Data & Policy (2021), 1–4

RESEARCH ARTICLE

Covid-19: An exploration of consecutive systemic barriers to pathogen-related data sharing during a pandemic

Yo Yehudi[1]*, Carole Goble[1] and Caroline Jay[1]

[1]Department of Computer Science, University of Manchester, Oxford Road, M13 9PL, United Kingdom.
*Corresponding author. E-mail: yochannah.yehudi@postgrad.manchester.ac.uk

Abstract
In 2020, the COVID-19 pandemic resulted in a rapid response from governments and researchers worldwide. As of March 2021, there have been over 122 million confirmed cases, with over 2.7 million people dying as a result of COVID-19. Despite this staggering toll, those who work with pandemic-relevant data often face significant systemic barriers to accessing, sharing or re-using this data. In this paper we report preliminary results for an in-progress qualitative study, where we interviewed data professionals working with COVID-19-relevant data types including social media, mobility, viral genome, testing, infection, hospital admission, and deaths. These data types are variously used for pandemic spread modelling, healthcare system strain awareness, and devising therapeutic treatments for COVID-19. Barriers to data sharing include cost of access to data (primarily certain healthcare sources and mobility data from mobile phone carriers), human throughput bottlenecks, unclear pathways to request access to data, unnecessarily strict access controls and data re-use policies, unclear data provenance, inability to link separate datasources that could collectively create a more complete picture, poor metadata standards, and a lack of computer-suitable data formats.

Policy Significance Statement
Preventing the spread of a global pandemic requires effective and swift access to high-quality and up-to-date data. Despite this urgency, non-private and semi-private data are often stored behind access-controlled systems that prevent effective re-use, and human time constraints for access requests result in bottlenecks, especially in high-demand scenarios instigated by a pandemic. Even when data can be accessed, poor data quality often makes it difficult or impossible for researchers to effectively use this data, and may necessitate lengthy workarounds or "best guesses" in order to create computational models that inform policy. To avoid such a costly death toll in future epidemics, it is imperative to implement effective computational data sharing pipelines and permissive re-use policies before another outbreak occurs.

1. Introduction

Barriers to effective use of data to inform policy are not new, but become noteworthy during a global pandemic that has already infected over 120 million and taken more than 2.7 million lives (Max Roser and Hasell (2020)). We discuss consecutive systemic barriers professionals encounter when working to address the effects of the pandemic, with intent to inform policy and create effective sharing mechanisms before another such deadly event occurs.

2. Research Question

In times of a pandemic or epidemic when rapid response is required, what are attitudes towards pathogen-related data sharing? In particular, what barriers do researchers encounter, and what do they do, if anything, to get around these barriers?

3. Methods and data used

To gain nuanced understanding of the barriers to data access, sharing, and re-use, we conducted a qualitative study, interviewing professionals who work with COVID-19 related data. Participants were researchers based in Europe or North America, and came from a range of domains, including data scientists, health professionals, epidemiology modellers, data policy experts, social media researchers, and computational genomics researchers. This study is still accepting new participants and aims to interview between ten and twenty individuals in total, with seven interviewed to date. For this preliminary analysis, interview answers were coded in NVivo to identify themes. A high-level theme overview is presented below.

4. Key findings

Throughout data lifecycles, from creation to analysis and dissemination, researchers repeatedly encounter barriers that hinder or prevent data use, battling inefficient systems that fail to put data in the hands of researchers and policymakers who need it most. We identify three key barrier stages: 1. initial data access, 2. difficulty using the data once access has been granted, and 3. prohibition of further data sharing and re-use.

4.1. Barriers to data access

The first hurdle researchers face is gaining access to relevant datasources. While this study focuses on access to non-private and non-clinical data, we still found that researchers faced systematic access barriers. **Cost** can range prohibitively from thousands to tens of thousands or more for mobile-phone generated mobility data. **Culture** is another barrier - sometimes access control was treated as a default without questioning if the data actually *needs* privacy-preservation measures. Third, **human throughput** creates bottlenecks: humans are gatekeepers for access control, and many data pipelines require intervention to run or human dissemination. These people may have been resource-stretched even before COVID-19. One participant asserted:

“If the mechanism for accessing data is "email someone and wait for them to respond", then if you don’t ideally have both professor and OBE involved in your name, you’re going to struggle.”

Access barriers may occur sequentially rather than concurrently, further delaying useful responses. Data access requests and grant applications typically turn around in weeks or months. If a pandemic modeller must first apply for a grant to purchase access to data, and only later request access to the data *if* funding is approved, the incumbent pandemic wave may pass before researchers can create realistic prediction models.

