COVID-19 Data Representativeness versus Actionability: An Avoidable Trade-Off?

Nathanael Sheehan & Sabina Leonelli

2023-01-16

Abstract

This paper discusses a tension between actionability and representativeness that is often found in data science initiatives within the biological and biomedical sciences. On the one hand, engaging in data sharing efforts only makes sense if data can be used to support discovery, thereby becoming 'actionable' – a consideration that leads some advocates of open data to argue that any constraint on data circulation constitutes an obstacle to knowledge production. On the other hand, making data actionable for knowledge development presupposes that the data are representative of the phenomena being studied. This in turn assumes that: (1) enough data are contributed by a wide and diverse set of relevant sources; and (2) mechanisms of feedback and inclusion are set up to ensure that data contributors can participate data governance and interpretation, thereby helping to adequately contextualise data. All too often, the requirements for actionability and representativeness of data are conceptualised as incompatible and leading to a trade-off situation where increasing one will unavoidably decrease the other. Through an analysis of two different platforms used to share genomic data about the SARS-COV-2 virus, we critique this framing as damaging to data initiatives and infrastructures. We argue that the tension between actionability and representativeness can be negotiated by a model of responsible data sharing that enhances users' ability to work with data without sacrificing data protection measures and mechanisms of fair equitable governance, leading to an inclusive approach that maximises both representativeness and actionability. Crucially, such a model can only work when accepting that data do not need to be transparent, or even easily accessible, in order to be actionable; and that including a variety of contributors in efforts of data governance and interpretation may slow down the pace of discovery while boosting the robustness and quality of outputs.

Introduction

On the 29th of January 2021 the governing board of the European Bioinformatics Institute (EBI) posted a public letter in Nature calling for a greater "openness" in sharing SARS-CoV-2 genome data. The letter argued that "to unleash the fast flow of research advances" the scientific community must remove all formal barriers which restrict data sharing and share all SARS-CoV-2 genome sequences to one of a triad of state genomic surveillance programs (EBI, The GenBank of USA and the DNA Data Bank of Japan). The letter was signed and promoted by Nobel Laureates, Directors of Bioinformatic programs and many researchers at the cutting edge of genome sequencing. At the same time, the Global Initiative on Sharing Avian Influenza Data (GISAID) had just overtaken the EBI's European COVID-19 Data Portal (CV19-DP) in the volume of genome sequences being shared to open access databases. GISAID was launched in 2008 to monitor global influenza outbreaks and from the offset positioned itself as an alternative to the public domain sharing model. Its policy requires users to authenticate their academic identity and agree not to republish or link GISAID genomes with other datasets without permission from the data producer. This requirement stems from the recognition that some researchers – often working in low-resourced environments and/or less visible research locations – are reluctant to share data due to fears of better-equipped researchers building on such work

without due acknowledgment. Indeed, the GISAID model has fostered trust and information exchange among groups that differ considerably in their geo-political locations, funding levels, material resources and social characteristics, thereby expanding the range of data sources shared online. This proved decisive when, at the beginning of 2020, GISAID launched the EpiCov database which stores, analyses and builds evolutionary trees of SARS COV-2 genome sequences – now the leading open access database for SARS-CoV-2, with over 13 million genomes sequenced by November 2022. At the same time, limiting the extent to which data can be accessed and linked can negatively affect the insight, pace and breadth of future research – leading to the backlash by hundreds of leading researchers concerned about the urgency of an effective pandemic response.

This episode, and the ongoing debates underpinning the data sharing efforts by both GISAID and CV19-DP, signals a tension between actionability and representativeness that is often found in data science initiatives within the biological and biomedical sciences. On the one hand, engaging in data sharing efforts only makes sense if data can be used to support discovery, thereby becoming 'actionable' – a consideration that leads some advocates of open data, such as the signatories of the above-mentioned Nature letter, to argue that any constraint on data circulation constitutes an obstacle to knowledge production. On the other hand, making data actionable for knowledge development presupposes that the data are representative of the phenomena being studied. This in turn assumes that: (1) enough data are contributed by a wide and diverse set of relevant sources; and (2) mechanisms of feedback and inclusion are set up to ensure that data contributors can participate data governance and interpretation, thereby including domain/location-relevant expertise in specific instances data re-use, thus helping to adequately contextualise data. All too often, the requirements for actionability and representativeness of data are conceptualised as incompatible and as leading to a trade-off situation where increasing one will unavoidably decrease the other.

In what follows, we critique this framing and argue that the tension between actionability and representativeness can be negotiated by a model of responsible data sharing that enhances users' ability to work with data without sacrificing data protection measures and mechanisms of fair governance, leading to an inclusive approach that maximises both representativeness and actionability. Crucially, such a model can only work when accepting that data do not need to be transparent, or even easily accessible, in order to be actionable; and that including a variety of contributors in efforts of data governance and interpretation may slow down the pace of discovery while boosting the robustness and quality of outputs.

