Open Data, Open Science and Transparency in the time of COVID 19

William P Ball:

Introduction

A novel coronavirus now known as SARS-CoV-2 was first reported in Wuhan, China in December 2019. It targets the respiratory system, with a wide range of symptom severity and results in a comparatively high level of mortality. Crucially, it has rapidly spread across the globe, affecting people living on all the majorly populated continents.

The rapid spread, high mortality, range of severity and other unknown factors, have resulted in huge uncertainty. This means there is an urgent necessity to understand the characteristics of the virus and develop strategies to reduce its impact. To make the best decisions in this developing situation, we need information.

COVID-19, the disease caused by the SARS-CoV-2 virus, has fundamentally changed the way society interacts with health data. It has become a huge and dominating focus for both public and media interest and now guides a large proportion of research effort. Our media (both traditional and social) is dominated by daily updates on new figures, visualisations and discussion.

At the same time, many academics have shifted or pivoted their work towards studying the ongoing pandemic, as funding calls from major sources are seeking to invest large sums of money into COVID-specific research projects. As interest and concern have escalated, our requirement for information to learn more and inform decision-making has also increased. Concurrently, the production and use of Open Data and wider Open Science practices has accelerated at an incredible rate.

Open Science and Open Data Before COVID-19

Open Science proponents advocate for greater transparency, collaboration and access in the scientific process. Open Data refers to data which is made freely available for use, re-use and redistribution, normally only with the requirement to attribute the source.

Before COVID-19, campaigns to promote Open Data and wider Open Science practices in academic research had been steadily gaining interest. Journals and publishers such as Nature, Science and PLOS have recently made data sharing a prerequisite for publication. Open Access publishing was becoming more commonplace and alternative routes of peer-review and collaboration, such as pre-prints, were also increasingly being adopted by authors.

However, despite favourable opinions from researchers of open data practices in general, they have been shown to be less favourable when considering applying it to their own research (Houtkoop et al., 2018). Many data holders have traditionally been unwilling to allow access without significant restrictions, in part due to legal implications of data protection and the ethical implications in risking confidentiality. Researchers have reported that many of these access restrictions have now been relaxed to allow rapid access and linkage of data in projects related to COVID-19.

During COVID-19

The current context has expedited the adoption of many of these 'Open' practices. Publishers Elsevier, Springer and Wiley, among others, have allowed near-universal Open Access for emerging COVID-19 journal articles. Open Data has been used by multiple institutions to track cases, deaths and other information in interactive, online dashboards. Pre-prints have also been adopted at an amazing rate. Since January 2020, the bioRxiv/medRxiv pre-print repositories alone list 2490 preliminary reports (as of 30/04/2020), presumably as a method of rapid dissemination and an avenue for collaboration and informal peer-review.

Numerous prospective trials and surveys are also underway to help understand the virus, it's spread and the possible impacts it will have in the medium and long-term. Rapid response teams have been set up to appraise current evidence and modellers are attempting to predict the outcomes of potential strategies to reduce the impact of the virus. Data is central for all these efforts and in this fast-developing and time-critical context, Open Data should become increasingly fundamental to our approach to quickly generating new knowledge.

Approaches to supplying Open Data

A huge amount of health data has been supplied internationally in an 'Open' format for use or reuse by researchers, media outlets and others. These datasets range from key 'headline' figures (e.g. testing numbers and deaths) to regional breakdowns. To date, they have been used by modellers to predict disease spread as well as to identify potentially vulnerable populations or areas. Interactive platforms and maps have enhanced the data available, facilitating a better understanding of the ongoing situation.

The UK Government provides daily updates for key testing figures and mortality tracking. The Scottish Government goes further and in addition to the above, they supply information on ambulance service activity, intensive care occupancy, staffing and relevant numbers from care homes.

The wider European approach has been very enthusiastic. At the time of writing, a search of the European Data Portal, a repository for Open Datasets published by EU countries and national institutions, lists 307 datasets which mention 'Covid' or 'Corona' and this total is rising daily.

Missing Information

It may be an aphorism, but it's true in this context to highlight that we should 'measure what matters.' To date, very few institutions have prioritised collecting or sharing socioeconomic information related to the virus. Despite this, inequalities in the impact of COVID-19 are already apparent. Higher than expected mortality has been observed in BAME groups and there is a social gradient (by Index of Multiple Deprivation) in illness severity in intensive care units (ICNARC, 2020).

