Data challenges for international health emergencies: lessons learned from ten international COVID-19 driver projects

Sally Boylan, Catherine Arsenault, Marcos Barreto, Fernando A Bozza, Adalton Fonseca, Eoghan Forde, Lauren Hookham, Georgina S Humphreys, Maria Yury Ichihara, Kirsty Le Doare, Xiao Fan Liu, Edel McNamara, Jean Claude Mugunga, Juliane F Oliveira, Joseph Ouma, Neil Postlethwaite, Matthew Retford, Luis Felipe Reyes, Andrew D Morris, Anne Wozencraft

The COVID-19 pandemic highlighted the importance of international data sharing and access to improve health outcomes for all. The International COVID-19 Data Alliance (ICODA) programme enabled 12 exemplar or driver projects to use existing health-related data to address major research questions relating to the pandemic, and developed data science approaches that helped each research team to overcome challenges, accelerate the data research cycle, and produce rapid insights and outputs. These approaches also sought to address inequity in data access and use, test approaches to ethical health data use, and make summary datasets and outputs accessible to a wider group of researchers. This Health Policy paper focuses on the challenges and lessons learned from ten of the ICODA driver projects, involving researchers from 19 countries and a range of health-related datasets. The ICODA programme reviewed the time taken for each project to complete stages of the health data research cycle and identified common challenges in areas such as data sharing agreements and data curation. Solutions included provision of standard data sharing templates, additional data curation expertise at an early stage, and a trusted research environment that facilitated data sharing across national boundaries and reduced risk. These approaches enabled the driver projects to rapidly produce research outputs, including publications, shared code, dashboards, and innovative resources, which can all be accessed and used by other research teams to address global health challenges.

Lancet Digit Health 2024; 6: e354–66

Health Data Research UK, London, UK (S Boylan MSc, E McNamara LLM, N Postlethwaite BSc, M Retford BCompSci, Prof A D Morris MD, A Wozencraft PhD)**; Department of Global Health, Milken Institute School of Public Health, George Washington University, Washington, DC, USA** (C Arsenault PhD)**; Center for Data and Knowledge Integration for Health, Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, Brazil** (M Barreto PhD, A Fonseca PhD, M Y Ichihara PhD, J F Oliveira PhD)**; Evandro Chagas National Institute of Infectious Disease, Oswaldo Cruz Foundation, Rio de Janeiro, Brazil** (F A Bozza PhD MD)**; Aridhia Informatics, Glasgow, UK** (E Forde PhD)**; St George's, University of London, London, UK** (L Hookham MBBS, Prof K Le Doare PhD)**; Green Templeton College** (G S Humphreys PhD) **and Nuffield School of Medicine** (L F Reyes PhD MD)**, University of Oxford, Oxford, UK; Makerere University John's Hopkins University Research Collaboration, Kampala, Uganda** (Prof K Le Doare, J Ouma PhD)**; Department of Media and Communication, City University of Hong Kong, Hong Kong Special Administrative Region, China** (X F Liu PhD)**; Partners in Health, Boston, MA, USA** (J C Mugunga MD MS)**; Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, USA** (J C Mugunga)**; Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA, USA** (J C Mugunga)**; Department of Mathematics, Centre of Mathematics of the University of Porto, Porto,**

Introduction

Data are at the core of patient care, population health management, health-service planning, and research. Amid the loss of health and life to COVID-19, researchers and policy makers engaged with data and information more intensively than ever, but major challenges to international data access and sharing were exposed. These challenges led the G7 nations to prioritise the establishment of health data as a global public good.[1] As such, the COVID-19 pandemic highlighted the importance of effective, equitable, and ethical data sharing, and the power of timely data sharing to provide crucial new insights and improve health-care outcomes.[2] Initiatives and systems to enable sharing of health-relevant data were established,[3] and many researchers, particularly in lockdowns, focused their efforts on maximising the benefits from secondary data.[4] In contrast to primary data collection, secondary data refers to information collected for purposes other than the intended research.[5] These data can come from a wide range of pre-existing sources including disease registries, health management and planning systems, clinical care records, and epidemiological surveillance tools. Due to the broad spectrum of secondary data sources, accessing, preparing, and analysing these data for research purposes presents numerous and distinct challenges.[6]

The International COVID-19 Data Alliance (ICODA) programme was convened by Health Data Research (HDR) UK in July, 2020. Its aim was to make existing, health-relevant research data accessible to researchers everywhere, enabling them to address key questions relating to COVID-19 and provide new insights that led to improved health outcomes for all, with a particular focus on low-income and middle-income countries. To achieve this vision, it assembled an open, international alliance of partners that brought together stakeholders, including community and patient representatives, to shape the programme and show trustworthiness. This approach built on that of other initiatives that have brought together international communities of health-care and research organisations to develop agreed standards and frameworks for the ethical use of health data to address global health challenges, such as the Global Alliance for Genomics and Health[7] and the International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC).[8]

The ICODA initiative made use of the Five Safes framework,[9] first developed by the UK's Office for National Statistics and adopted by HDR UK, as an approach to enable the right to privacy while unlocking the power of data for research (panel 1). ICODA focused on improving data discoverability through the ICODA Gateway, a dataset catalogue and access tool, and on providing researchers with a trusted research environment (TRE) called the ICODA Workbench, a highly secure and controlled computing environment that allowed approved researchers from authorised organisations a safe way to access, store, and analyse sensitive data remotely.[12] Working in partnership with data curation and databank providers, the Workbench enabled secure data access and analysis for data partners and researchers from over 70 countries, as well as a collaborative space for their projects.

Central to ICODA's approach was a cohort of 12 exemplar or driver projects, for which the core team

Key messages

- Appropriate data infrastructure, governance, and support should be provided early, to enable insights to be generated rapidly.
- Standard data sharing agreements and templates can speed up what can typically be a lengthy process.
- Use of pre-provisioned trusted research environments can go a long way to opening up data sharing across national and regional boundaries; expediting this process can be crucial in research areas such as rare diseases, where national datasets might be too small to give rise to significant results. It also provides a good mechanism for reducing the risk involved in data sharing, as the data remains within a secure environment at all times.
- Use of data curation expertise early on in initiatives can accelerate progress as this step is typically time-consuming and often underestimated. As part of this curation, considering making data findable, accessible, interoperable, and reusable at the same time and considering field labelling and units can reduce the work involved in sharing metadata.
- Community and wider stakeholder engagement requires an investment of time and resources, but is crucial to building trust and ensuring that relevant research questions are addressed and insights taken up into policy and practice.
- The Five Safes framework resonates with a global audience and provides an understandable and readily accessible framework with broad applicability.
- Willingness to conduct open science is crucial not only to making datasets discoverable to other researchers, but also to share code, methods, and lessons learned. It can also improve trust and transparency and, in future, promote easier sharing of data.
- Bringing together cohorts of researchers, even if working on disparate research questions, results in synergy and rapid identification of problems that need to be resolved for multiple parties.

