simpson’s Paradox). Training should be provided both for those conducting analysis, and also those interpreting it - ideally, management should be sufficiently literate to judge the reasonableness of an analysis for reaching a particular conclusion. This is certainly more advanced than the typical training on offer, but in our view it should be seen as essential to an effective data democratisation strategy.

5.5 Internal promotion and awareness

Finally, if significant resources are expended on good data democratisation within an organisation, it is also important that people are made aware of the resources available. It has been common, in the authors’ experience, for a subdepartment to undertake a data exercise only to discover they could have obtained the data more quickly (and more consistently) from an existing location that they were not aware of.

Good promotion can include awareness campaigns, but more commonly is aided by good design of internal information or communication platforms such-as intranets - enabling self-discovery of useful resources in a natural way. This may require those delivering data products to engage with IT and communications to ensure resources are properly incorporated into other staff resources.

6 Conclusion

We believe that there is much that organisations can learn from the way data was used and abused during the pandemic. While we have spent time labouring the ‘bad and ugly’, we are ultimately very optimistic on the value of data democratisation. Done well, more evidenced-based decisions will be used across organisations with less confusion and with more empowered staff. Done poorly, it can add to any existing confusion or politicking. As people charged with effective use of data, actuaries have an important role in ensuring democratisation efforts can accentuate the ‘good’.

7 References

- Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour*, *6*(2), 236-243.
- Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., ... & Cobey, S. (2020). Practical considerations for measuring the effective reproductive number, R t. *PLoS computational biology*, *16*(12), e1008409
- Islam, N., Sharp, S. J., Chowell, G., Shabnam, S., Kawachi, I., Lacey, B., ... & White, M. (2020). Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. *bmj*, *370*.
- Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., ... & Klein, D. J. (2021). Covasim: an agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, *17*(7), e1009149.
- Lane, J. (2021). Democratizing our data: A manifesto. MIT Press.
- Layne, K., & Lee, J. (2001). Developing fully functional E-government: A four stage model. Government information quarterly, 18(2), 122-136.
- Lefebvre, H., Legner, C., & Fadler, M. (2021). Data democratization: toward a deeper understanding. In Proceedings of the International Conference on Information Systems (ICIS).
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological), 13(2), 238-241.
- Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and program planning, 2(1), 67-90.
- Goodhart, Charles (1975). "Problems of Monetary Management: The U.K. Experience". Papers in Monetary Economics. Papers in monetary economics 1975 ; 1 ; 1. - [Sydney]. - 1975, p. 1-20. Vol. 1. Sydney: Reserve Bank of Australia.