Finally, sometimes different datasets have the potential to meaningfully answer a question present in the other dataset, but nevertheless cannot be combined. This barrier may be technical - i.e. there is no common unique identifier with which to harmonise records across datasources - or it may be a sociolegal barrier: some licence terms prevent data remixing. One participant illustrates their experience:

"Having access to highly protected things, like healthcare data, for half of my work is bizarre, because for example, we need to know what proportion of people who enter the hospital over the age of seventy dies? What is that rate? I know it, because I generated it for one of my models, but I can't use it for the other model."

4.2. Hard-to-use data, even once available

If a researcher successfully gains access to privileged data, they face a new challenge. Access alone doesn't guarantee data is well-documented or easy to understand. As one participant asserts:

"Essentially, you had a treasure trove of information that was not at all mapped to each other, that a lot could have been done with, which was being heavily access managed and not at all curated."

Data may not be designed for computational analyses. Often data exists as Excel "data dumps" that must be manually downloaded by a human, or as PDFs, which aren't easily read by a computational script. Data quality issues are rife and "data cleaning" is time consuming. Additional reported examples include: no metadata or indexing (making it hard to find the right data to answer questions); data formats changing incrementally over time; multiple columns with the same names (and no explanation); and raw data existing in hidden sheets.

Geographical region definitions posed a noteworthy challenge: different sources may name the same regions differently - e.g. "Cambridgeshire" and "Cambs" are equivalent to human readers but stymie a computer script. Regions also change over time as boundaries and districts are re-defined, making it hard to reliably define regions without significant temporal context. Discussing difficulties converting from one geographical coordinate set to another, one participant stated "Guides on how to do that from the government come with a big warning sheet saying you should under no circumstances use this for any software that informs policy, which is not very helpful... because we have to."

The more quality issues researchers face, the more time required to process and clean data. Chances of errors in policy-informing analyses rise, with each error increasing the risk that researchers may fail to spot a serious problem in the data. Establishing a baseline of well-formatted and computationally-accessible data would be a big step towards mitigating data quality concerns and human throughput bottlenecks.

4.3. Barriers to sharing re-used data and data analyses

Once a researcher overcomes access and quality barriers, they may face difficulty disseminating the analyses they have created. Restrictive datasources may disallow redistribution of their data, which can result in analyses being impossible to reproduce or researchers simply abandoning the datasource entirely in favour of less restrictive datasources. Two researchers weigh in on datasources with strict or prohibitive re-use policies:

On Strava Metro, a city mobility datasource, which has a policy requiring all re-use to be approved by Strava:

"I spoke to one person from local government who just went "we're just not going to use it then" - there's just no value to it if you're going to provide that amount of restriction."

On GISAID, a viral genome sequence database: "The GISAID database which is containing the SARS-CoV-2 sequences is blocked off, and it's kind of not worth the hassle to use that data, even though it might be informative, because if we use it we then can't share our data, which is a derivative."

The European Commission with ELIXIR has established a pan-European open COVID Data Portal (ELIXIR (2021)) for public data deposition and access, and hosts an open letter with over 750 signatories, calling for open and FAIR data sharing (COVID-19 data portal (2021)).

5. Conclusion

The human cost of COVID-19 so far has been staggering. If we wish to ensure that we are in a position to mitigate future pandemics, we must collectively dismantle systemic barriers, and build policies which not only enable efficient re-use of non-private data, but also ensure the data itself is intentionally designed to be accessed computationally with a minimum of human input, reducing costly errors and deadly bottlenecks to informed decision making.

Competing Interests. None

Funding Statement. This work was supported in part by grant EP/S021779/1.

Data Availability Statement. Data for this study are stored on University of Manchester infrastructure, but are not available or deposited publicly in order to protect the privacy of the individuals who were interviewed.

Ethical Standards. The research was approved by the University of Manchester meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions. Y.Y. designed the study, gathered the data, wrote the first draft, and approved the final version of the manuscript. C.J and C.G. supervised ideation, study design, revised the manuscript and approved the final version.

Supplementary Material. State whether any supplementary material intended for publication has been provided with the submission.

References

COVID-19 data portal (2021). Open letter: Support data sharing for COVID-19. [Online; accessed 22. Mar. 2021].

ELIXIR (2021). ELIXIR support to COVID-19 research. [Online; accessed 22. Mar. 2021].

Max Roser, Hannah Ritchie, E. O.-O. and Hasell, J. (2020). Coronavirus pandemic (covid-19). *Our World in Data*. https://ourworldindata.org/coronavirus.