Panel 1: International COVID-19 Data Alliance (ICODA) implementation of the Five Safes framework

(1) Safe people: ICODA implemented a proportionate researcher accreditation process to ensure data are only accessed by those who are trained and trusted to use it appropriately[10]
(2) Safe settings: a secure trusted research environment, the ICODA Workbench, was provided for use by driver projects
(3) Safe projects: the Grand Challenges ICODA open funding call review identified projects with rigorous project management and delivery plans in place
(4) Safe data: teams were supported in ensuring project data were de-identified
(5) Safe outputs: an output review process was developed to ensure outputs were non-disclosive and, where appropriate, validated the scientific integrity of results[11]

Portugal (J F Oliveira); **Universidad de La Sabana, Chia, Colombia** (L F Reyes)

Correspondence to: Sally Boylan, Health Data Research UK, London NW1 2BE, UK
sally.boylan@hdruk.ac.uk

For more on **ICODA** see https://icoda-research.org/about/about-us/

For more on the **ICODA Workbench** see https://icoda-research.org/research/our-research/

For more on the **ICODA driver projects** see https://icoda-research.org/research/driver-projects/

worked in close partnership with data contributors and the research community. Each driver project addressed important COVID-19-related research questions and ran at the same time as the ICODA programme was being developed, serving to test and shape ICODA's processes and tools as overall approaches were being developed for the long term. The projects allowed ICODA to bring together data, make them accessible, and generate valuable insights. The initial driver projects, Efficacy and Safety of COVID-19 Treatments and International Perinatal Outcomes in the Pandemic, were the first to co-create and test these approaches and provided valuable learning that benefited subsequent driver projects.[13,14]

Following an international call for proposals, a further ten driver projects were identified and established in July, 2021, through 12-month Grand Challenges ICODA research awards, and these are the main focus of this Health Policy (table). These projects addressed major research questions relating to the COVID-19 pandemic, such as the effectiveness of community-based vaccination programmes and the effect of the pandemic on health-service delivery. All used innovative data science approaches applied to existing health data and aimed to help address global inequity in access, the quality and use of data, test new approaches to support ethical and trustworthy health data use, and deliver rapid outputs and insights. Some of the projects involved large teams and datasets from multiple countries, with many researchers based in low-income and middle income countries (figure 1).

Despite the many challenges that remain in health data reuse, the ICODA programme supported 135 researchers from 19 countries to access and analyse a broad range of data types, with the ten Grand Challenges ICODA driver projects enabling access to datasets from over 70 countries. To date, this cohort of ten driver projects has generated 57 outputs, including publications, processes, dashboards, datasets for secondary analysis, and code.[15] Following review of the data research cycle used by each driver project, the lead researchers and ICODA team have worked together to identify common challenges in using secondary, health-relevant data in a global context to provide rapid insights and set out solutions they used that could be applied by others across other health challenges and data types.

Methods

The cohort of ten Grand Challenges ICODA driver projects all focused on major research questions related to the COVID-19 pandemic, but used different types of secondary data from a wide range of sources and contexts. As the projects were established at similar times, the ICODA team worked with the project leads to conduct a retrospective analysis of the challenges that each project faced in going through the relevant stages of the health data research cycle, and to collect project timeline data that clearly identified where common challenges and barriers existed across the cohort. The teams also worked together to identify the approaches and solutions that helped address these challenges, and which could be used by other researchers taking data science approaches to address a wide range of different health challenges.

Common process steps were selected for measurement in each of the projects, following the broad steps for the

	Main aims	Types of data and data source	Country	Key challenges	Key solutions
The PRIEST study for low- and middle-income countries (DP-PRIEST)	To ensure hospitals in low-income and middle-income countries are not overwhelmed during the COVID-19 pandemic by developing a risk assessment tool for clinicians to quickly decide whether a patient needs emergency care or can be safely sent home.	Existing data for 50 000 patients with suspected COVID-19 and who sought emergency care.	UK, South Africa, and Sudan	Obtaining additional approvals; linking and cleaning the datasets took longer than anticipated; projects involving teams distributed across the world require different methods for co-ordination than when all located in one place.	Team ultimately required a 3-month, no-cost extension to complete all planned analysis—a key lesson for use of any routine datasets. The ICODA initiative made expert curation support available and the use of natural language processing in data curation meant that the standardisation of the datasets could be achieved; the ICODA Workbench was extremely useful in providing a central repository of datasets to facilitate standardisation and analysis. There was a clear division of labour and responsibility between the different teams.
Addressing critical COVID-19 questions through research using linked population data (DP-ACCORD)	To understand COVID-19 evolution and impact, also on pregnancy and chronic diseases, by applying a data science approach to health data to study the clinical epidemiology and evolution of a new SARS-CoV-2 variant, which emerged in South Africa.	Anonymised COVID-19 health data from the government health department including >1 million tests and 60 000 hospital admissions in the Western Cape province of South Africa.	South Africa	The biggest challenge was dedicating person-time to the analyses when there were resurgences and competing service priorities. This challenge could be unique to the project being led by health-service staff. Outcome (COVID-19 relatedness of morbidity and mortality) and exposure ascertainment (previous infection) became increasingly challenging over time.	Where the team were able to automate the updating of analysis datasets, repeating analyses as the epidemic progressed became progressively easier. This enhancement is reflected in the severity analyses based on a standard case cohort, which was updated daily.
Effectiveness of COVID-19 vaccination in Brazil using mobile data (DP-EFFECT)	To quantify the real-world value of COVID-19 vaccines for protecting individuals from severe disease, and for protecting the entire population from being infected.	Data from the national vaccination programme, as well as deaths and cases at a municipality level and from 43 hospitals.	Brazil	Research teams in low-income and middle-income countries might not always have worked within what are considered international best practice and standards for data governance and data sharing; access and analysis of health data are restricted to researchers and government representatives, and not available for the health workers and communities to rapidly respond to the crises.	The project provided an opportunity to update the team on current best practices in data governance, privacy, and sharing guidelines, using an open science approach. This was important to maintain good collaborations with international and national networks, based on research reproducibility and transparency; the team also developed dashboard monitors that enabled the local health and research teams and the local community to easily access and visualise data from the cohort studies. Using a dissemination strategy plan, the team communicated the results of their research to local health practitioners and residents during meetings and conferences.
Routine assessment of infections, prevention, and control of SARS-COV-2 in unequal populations (DP-RASUP)	To study COVID-19 transmission issues in socially and economically unequal populations, accounting for human behaviour, non-pharmaceutical interventions, and vaccine strategies. By developing a user-friendly surveillance platform, the community could follow up its risk and jointly contribute to decrease cases and mortalities.	Five datasets were reviewed: de-identified surveillance data—daily time-series of cases and deaths of confirmed SARS-CoV-2 infections (n=2 005 200); de-identified information on gender, age, comorbidities, and infectious status (n=345 281); socioeconomic determinants data at summary level—classifies municipalities according to welfare benefits measured by income, literacy, and housing (n=5570); human mobility data at summary level—variables affecting human mobility, given by both Google trends and the historical average daily flux data throughout the country using road, air, and fluvial networks (n=65 638 [daily flux data]; n=754 095 [total state time-series]); stringency index at summary level—a metric that the team constructed that summarised the level of governmental measures enacted by local states (n=803).	Brazil	Processing the data for real-time analyses and providing results to the community in an optimal way; data access is still a challenge. The team collected a large amount of data to understand the COVID-19 pandemic in Brazil. However, statistical and mathematical modelling is challenging, due to the lack of studies that adequately account for different sub-populations; development of communication materials tailored for targeted populations with different backgrounds and priorities.	Invest time in building a systematic data pipeline tailored to the needs of the project; redesign aspects of the project to incorporate real-world variables influenced by population inequalities into the literature; initially, journalists were primary consumers of the team's results, aiming to disseminate scientific knowledge about the spread of SARS-CoV-2 in the country to the public. However, recognising the importance of understanding the priorities of the local community, the team started to work with local communicators in the favelas of Salvador.