To exemplify our argument, we focus on the ongoing efforts to share genomic data about SARS-COV-2 and compare the characteristics of the GISAID and C19DB platforms with respect to both their representativeness and actionability. We use data and metadata on genetic sequences posted on the two platforms in the period between March 2020 and October 2022, thus capturing the period of arguably most acute response to the COVID-19 pandemic. Through a variety of aggregation methods, we compare the extent to which the two portals succeeded in attracting data submissions from different countries around the globe, and look at the ways in which submission rates varied over time. We also analyse the structure and underlying architecture of the platforms, reviewing how they organise data access and use, the types of metadata and version tracking they provide, and the scope they offer for data re-use. We conclude that GISAID and C19DB are both carrying out an extraordinary public service, with much that can be learnt from each around how to maximise both the actionability and the representativeness of the data in question, with great benefit not only to ongoing research on the COVID-19 pandemic, but also to future pandemics for which reliable and responsive data infrastructures may be urgently required.

Background

General background of urgent response requiring comprehensive and reliable data sources

`push for Open Science initiatives, including OA and OD`

`immediate focus on virology and research on SARS-COV-2 variants (Dupré and Leonelli 2021).`

Genetic data are relatively easy to share from a technical perspective, given their digital format and the fact that current initiatives build on decades of efforts by molecular biologists and bioinformaticians towards the fast and free sharing of sequencing data (Leonelli and Ankeny 2012, Maxson-Jones et al 2018, Strasser 2019).

Among most successful initiatives is GISAID, the Global Global Initiative on Sharing Avian Influenza Data. GISAID was launched in 2008, on the anniversary of the Spanish influenza, to foster the sharing of influenza genomic data securely and responsibly. Data sharing was immediately conceptualised not as straightforward opening up of the data by placing them all online without restrictions, but rather as an alternative to the public sharing model, whereby users agree to authenticate their academic identity and not to republish or link GISAID genomes without permission from data producer. The existence of a governance structure regulating data access and re-use is meant to foster trust and information exchange among groups that differ considerably in their geo-political locations, funding levels, material resources and social characteristics, thereby expanding the range of data sources shared online.This model proved very successful in relation to influenza research, and since its launch GISAID plays an essential role in the sharing of data among the WHO Collaborating Centers and National Influenza Centers for the bi-annual influenza vaccine virus recommendations by the WHO Global Influenza Surveillance and Response System (GISRS). It is no surprise therefore that GISAID was swiftly redeployed, in early 2020, to include SARS-COV-2 data

The EpiCov database stores, analyses and builds evolutionary trees of SARS COV-2 genome sequences. GISAID is now the leading open access database for SARS-CoV-2, over 13 million genomes sequenced by November 2022. It plays a key role in identifying and studying variants of interest, variant evolution and lineages, and spread in real-time; and indeed it features as key data provider for a wide variety of consortia, initiatives and projects devoted to the analysis of COVID-19 variants of interest (some of which listed on this page: https://gisaid.org/collaborations/enabled-by-hcov-19-data-from-gisaid/, many though not all funded and based in the Global North). Accordingly, GISAID is funded by a wide consortium of public and private bodies, including the Federal Republic of Germany, who first backed the project at its main site in Geneva, as well as public-health and academic institutions in Argentina, Brazil, China, Republic of the Congo, Ethiopia, Indonesia, Malaysia, Russia, Senegal, Singapore, South Africa and many other countries, as well as several donors and partners garnered under the label of "Friends of GISAID".

In April 2020, the European Commission, through the auspices of EMBL and Elixir, launched another platform for sharing scientific – including genomic – information of relevance to the biological study of COVID-19: the COVID-19 Data Portal (CV19-DP). CV19-DP is modular in nature, allowing EU countries to quickly develop their own versions of the data portal, such as https://covid19dataportal.es for Spain or https://covid19dataportal.pl for Poland. However, a main site also exists at https://covid19dataportal.org. The CV19-DP's stated goal is to "accelerate research through data sharing" and its primary strategy is to curate data to improve interoperability with other platforms, such as ENA, UniProt, PDBe, EMDB, Expression Atlas, and Europe PMC. The CV19-DP also facilitates rapid and open data sharing through a top-level API or direct bulk downloads, with minimal user tracking, in order to encourage linking and cross-analysis of a wide range of data types, including protein, expression, networks, imaging, and socio-economic data. The CV19-DP is committed to following the FAIR data principles, which prioritize making data Findable, Accessible, Interoperable, and Reusable.