We should also highlight the importance of not just the effects of the virus, but also the impact of social distancing measures. Both are likely to disproportionately affect the most disadvantaged and vulnerable groups in our society. To ensure that our response to this challenge is equitable, we require more detailed socioeconomic and demographic information to be recorded and shared (Douglas et al., 2020).

Transparency in decision-making

The importance of openness and transparency in our scientific approach to the COVID-19 virus spreads beyond just the researchers and modellers and equally applies to the governments who are supplying much of the Open Data but also applying the subsequent evidence in their decision making.

The UK Government have repeatedly stated their decisions are based on 'The Science' but have neglected to acknowledge the nuance that science is uncertain and often contested. The Cabinet Office guidance for the Scientific Advisory Group for Emergencies (SAGE), a team organised to provide scientific and technical advice and ensure informed governmental decision-making specifically states that:

"Transparency is an important element of democratic decision making and the evidence used to inform decision should be published."(Cabinet Office, 2012)

Despite this clear affirmation in support of openness, their approach has been less than forthcoming, exemplified by the government only officially announcing the names of those attending SAGE meetings on the 4th May after significant media pressure. The content of these meetings, which inform the UK government response have remained closed. Contrast this with New Zealand's Epidemic Response Committee, whose meetings are live-streamed online and later made available to watch on-demand and it is clear to see the range of approaches to transparency from national governments.

The push for transparency at this level is not simply a matter of curiosity but can also promotes democratic accountability. The decisions made during this crisis will affect us all and may cause or avoid significant harm at a population level. The policies implemented influence social, political and economic forces which have huge impacts on society and individuals.

Potential Drawbacks

There are of course some major drawbacks to providing Open Data which we need to remain vigilant about. Not least, to balance transparency against personal privacy. In our rush to promote access to data, we require a robust ethical justification as the potential to breach confidentiality increases.

The additional interest and urgency brought about by this situation also risks scientific rigour. The hugely increased availability of open datasets may encourage analysis by people who are acting beyond their competency as researchers or who may be drawing conclusions which are not justified by the data.

The huge increase in the uptake of pre-print approaches to publication reflects, at least in part, a desire to share results quickly and influence decision-making early on. Whether these publications acknowledge the limitations in emerging data sources or indeed are properly scrutinised, they are increasingly being shared on mainstream and social media platforms. Even if produced and shared in good faith, there remains a risk of amplifying misinformation or errors. Expert analysis and data from trusted sources should remain at the heart of our approach to tackling this crisis.

Despite these risks, there are a great many benefits to growing interest in and adoption to open access to data. One which is often not considered is that non-use of healthcare and routine administrative data has the potential to increase harm (Jones et al., 2017).

Without the capacity or will to analyse certain types of data, potentially useful information may go unused. Closed or incomplete datasets, which omit relevant information, risk poorly informed decision-making and negative outcomes.

It is imperative in our current landscape that data which could inform action to avoid harms are available for analysis. If non-use results from a lack of capacity or technical ability to analyse internally by institutions or governments, Open Science approaches provide a solution.

Conclusion

COVID-19 has had and will have a significant impact on people's lives; physically, socially and economically. The effects of both the virus and our measures to mitigate the virus will have serious consequences across the globe, although we cannot be certain for how long; and at a very human level, as friends or loved ones are put at risk.

The urgency to address this situation has massively accelerated the adoption of Open Science approaches, which offer huge potential when applied within appropriate ethical boundaries and in considering potential unintended consequences. Whether these changes, promoting transparency and collaboration in science and policy, are sustainable after the immediacy of this current crisis subsides, remains to be seen.

References

Cabinet Office. 2012. Enhanced SAGE Guidance: A strategic framework for the Scientific Advisory Group for Emergencies (SAGE).

Douglas, M., Katikireddi, S. V., Taulbut, M., Mckee, M. & McCartney, G. 2020. Mitigating the wider health effects of covid-19 pandemic response. BMJ, 369, m1557.

Houtkoop B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E. & Wagenmekers, E.-J. 2018. Data Sharing in Psychology: A Survey on Barriers and Preconditions. Advances in Methods and Practices in Psychological Science, 1, 70-85.

ICNARC (2020).Report on COVID-19 in critical care: April

Jones, K. H., Laurie, G., Stevens, L., Dobbs, C., Ford, D. V. & Lea, N. 2017. The other side of the coin: Harm due to the non-use of health-related data. Int J Med Inform, 97, 43-51.