(Table continues on next page)

	Main aims	Types of data and data source	Country	Key challenges	Key solutions
(Continued from previous page)					
Evaluating social inequalities and their effects on the COVID-19 pandemic in a low-income and middle-income country (DP-IDS-COVID19)	To measure the social disparities to understand the extent to which, in terms of socioeconomic factors, demographic factors, and access to health care, vulnerable people become unwell and die from COVID-19.	Administrative data collected routinely by government institutions and publicly available. These included: Brazilian Census 2010, National Register of Health Facilities, Influenza Case Notification System, National Total Population and Estimated Age Groups for 2020, and Brazilian Index of Deprivation database.	Brazil	The main challenge was in the construction of the COVID-19 social disparity index. This was designed to define indicators that were statistically correlated with each other and which provided evidence of social determinants for COVID-19 infection. This was presented as a publicly available and interactive dashboard and there were challenges in ensuring best user experience. The short project duration was also challenging.	The team used the regionalisation of health services categorisation used by Brazil's publicly funded health care system, Sistema Unico de Saude. The social disparity indicators that were identified were reviewed by a range of stakeholders and this helped ensure consistency between regions. The team also carried out usability tests for the online dashboard to enhance the user experience.
Disruptions in clinical outcomes and care among patients with chronic conditions: a four-country retrospective cohort (DP-PIH.CovCo)	To understand how the COVID-19 pandemic has affected care provision, care use, and health outcomes among chronic care patients—specifically those with HIV, cardiovascular disease, or diabetes.	A retrospective cohort study among patients with diabetes, HIV, or hypertension receiving care at Partners in Health-supported facilities. The dataset included 111 252 electronic medical records of chronic care patients.	Haiti, Malawi, Mexico, and Rwanda	Timeline for extracting electronic medical records; navigating complex electronic medical records datasets and in varying formats.	Frequent discussion with multiple-site level and cross-site level team members to assist in facilitating the data. Also creating a staggered analysis plan to focus first on HIV patients and then on cardiovascular and diabetes patients; programming rules were designed to assist with data cleaning, while also discussing metrics among site-level experts.
Characterising COVID-19 transmission chains for precision mitigation using epidemiological survey data (DP-CHAIN)	To reconstruct transmission chains between individuals in households and communities, and study COVID-19 transmission patterns from the reconstructed transmission chains.	Line-list datasets included confirmed cases of COVID-19 with detailed information enabling linkage to other cases, such as close contacts and simultaneous presence in the same location. Data were compiled from publicly available case reports, released by governments, or extracted from published studies. The dataset encompasses approximately 40 000 cases from four Asian countries and regions.	China, Taiwan, Hong Kong, and Singapore	The initial challenge lay in the data merging process. This involves integrating various datasets that adhere to different standards and organisational structures. The task required identifying common elements within these disparate datasets and successfully combining them; the information extraction process, which involved extracting specific details such as dates, named entities, and the relationships between these elements from the texts. The process is both time-consuming and resource intensive.	The first solution involved a dedicated team of 20 data collectors who monitored government announcements daily to gather reports. This systematic approach ensured consistent and up-to-date data collection, using a high-standard human coding system. This system involves two independent coders for initial data processing, followed by a resolver to address any discrepancies. This method enhances the accuracy and reliability of the data, using computer-aided coding. This involves using the data processed by human coders to train neural networks. The performance of these neural networks is on par with that of human coders, thereby combining the benefits of both manual expertise and automated efficiency.
Assessing the resilience of health systems during COVID-19 using routine data (DP-REHCORD)	To assess the magnitude of disruptions for non-COVID-19 essential health services during the COVID-19 pandemic in ten countries.	In each country, the team compiled administrative or routine health information system data on the number of health services provided from 2019 to 2021. In Ethiopia, Ghana, Haiti, Laos, Nepal, and South Africa, the data were extracted from health management information systems. In the other countries, the team used data from the Sistemas de Información del Ministero de Salud (Chile), Sistema de información del Instituto Mexicano del Seguro Social (Mexico), the National Health Insurance Service Health Facility Claims Database (South Korea), and the National Health Database of the Ministry of Public Health (43-folders dataset; Thailand).	Chile, Ethiopia, Ghana, Haiti, Laos, Mexico, Nepal, South Africa, South Korea, and Thailand	Data harmonisation: several indicators had different definitions across countries; data cleaning: missing values continue to be an important problem in health management information systems data in many countries; lack of master facility lists: in some countries, it was difficult to obtain an official count for the number of facilities that should be reporting every month.	Data harmonisation: multi-country codebooks with clear definitions highlighting differences by country were developed; data cleaning: a standardised data cleaning process was developed to exclude health facilities with sparse reporting. The data cleaning code was made available on GitHub for improved transparency and reproducibility; lack of master facility lists: to assess completeness, the team used the maximum number of facilities reporting in any given month as the estimated maximum number of facilities. In some countries, the team had to rely on aggregate analyses at district or provincial levels.