The portal offers two key data visualization tools for the rapid analysis of the large volume of data stored on the system. The first is an open source phylogenetic tree that illustrates Covid-19 sequences that represent 98% of the SARS-CoV-2 template, including PANGO lineages, and are stratified by WHO regions. The second is the CoVeo browser, a proprietary software that performs a systematic analysis of raw reads from CV19-DP and presents the results in graphical and summary form for global regions, specifically examining VOCs and VOIs.

The tensions between GISAID and COVID-19 Data Portal: a trade-off between actionability and representativeness?

It can be argued that GISAID and CV19-DP represent two distinct and potentially conflicting strategies in the realm of data sharing, as reflected by their respective organizational structures and priorities. GISAID places a strong emphasis on data representativeness, or the extent to which its data constitutes a comprehensive and credible sample of the populations under study. This approach is based on the belief that such data is necessary in order to build trust among "sequence donors" and ensure the accountability of the organization. However, it is important to acknowledge that this focus on representativeness may come at the expense of actionability, or the ability to utilize data for a variety of research purposes and goals.

In contrast, CV19-DP prioritizes data actionability, utilizing interoperable standards and tools to facilitate the unexpected repurposing of its data. While this approach has the potential to generate novel insights, it is worth considering whether it may result in a lack of emphasis on data representativeness, or the comprehensiveness and credibility of the sample populations. This tension between representativeness and actionability raises important questions about the potential trade-offs involved in prioritizing one over the other, and the impact that these choices may have on the utility and reliability of the data produced by these organizations.

This is by no means limited to this case. Efforts to apply data protection laws such as GDPR to datasets used for research have also encountered strong resistance, particularly in the form of concerns that limiting access and exploration of existing datasets would result in much diminished capacity to use those data as evidence to support novel insights and research directions. In the absence of a sophisticated governance structure, and the funding to maintain it, this has indeed come to pass: many research institutions (such as hospitals, for instance) have become more conservative over their data sharing policies, making it extremely hard for researchers to be granted access and virtually impossible to simply 'explore' the data without a precise and pre-existing commitment to specific forms of re-use. This is a problem particularly for exploratory research based on data mining techniques, where access to the data is seen as a foundational requirement to be able to 'look around' and fish for surprising findings, promising correlations, and new hypotheses.

An ongoing challenge, at once ethical and epistemic: Different interpretations of openness and responsible research.

Comparing the two data infrastructures

We now compare the two data infrastructures on the basis of metadata acquired from the two portals between 2020 and 2022.

Methods

Correlation between the two databases

In order to examine the relationship between GISAID sequence submissions and COVID-19 data portal sequence submissions at United Nations geographical regions, we employed a Pearson correlation analysis. The data for this analysis was gathered from publicly available sources and included the number of GISAID sequence submissions and COVID-19 data portal sequence submissions for the specified geographical regions. We chose to utilize Pearson correlation due to its ability to measure the strength and direction of a linear relationship between two variables. A scatterplot was generated to visualize the data, and the Pearson correlation coefficient was calculated and interpreted to determine the nature of the relationship between GISAID sequence submissions and COVID-19 data portal sequence submissions in these regions. It is important to note that Pearson correlation can only assess linear relationships and may not accurately represent the relationship between the variables in other forms. Despite this limitation, this methodology

allowed us to investigate the potential association between GISAID sequence submissions and COVID-19 data portal sequence submissions in United Nations geographical regions.

Global aggregations of Sequence and Epidemiological Data

Recent studies have begun to explore the heterogeneity of the global genomic surveillance landscape for the SARS-CoV-2 virus using metadata from surveillance databases (Khare et al. 2021; Samlali 2021; Brito et al. 2022). To date, no explicit comparison has been made between submissions to the GISAID database and its counterpart, the Covid-19 Data Portal. Considering the independent commission report from the UK government often cites the Chen et al. (2021) paper, despite the lack of consideration for the Covid-19 Data Portal or the European Neucaloide Archive as data sources, this comparison is especially pertinent. To address the gap in existing academic research, our analysis collected over 19 million metadata points for the respective databases between the epidemiological weeks of 23 December 2019 and 1 October 2021. Epidemiological data from the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (JHU) (https://github.com/CSSEGISandData/COVID-19) was used to report on global new case counts. The data sources were linked together using country codes defined by ISO 3166-1 alpha-2. Percentages of Covid-19 cases sequenced per country - cumulatively and weekly - were calculated by filtering the data into weekly submissions and aggregating counts per country, continent and income group, using a custom R script (https://github.com/natesheehan/OPEN-GM/blob/main/code/data-wrangle.r). Visualisations were created using the ggplot2 visualisation package and are accessible via the following open access scripts (https://github.com/natesheehan/OPEN-GM/blob/main/code/plot-temporal-submissions.r, https://github.com/natesheehan/OPEN-GM/blob/main/code/plot-treemap.r, and https://github.com/natesheehan/OPEN-GM/blob/main/code/plot-continent-landscape.r).