(Table continues on next page)

health data research cycle set out in figure 2. Some of these steps were time limited for the projects to complete, such as data curation or initial analysis. Other steps were not key to project completion but were mandatory in the ICODA framework and principles; for example, making data findable, accessible, interoperable, and reusable (FAIR) via publication of metadata and access request routes, so other researchers are able to use it in the future. The key process steps tracked are outlined below.

	Main aims	Types of data and data source	Country	Key challenges	Key solutions
(Continued from previous page)					
Data descriptor, reference coding, and characterisation of the systemic complications of critical care patients included in the ISARIC COVID-19 dataset (DP-ISARIC)	To identify and develop tools and strategies to enhance the extraction and comprehensive utilisation of data within the ISARIC COVID-19 dataset, with the specific aim of identifying risk factors associated with systemic complications and assessing their effect on clinical outcomes.	Clinical data from patients hospitalised with COVID-19 globally shared as a part of the ISARIC Clinical Characterisation Group collaboration. ISARIC has assembled the world's largest global database on COVID-19 clinical data with detailed individual patient data on 657 312 hospitalised individuals from 1297 institutions across 45 countries. This database includes data from more than 705 000 patients and 1500 centres worldwide.	>60 countries	Standardisation and mapping of the data encoded in open text fields of the dataset; due to the large dataset size, this was divided into 16 individual large tables that were hard to manipulate.	Team developed a computational code that identified the most frequently used texts and mapped them to standard codes with relevant clinical meaning. They then reviewed these codes manually to ensure they were accurate, capturing the original text reported in the dataset; the team designed a computational strategy that allowed them to identify each subject in each table to extract the individual data registered in each table. Then, they created a smaller dataset that allowed them to perform comprehensive statistical analyses. The code used to generate these datasets was registered in an open platform that other researchers could use to conduct their extractions and analyses once they had access to the dataset.
Using routine data to understand adverse pregnancy and neonatal outcomes associated with the COVID-19 pandemic in Kampala, Uganda (DP-IROC)	Incidence and risk factors for COVID-19 among pregnant and lactating women and their infants.	Quantitative data from the electronic medical records system.	Uganda	Team had a lot of unstructured data collected using free text fields in the variables that were required for the analysis dataset; working in the trusted research environment was challenging, due to limitations on data table editor and bandwidth.	The team, with support from MMS Holdings, completed additional data curation processes to produce an analysis-ready dataset; team held interactive sessions with the ICODA Workbench provider to inform expansion of workbench space and utility by using a Windows Virtual Machine.

ICODA=International COVID-19 Data Alliance. ISARIC=International Severe Acute Respiratory and Emerging Infection Consortium.PRIEST=Pandemic Respiratory Infection Emergency System Triage.

Table: **Grand Challenges ICODA driver projects**

Signing a data processing agreement

These agreements were between the data custodians (institutions or organisations sharing the data) and the provider of the ICODA Workbench, Aridhia Informatics (Glasgow, UK). The agreements covered the legal obligations and data governance responsibilities for adding data to the ICODA Workbench, and all ICODA driver projects were provided with a data processing agreement by Aridhia Informatics, along with support from the core programme team to aid completion.

Gaining researcher accreditation

This step involved the project teams completing the accreditation process for all researchers requiring access to the research environment, in line with the safe people principle (panel 1). Researchers' suitability to be given access to the ICODA Workbench was assessed against the following criteria: all requested information was provided; researcher is affiliated to a legitimate organisation conducting research and a bona fide researcher; and researcher has professional qualifications and experience to work with health data. This suitability was assessed by ICODA's research manager.

Preparing or curating the data

This step encompassed activities related to loading and preparing the data for analysis. Two of the projects received support from a commercial company specialising in data curation to expedite the process and address the challenges of harmonising datasets from multiple countries. This process included missing data and non-standardised records, especially those in free text format.

Sharing the metadata

This step involved the research teams publishing the details of the data (metadata) used for their research project. The FAIR principles[16,17] underpinned the ICODA initiative, and projects were able to access support to list metadata descriptions of all the data used—including associated data access requirements—on an accessible metadata catalogue, with digital object identifiers that could then be cited in associated publications.[18] The core ICODA team assisted project members to perform this step, which often was left until the latter part of the projects.

Initial analysis

This step included the time to complete the first iteration of analysis of the research data. Several analysis tools

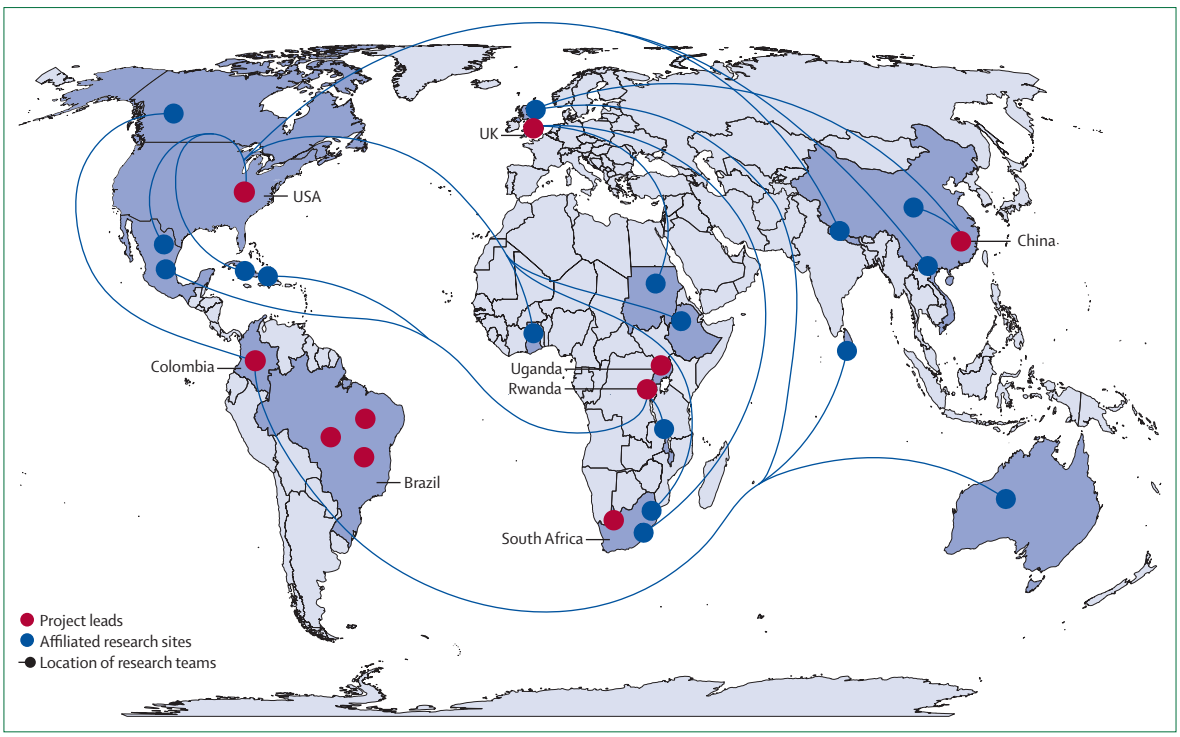

Figure 1: **Global scope of ten Grand Challenges ICODA driver projects**
The countries shaded dark blue indicate the geographical scope of the ICODA initiative.

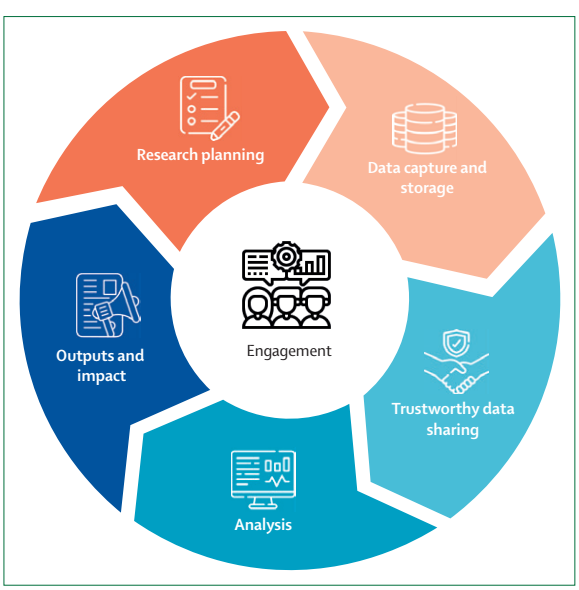

Figure 2: **Health data research project lifecycle**

were provided for use in the ICODA Workbench; in addition, bespoke tools were provisioned where required.

Sharing outputs

Each project addressed major research questions relating to the COVID-19 pandemic and aimed to generate rapid insights within 12 months. This step included activities related to initial publications for each research project as well as other outputs such as dashboards, community engagement materials, videos, and documentaries.

Engagement with and involvement of local communities, health practitioners, and policy makers

From the initial project design and through each of these process steps, the driver project research teams engaged with a range of stakeholders, including community members, health practitioners, and policy makers, to ensure research questions were relevant, the value of the research was understood, and outputs benefited the health of communities. Several projects invested substantial time and resources in engaging with local communities including: participation in a vaccine programme alongside a local non-governmental organisation; working with community leaders, community groups, and young influencers to communicate the purpose and value of the research; and engagement with local media.

A quantitative and qualitative analysis of the common process steps and the length of time required for each step was then carried out for the ten Grand Challenges ICODA driver projects, to identify common barriers and bottlenecks as well as possible solutions. The length of time taken for each process step was self-reported, in response to requests for this information. This information was verified through cross-referencing relevant correspondence and system-generated timestamps

where available. Although this information offered insights into the effort required at each process step, there are limitations to the data obtained, which can introduce bias and variability. These limitations include differences in the types of data, the analytical methods and the research approach being used by each driver project, as well as different interpretations of definitions of each process step, and the fact that the information was collected retrospectively. As can be observed, the time for each process step varied considerably between each project team with process step duration being shown in figure 3 (appendix).

Qualitative analysis was supported through the quarterly monitoring reports and final report provided by each driver project team, which specifically requested feedback on any lessons learned. Online meetings during the funding period and a mid-project convening of the cohort (focused on community and stakeholder engagement) also provided opportunities for feedback from the research teams.

Results: challenges and solutions for international health data access and sharing

Challenges within the health data research cycle

Analysis of the time taken for the ten Grand Challenges ICODA driver projects to undertake each process step in the health data research cycle is set out in figure 3. This analysis highlighted some common challenges and barriers, and suggested possible solutions in relation to putting data processing agreements in place, researcher accreditation, data curation and preparation, ensuring metadata was prepared and published appropriately following FAIR principles, and undertaking analysis of secondary data. These and other challenges that research teams using data science approaches to address health challenges might have will be explored further.

Identifying data sources

A secondary data use project usually starts with identifying existing data to answer the research question. This process often involves a literature search and scanning other sources where available datasets are listed, such as repositories and data catalogues. Despite improvements in digital cataloguing, this stage is still challenging due to a lack of interconnectedness in the data sharing ecosystem and inconsistent use of metadata standards around the world. The Grand Challenges ICODA driver projects required data to have been already identified at the application stage, so data source identification time was not included within the project timelines.

Data processing agreements

A median of 1·5 weeks of project time was taken up with finalising data sharing agreements. This figure was calculated using data reported by seven of the ten project teams. Research contracts are recognised to be time-consuming, as multiple departments get drawn in to reviewing and negotiating terms. Contracts relating to data reuse are particularly complex due to additional regulatory requirements, such as data protection obligations. One solution to reduce the time taken for this stage is to standardise data sharing agreements as much as possible, as was done for the ICODA programme. Standardisation would ideally use an internationally agreed set of minimum criteria that could be used by data custodians and data users around the world.

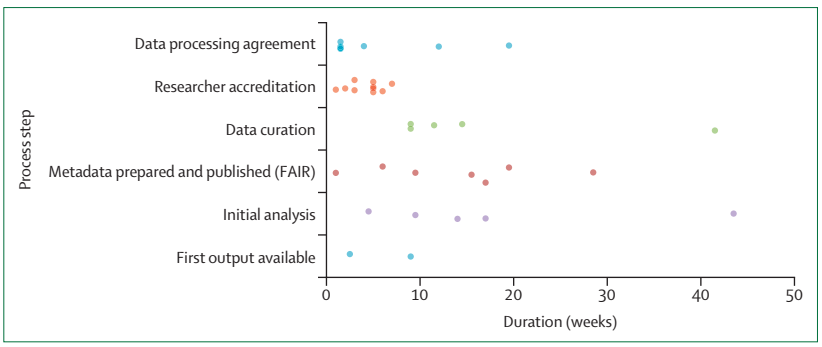

Figure 3: **Strip plot showing the time taken in weeks across the ten driver projects to complete each process step**
FAIR=findable, accessible, interoperable, and reusable.