TCP/IP stack fingerprinting

TCP/IP stack fingerprinting (OS fingerprinting) is a function often used by hackers - ethical and not so ethical - to find out the characteristics of a system they may or may not have access too. As one author in the popular open source security package `nmap` puts it: "The legal ramifications of scanning networks with Nmap are complex and so controversial that third-party organizations have even printed T-shirts and bumper stickers promulgating opinions on the matter". OS fingerprinting works by remotely accessing a number of features in the TCP/IP stack implementation and comparing them to previously defined combinations of parameters to infer matches. For the purpose of this study, we deploy this function to uncover the various degrees of openness each database is designed with, as well as the various services and steps in place to access SARS-CoV-2 metadata. The open source `whatweb` command line tool was employed to retrieve data regarding geographical location, author, type of server, and different types of plugins/libraries present in the system. The following command was used for each portal on October 31st 2022: `./whatweb -v https://www.dataportal-url.org -a 1`, where -v provides a verbose output of the results and -a 1 represents a soft level of pen test. The complete output for each data portal can be found in the supplementary materials.

Data and Metadata mapping

On October 30th, 2022, a systematic manual collection of data and metadata was conducted across multiple platforms in order to access and analyze these records. In order to standardize the collection process, a consistent sequence was employed for each portal, and the number of "hops" was recorded in order to trace the infrastructural lineage of the data. In addition to the primary data, supplementary metadata associated with each sequence was also carefully documented. Access to data through GISAID was strictly governed by their terms of use policy.

Results

The GISAID database and the Covid-19 Data portal are both important resources for understanding the spread and impact of the SARS-CoV-2 virus. A closer examination of the data, however, reveals significant imbalances in the distribution of submissions between the two databases. As a data user, one's initial point of access to the GISAID platform will likely be through its hosting on free/open source software in Germany. In order to utilize the EpiCov3 database, which serves as the primary means of accessing GISAID data, it is necessary to register. This database is hosted on closed software in the United States, and presents aggregated data in nine columns that correspond to three submetadata categories. It is noteworthy that there is a direct mapping between the data and metadata, with each column in the aggregated data typically corresponding to a submetadata category containing more detailed, potentially sensitive information. In addition, there is a separate metadata category that includes highly personal information about the sequence submitter. The Covid-19 Data portal, hosted on a secure LAMP stack in the United Kingdom, serves as an alternative means of accessing sequence data. While queries made through this portal are directed to the European Nucleotide Archive database and returned as aggregated data, the portal does not directly host any data. To view the metadata associated with a particular sequence, a data user will be redirected to the ENA's application. It is worth noting that there is a disconnect between the naming conventions used for aggregated data and metadata within this system. While certain labels remain consistent, such as the accession ID, others are altered upon being pooled into the Covid-19 Data portal, e.g. "center name" (data) becomes "collection institution" (metadata). This linking strategy obscures the direct correlations between the data and the five submetadata categories, although two of the metadata categories hosted by the ENA do contain information about linked studies, samples, and taxa.