See **Online** for appendix

Contract negotiation difficulties stemmed from both the complexity of agreements and models to enable data sharing. Each institution was responsible for uploading their data into the TRE, and driver project teams were encouraged to make their data FAIR. Aridhia Informatics, the TRE provider, provided their standard data processing agreement, which was used for projects to access the TRE. Projects were responsible for their own data uploaded to the TRE.

Each institution was recognised as a data controller of the data they contributed, enabling the use of their data within the TRE; however, this limited the ability to share data between projects.

Researcher accreditation

Researcher accreditation was a quick process stage for researchers to complete. It was automated via the ICODA Gateway; data reported by all ten of the project teams indicated a median of 5 weeks for each project. The process also allowed for additional researchers to be brought into project teams quickly where required.

Data curation and preparation

Most datasets need to be curated before analysis can begin, which is typically a time-consuming step. The median time spent curating data was 11·5 weeks, but there were considerable outliers. The median time taken was calculated using data reported by five of the ten driver project teams.

A natural language processing approach was applied in the case of the DP-PRIEST study to address the challenges of converting free text to clinical codes. The

Panel 2: **Challenges and solutions from the ten Grand Challenges International COVID-19 Data Alliance (ICODA) driver projects**

Identifying available data

Challenges

It is difficult to identify available data globally, with limited connections between repositories and data catalogues, and no internationally agreed minimum metadata standards.

Published articles regularly have limited or no data sharing statements included.

Solutions

Researchers should always list metadata on a publicly accessible repository and include clear data sharing statements with as much detail as possible about access request requirements. By the end of their grants, nine of ten Grand Challenges ICODA projects had listed their metadata on the ICODA Gateway in a standard format to enable other researchers to understand the contents of the datasets and request data access.[6]

Each metadata description includes a digital object identifier which can be cited in relevant articles to ensure appropriate attribution.

All articles published by the teams should include data availability information, along with a digital object identifier.

Data sharing agreements

Challenges

Getting contracts signed is a protracted process and terms are often debated.

Solutions

Encourage use of standard templates and share these with relevant contract departments as early as possible.

Simplified language is key, with agreements needing to be approachable and clear while serving the required purpose.

When carrying out research in an international context, it is important to identify and share best practice in data governance and privacy and share guidelines with teams, especially when taking an open science approach.

Accrediting researchers

Challenges

Protecting the privacy of participants and preventing misuse of patient data are critical when using health data for research and it is key that trustworthy and transparent systems are built into all activities undertaken by researchers.

Solutions

ICODA made use of the Five Safes framework first developed by the UK's Office for National Statistics and adopted by Health Data Research (HDR) UK.[10]

ICODA researchers were assessed against Safe People criteria and this is an approach we would recommend for adoption by others as a proportionate review process for data access to ensure data is only accessed by trained and accredited researchers who are trusted to use it appropriately.[11]

Data curation and challenges preparing data

Challenges

Lack of data standardisation, even within single health datasets, means that combining data from multiple sources can be extremely difficult and time-consuming to convert and transform.

Solutions

For ICODA projects, the initial aim of working to a standardised data dictionary across the spectrum of projects was realised to be too ambitious and was therefore refocused on preparing dictionaries for individual projects, rather than across the initiative.

Data curation tasks were aided by partnerships with expert teams, initially with Aridhia Informatics and Cytel for the first ICODA driver project, and then with MMS Holdings for the ten Grand Challenges ICODA driver projects. These projects had mixed data curation needs and being able to ask questions of the statistical expert group and expert curation partners helped formulate the project teams' thinking.

Having professional partnerships available to all projects earlier would have facilitated progress of research and been a worthwhile investment.

Sharing tools for transforming data, particularly in open-source format, can help the secondary data use community. ICODA supported shared tools within the trusted research environment, and shared tips and approaches for data curation through webinars and workshops.

Realistic costs for data curation and transformation need to be included upfront and planned for in grant budgets, along with sufficient time for this activity built into the project plan.

Making metadata available

Solutions

HDR UK's existing Innovation Gateway was repurposed to support the ICODA initiative, creating a new Gateway instance with associated branding.[6] The ICODA Gateway enabled metadata to be made visible, while also implementing key processes, such as researcher accreditation and data access requests. Researchers entered metadata into the metadata catalogue provided within the trusted research environment which was then federated to the ICODA Gateway, making metadata publicly visible.

This reuse of HDR UK's existing software assets accelerated time to delivery through customising, rather than coding from scratch, which was cost-saving and time-saving. Piloting metadata federation with the ICODA trusted research environment partner resulted in functionality that has subsequently been implemented more broadly within HDR's Innovation Gateway and is being rolled out across the HDR UK ecosystem.

(Continues on next page)

(Panel 2 continued from previous page)

Ensure training is available for teams to upload metadata to catalogues and project time is planned in for curating the metadata or data dictionary itself. Many teams left this step until the later stages of their projects and data could have potentially been reused within the cohort of projects and more broadly, had this been done earlier.

Enabling data analysis

Challenges

Still difficult for sensitive data or multiple large datasets or data spanning country borders.

Solutions

Researchers in the ICODA programme were provided with access to the COVID-19 Workbench, a cloud-hosted trusted research environment delivered by our partner, Aridhia Informatics. This provided researchers with a secure space to perform collaborative research, with controlled access, scalable compute power, tooling, assistance with data provision, hosting, and support.

Use of a turnkey, cloud-hosted trusted research environment jumpstarted the initiative and is a route to be considered for future projects. It was key in enabling collaboration, saved the creation of multiple instances hosting the same data (important when working on large datasets across geographical locations), provided data custodians with a high level of confidence in security of data, and proved accessible from low bandwidth settings, providing high end computational power to research teams where required.

Effective community and stakeholder engagement

Challenges

Teams often had limited experience of community and patient engagement in shaping research projects, and pandemic restrictions made engagement even more challenging.

The limited time scale (12 months) of the projects also proved challenging.

Solutions

The ICODA team ran workshops and question and answer sessions on stakeholder and community involvement and engagement for long-listed research teams to further develop their detailed plans and, working with expert groups, convened a community, public, and patient review panel for the Grand Challenges ICODA open funding call.

A halfway convening workshop was organised for all ten Grand Challenges ICODA driver project teams, which focused on their community, public and patient involvement and engagement plans, progress, and challenges, enabling knowledge sharing across the cohort. With guidance from ICODA's ethics advisory council, an ethics and governance framework was developed for use by all project teams.[18]

Engaging stakeholders in setting research questions early and the communication of results proved valuable. This included establishing a dissemination plan with local stakeholders based on active listening with rapid communication, as used by DP-EFFECT.[19] Social media activities and partnership with local influencers, including primary care workers and community leaders, were cited as important.