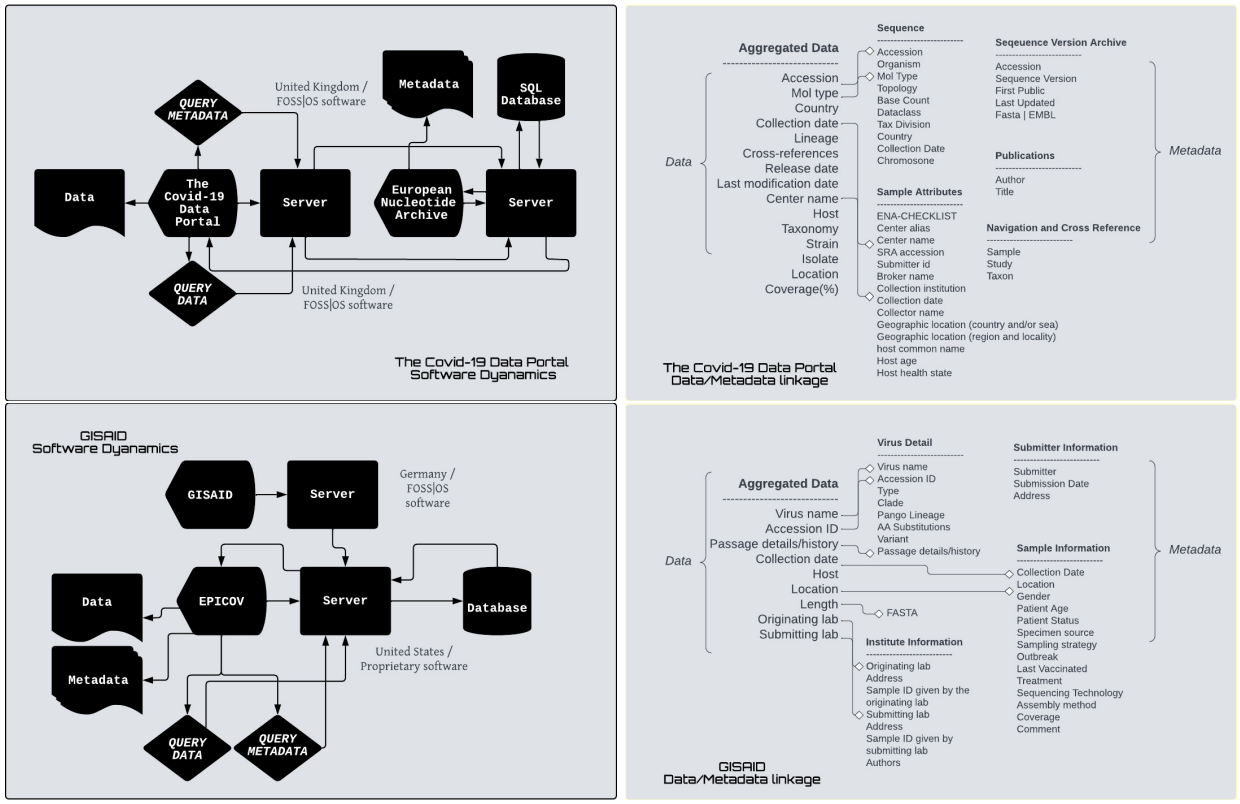

Figure 1: TCP Fingerprinting and data and metadata mapping for each database.

In our analysis of the metadata from the genome sequencing of SARS-CoV-2, we found spatial and temporal heterogeneity across the two databases. At the global level, there is a strong positive correlation between the two databases in terms of the weekly average number of submissions (see figure 1). For instance, an increase

in submissions to GISAID is correlated with a corresponding increase in the number of sequences submitted to the COVID-19 data portal at the global level. At sub-regional geographical scales (as seen in the UN dataset), there are highly significant correlations in several regions, including Northern America, Northern Europe, Western Europe, Eastern Europe, South-eastern Asia, Sub-Saharan Africa, Western Asia, Micronesia, Northern Africa, and Eastern Asia. However, some regions, such as Latin America, North Africa, Australia and New Zealand, show a discrepancy between surveillance linkage, which may be due to the differences in volume between the two data portals. Specifically, GISAID receives 40% more viral sequences than the COVID-19 data portal, particularly from low- and middle-income countries and some high-income countries that submit more to GISAID than the covid-19 data portal.

According to recent studies, a minimum of 5% of SARS-CoV-2 positive cases should be sequenced in order to detect viral lineages at a prevalence ranging from 0.1% to 1.0% (REF). Using this benchmark, we evaluated the performance of the GISAID and Covid-19 Data portals in terms of achieving this minimum genomic surveillance requirement. We further analysed the geographical, regional, and income distribution of submissions to these portals. The GISAID database has 194 unique countries submitting SARS-CoV-2 sequence data, with 40.21% of these countries submitting over 5% of new cases. The top five countries achieving the 5% minimum were all high-income countries: Japan, Australia, the United Kingdom, Denmark, and Canada. These countries had outstanding genomic surveillance output, with submission rates of 100%, 99.21%, 99.21%, 96.70%, and 95.70%, respectively. At the regional scale, the most productive regions were North America, North Europe, and Western Europe, with all other regions representing less than 5% of submissions. In terms of income groups (see UN dataset for reference), the GISAID database is composed of 93% high-income countries, 5.50% upper middle-income countries, 2.30% lower middle-income countries, and 0.06% low-income countries. On the other hand, the Covid-19 Data portal has significantly fewer submissions, with only 114 unique countries contributing data. Of these, only 20.41% represent a 5% capture of new cases in a country. The leading countries in terms of submissions are Liechtenstein, the United Kingdom, New Zealand, Djibouti, and Iceland, with submission rates of 72.78%, 66.66%, 65.98%, 60.54%, and 54.42%, respectively. The most productive regions were North America, North Europe, and Western Europe, with all other regions representing less than 1% of submissions. In terms of income group distribution, the Covid-19 Data portal is heavily skewed towards high-income countries, with 98% of submissions coming from this group and all other income groups representing less than 1% of submissions.

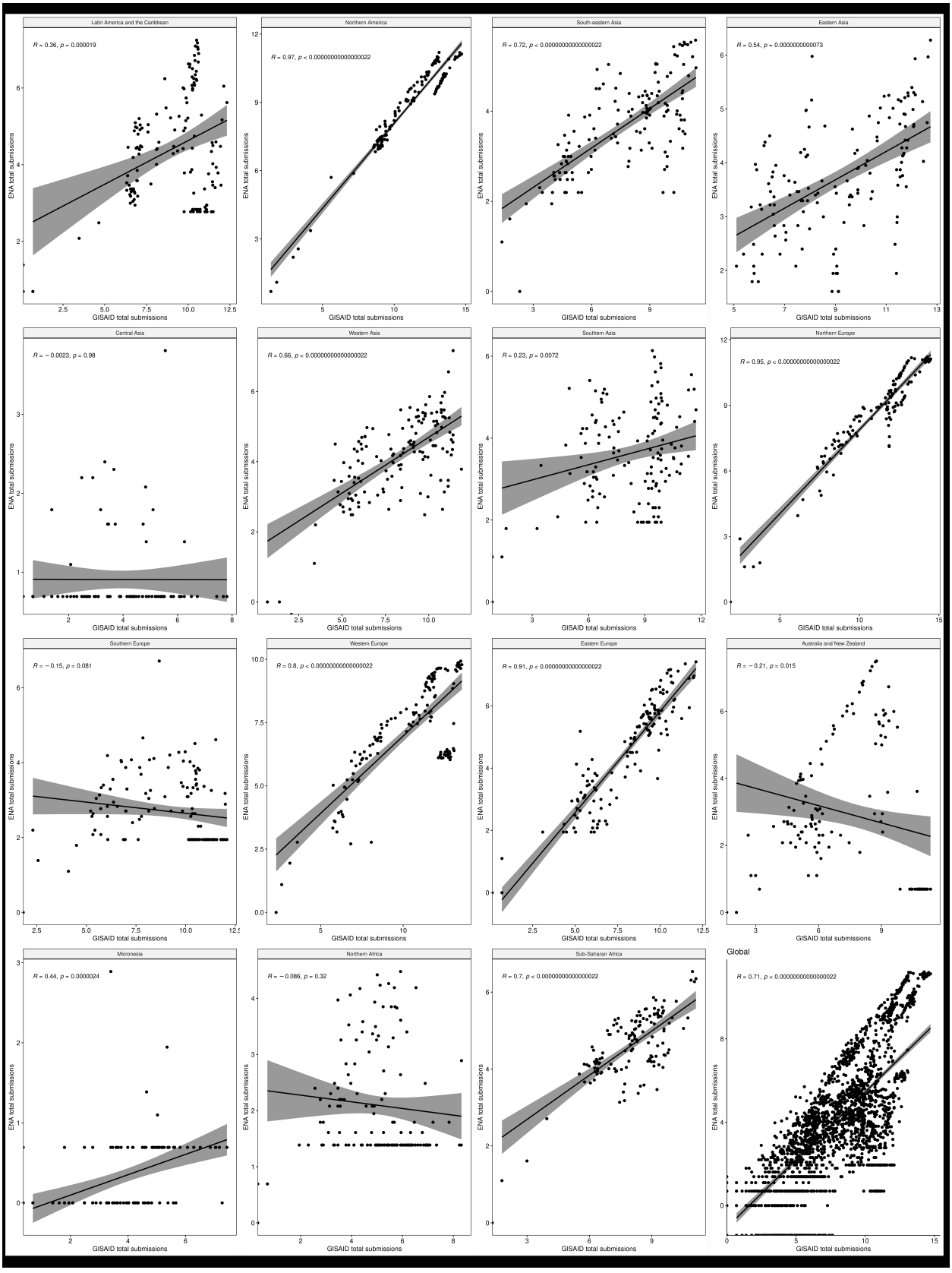

Figure 2: Figure showing the Pearson correlation between GISAID sequence submissions and COVID-19 data portal sequence submissions at United Nations geographical regions. The strong positive correlation indicates that as GISAID sequence submissions increase, COVID-19 data portal sequence submissions also tend to increase, suggesting a relationship between the two variables in these specific regions.