Creating different communication approaches for different audiences, including videos and webinars, was valuable, and the DP-IDS-COVID-19 team highlighted the importance of avoiding use of technical terms.[20] The best tool to be used depends on the objective of engagement with participants.

Budget for engagement activities and ensure that there are members in the team with the appropriate skills to support this work.

Partner with local civil society organisations and the public and private sectors to ensure local needs are embedded in the research.

Engaging policy makers in using research outputs

Solutions

Plan for this element from project initiation and engage early to ensure outputs are in a useful and consumable format for policy makers. Several teams produced dashboards for use by policy makers or communities, including DP-IROC[21] and DP-ACCORD.[22]

Tailor communications and create and circulate regular policy briefings.

Set up and hold regular briefing meetings to highlight the importance and relevance of the research to policy development, change, or implementation.

DP-CHAIN project also used natural language processing to curate their data and the natural language processing algorithm was published on iScience.[19] The DP-REHCORD team highlighted the importance of setting up a clear and standardised codebook when preparing datasets for analysis; for example, it was suggested that codebooks should include variable names, labels, and any skip patterns.[20] This method was found to be particularly important if working across multiple datasets, in large teams, or across multiple countries.

Another recommendation was to adopt a version control system for statistical code such as GitHub, to allow for collaborative and simultaneous work across large teams to ensure that code is not lost or overwritten. Making this code publicly available through such platforms also improves reproducibility and transparency. For studies in countries with disaggregated data, teams were advised to implement simple and standardised procedures for data cleansing. DP-CHAIN's codebook and data are published.[21] The DP-ISARIC team[22] applied a uniform data model to standardise the structures and ontologies in the ISARIC dataset to a harmonised format. All data were standardised to the Clinical Data Interchange Standards

Consortium Study Data Tabulation Model to facilitate pooled analyses.

Making metadata available

Median time spent making metadata available was 15·5 weeks, from data reported by seven of the ten project teams. This duration, although lengthy, reflects not only the time taken to upload the metadata to an online catalogue, but also the work involved in preparing the metadata itself. Organisation of variables, their names, descriptions, and valid ranges ended up being a substantial piece of work for many teams.

For more on **ICODA news and events** see https://icoda-research.org/news-and-events/

Initial analysis

Driver project teams experienced first-hand the challenges of using diverse health data collected from multiple sources and trying to combine and analyse them. One of the main difficulties encountered in combining data was the lack of data standardisation, quality, and structure, which reduced interoperability. Databases often had missing information and unstructured entries, such as free text information, which made analysis difficult.

The use of additional statistical and data science expertise to help execute data analysis was noted by a number of teams; the DP-RASUP team highlighted the importance of not underestimating the time needed for processing data, especially large and complex datasets.[23] Median time taken for the initial analysis stage was 14 weeks, using data reported by five of the ten project teams.

An example of how these challenges were overcome comes from the DP-CHAIN project, which reconstructed transmission pairs using epidemiological survey data published by governments around the world.[19] Different countries adopt different standards when tracing close contacts and report their findings in different ways. DP-CHAIN standardised global contact tracing data by categorising them into two types: individual contacts and contact clusters. The team further developed an algorithm to infer transmission pairs from contact networks. Epidemiological characteristics, such as the basic reproduction number (R_0) and dispersion, could then be calculated from the transmission networks and compared across countries. R_0 is the average number of secondary infections generated by a single infectious individual in a population where all members are susceptible to the infection, with higher values indicating an increase in infection in the population.[24] The dispersion parameter (k) quantifies the variation in the number of secondary infections caused by infected individuals, with lower values of k indicating a greater likelihood of superspreading events.[25] Taking an innovative approach, the DP-RASUP team documented their methods for data processing, data analysis, data visualisation, mathematical modelling, and statistical modelling in a series of YouTube videos, which are freely available for other researchers to view and learn from.[23]

For the **digital hub** see https://globalhealthdatascience.tghn.org/

For the **policies and processes** see https://icoda-research.org/research/publications/#genericgovernanceprocessesforreus

ICODA established several support mechanisms to enable teams to overcome analysis challenges. For example, a team of statistical experts were identified who provided advice and input to the research teams where needed; they commented on statistical analysis plans and developing and sharing analysis tools with them as the projects progressed. Furthermore, tools, code, and curation advice were shared between ICODA project team members within the analysis environment and more broadly through data science webinars and workshops, many of which are available to watch online on the ICODA website. A full summary of the challenges, solutions, and lessons learned by the ten Grand Challenges ICODA driver projects is presented in panel 2.

Outputs and impact

Across the ten ICODA driver projects, a wide range of project outputs were planned and delivered including: results manuscripts, code, methods papers, dashboards, community and stakeholder engagement materials and tools, videos, and documentaries. Despite the challenges outlined in this Health Policy paper, all teams were successful in delivering rapid insights and outputs, with some projects beginning to share findings as early as 6–7 months into the project. Most teams published their findings between 9 months and 15 months after the projects started, with all publications, code, and metadata being made open and accessible to other researchers.

Innovative outputs included those from the DP-PRIEST team who produced a validated triage tool for use by clinicians in low-income and middle-income settings for assessing whether patients should be admitted with COVID-19 to intensive care units.[26] Other teams have documented their methods, community engagement experiences and approaches, as well as tools for use by policy makers and health service leads. The DP-IDS-COVID19 team developed an index to measure social inequalities during the pandemic in Brazil, which was used by a council of representatives of state health managers to identify people vulnerable to COVID-19 and guide the planning of interventions. These outputs are shared more widely on the Global Health Data Science digital hub, to which the ICODA teams have contributed.

The ICODA initiative has sought to maximise research impact through making a range of outputs openly available, including transformational code, dataset metadata (on HDR's Innovation Gateway), community engagement materials, and governance policies and processes. These policies and processes are available on the ICODA website and have been genericised for wider reuse.

A less tangible but equally important outcome of the initiative has been that a global health data science community of practice has been established and continues to be active. The ten Grand Challenges ICODA

projects are now firmly embedded in the wider Grand Challenges community and their research teams continue to engage as part of this wider data science community of practice through the Global Health Data Science digital hub and HDR UK's Global programme.

Further perspectives and wider lessons learned

Engagement with local communities, health practitioners, and policy makers

Community and stakeholder engagement at all stages of the data reuse project cycle underpins relevant and quality research. It was a key element of the ten Grand Challenges ICODA driver projects, having been built into the design of the global funding call, and the subsequent support for the driver projects to deliver benefits to patients and better health outcomes for all. Levels and types of engagement varied across projects and involved direct engagement with local communities to raise awareness of the research and the potential positive effect on community health priorities through using data science-enabled insights to influence health policy and practice. Direct engagement with policy makers and practitioners also took place to shape research projects and ensure insights and outputs were taken up.