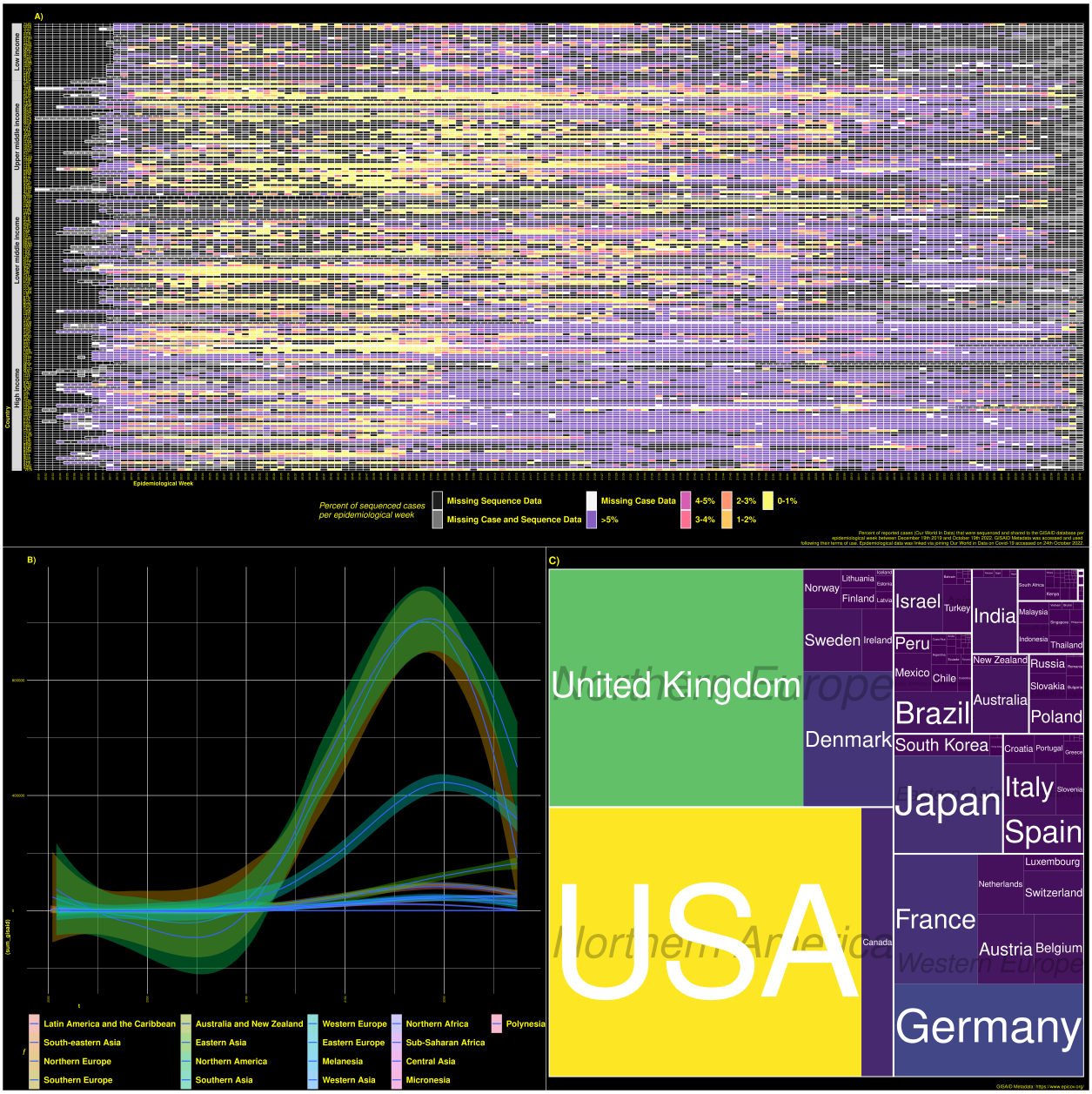

Figure 3: Epidemiological surveillance of COVID-19 through sequencing efforts. Panel A displays the percentage of cases sequenced for each country, organized by epidemiological week from February 23, 2020 to March 27, 2021. Panel B shows the distribution of submissions by geographical region. The geographical data provenance for the GISAID database is visualized in Panel C using a treemap representation.

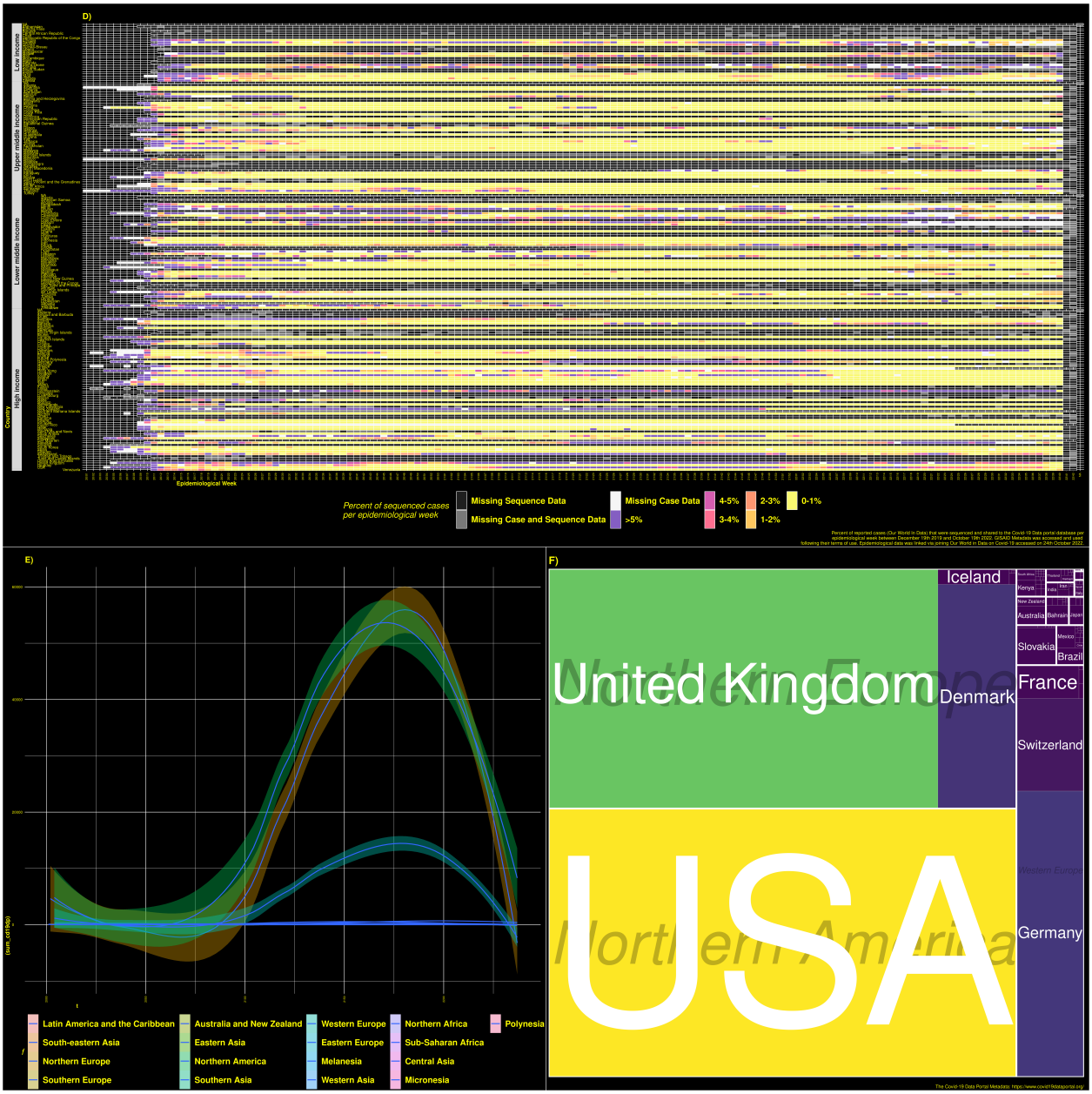

Figure 4: Epidemiological surveillance of COVID-19 through sequencing efforts. Panel A displays the percentage of cases sequenced for each country, organized by epidemiological week from February 23, 2020 to March 27, 2021. Panel B shows the distribution of submissions by geographical region. The geographical data provenance for the Covid-19 Data Portal is visualized in Panel C using a treemap representation.

Challenging the trade-off framing

It can be argued that GISAID and CV19-DP exemplify opposing strategies in data sharing, as demonstrated by their respective organizational structures and priorities. CV19-DP prioritizes data actionability, or the capacity to utilize data for novel research purposes and goals, while GISAID prioritizes data representativeness, or the extent to which its data constitutes a comprehensive and credible sample of the populations under study. However, it is important to note that there is often a perceived tension between these two approaches, with some arguing that a focus on representativeness may come at the expense of actionability, and vice versa.

In the case of GISAID, the organization places a strong emphasis on ensuring the comprehensiveness and accountability of its data, with the goal of building trust among its "sequence donors." However, it is worth considering whether the priority on representativeness may lead to an uneven playing field within the scientific community, with those who possess the most advanced methodological tools potentially having an advantage in terms of data inclusion. As such, it is important to consider the role of appropriate data governance and technological choices in ensuring the representativeness of GISAID's data samples.

On the other hand, CV19-DP prioritizes making data "reusable" rather than simply accessible, utilizing interoperable standards and tools to facilitate unexpected repurposing of the data. While this approach may have the potential to generate novel insights, it is crucial to consider the potential trade-offs involved in prioritizing actionability over representativeness.

Conclusion

References

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. The findings of this study are based on metadata associated with 8,949,097 sequences available on GISAID up to March 18, 2022, via 10.55876/gis8.220927xo.

Competing interests

The authors declare no competing interests.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101001145). This paper reflects only the author's view and that the Commission / Agency is not responsible for any use that may be made of the information it contains. We would also like to thank the Our World in Data and the Covid-19 Data Portal team for their efforts in collecting and curating the epidemiological data used in this study.

Author contributions

S.L. conceived and designed the study and wrote the manuscript. N.S. assisted in the design of the study, performed the analysis and discussed the results.