Teams had several challenges associated with engagement activities, but also found innovative solutions. The DP-EFFECT team used mobile app-based technologies to provide rapid access to free COVID-19 testing during the pandemic, showing the high potential of these e-health technologies in improving the access of vulnerable populations to health-care services.[27] Since completing the Grand Challenges ICODA research, the DP-EFFECT team has been developing a community engagement toolkit based on their stakeholder engagement experiences, from which other researchers can benefit. The DP-PRIEST team set up a public patient involvement and engagement group in the Western Cape, South Africa, including eight community members affected by COVID-19 (infected themselves or an immediate family member was infected or hospitalised). The group were kept informed about the study, and then given the opportunity to provide feedback. The feedback provided was particularly useful for gaining a public perspective on the use of anonymised routinely collected health-care data. Engagement was focused on raising awareness of the research and the findings and their potential to inform health policy.

The DP-IDS-COVID19 team invested substantial time engaging with a range of stakeholders and published their engagement experience in a journal article.[28] They describe how community members and policy makers made contributions to existing research through informing a new layer of information in the interactive social disparities index developed by the team, as well as improvements to the interactive index panel itself. Eight representatives of community groups and 29 policy makers participated in engagement activities during the project, more than 500 people engaged in open webinars about the project, and over 140 news items about this study were published in national and international media.

Research impact and uptake rely on the involvement of all key stakeholders through the research lifecycle. Challenges faced by the ICODA driver project teams included the fact that ministries of health were occupied with managing the COVID-19 burden and fast-moving policy development, making engagement difficult. Despite these challenges, the DP-IDS-COVID19 and DP-IROC teams developed dashboards that were accessible by ministries to improve targeting of COVID-19 interventions.[28,29] The DP-PRIEST research team engaged with a range of clinical academics from the South African and Sudanese health-care settings as well as local networks to discuss findings and the acceptability of triage tools developed.

Conducting research during a pandemic

Conducting any research in a pandemic is challenging and secondary health data use projects were no exception. COVID-19 restrictions limited in-person meetings and networking, something particularly important in building trust and productive teamwork across multiple locations, and the involvement of community and other stakeholders. Maintaining up-to-date data in a fast-changing context was difficult, as was implementing dynamic data collection with fast clinical and epidemiological variations.

A further challenge in conducting research for studies led by health service staff, was dedicating person-time to the analyses when there were resurgences of COVID-19 and competing service priorities. To address this, the DP-ACCORD team were able to automate the updating of analysis datasets, so that repeating analyses as the epidemic progressed became progressively easier.[30]

Conclusion

The ICODA programme itself has now completed, and achievements and outputs from the initiative and its driver projects have been significant and varied. To date, there have been 38 publications from its full cohort of 12 driver projects, with more papers submitted. All achievements and outputs are available to other researchers through open access publication and are indexed on the ICODA website.[15]

This Health Policy paper summarises the common challenges and potential solutions to accelerate the health data research cycle, based on the research process of a diverse range of ten research driver projects under the ICODA programme, and highlights where data infrastructure and governance could be improved to better enable secondary data use studies. The approaches, tools, and outputs from these driver projects are openly available and can be used by researchers for similar

studies using secondary data to address a wider range of health challenges, and the implementation of some of these solutions early on could help to accelerate the generation of insights and outputs.

Contributors

SB led on writing of the manuscript and GSH conceived and wrote the first draft. The principal investigators of the Grand Challenges ICODA driver projects (CA, FAB, MYI, KlD, XFL, JCM, JFO, and LFR) and team members (LH, JO, MB, and AF) contributed to the writing, reviewing, and editing of the manuscript. From the ICODA core team, MR led on data analysis from the studies and ADM, NP, MR, EM, and AW contributed to writing, reviewing, and editing the manuscript. EF contributed to the review, editing, and writing of the manuscript.

Declaration of interests

All coauthors (except EF) received funding from the Bill & Melinda Gates Foundation and Minderoo Foundation for the ICODA initiative. KLD received grant funding for the main project from the European and Developing Countries Clinical Trials Partnership (EDCTP) 2 programme supported by the European Union (RIA2020EF-2926 periCOVID Africa). LFR received grants from Pfizer and Merck Sharp & Dohme; consulting fees from Merck Sharp & Dohme, Pfizer, and GSK; payment for lectures and presentations from Merck Sharp & Dohme and GSK; and payment for expert testimony from Merck Sharp & Dohme, Pfizer, and GSK. MYI received grant funding from the Ministry of Health (decentralised executive term, process number 25000200517/2019-46), National Institute for Health and Care Research (NIHR), and the Wellcome Trust; and payment to participate in the International Population Data Linkage Network Conference from the NIHR Global Health Research Program Award. MB received grants from the London School of Economics, The Rockefeller Foundation, Global Challenges Research Fund Global Multimorbidity, and Google; has received consultancy fees from the Singapore Institute of Management; and has a patent from the Institute of Electrical and Electronics Engineers Standards Association. AF has received grants from the Ministry of Health, NIHR, and the Wellcome Trust. GSH received funding from the Gates Foundation contract for contract work unrelated to the contents of the manuscript; consultancy fees from Vivli clinical data sharing platform and ClinicalStudyDataRequest.com clinical data sharing platform; and fees from the National Institute on Ageing to present at their data sharing workshop. All other authors declare no competing interests.

Acknowledgments

We acknowledge and give special thanks for Joseph Ouma's contribution to the writing of this manuscript, the delivery of a webinar, and to the Grand Challenges ICODA initiative. Very sadly, Joseph passed away on Sept 2, 2023, a profound loss to his family and colleagues. We dedicate this paper to the life and work of Joseph Ouma and extend our deepest condolences to Joseph's family and loved ones. We also thank and acknowledge the significant contributions of the patients and public whose data informed the research projects; all members of the ten Grand Challenges ICODA research driver project teams who have participated in this initiative; and the data contributors who willingly shared their datasets for reuse and for the benefit of patients and the wider public. We also thank Steve Kern for his extensive guidance and commitment to the initiative; Kim Carter for his engagement and unwavering support; Amel Ghouila for her advice and inputs to the design and delivery of the Grand Challenges ICODA open funding call; Carl Marincowitz for his engagement in the ICODA initiative and participation in a webinar on Trusted Research Environments; Agklinta Kiosia for her review of the paper; and our technical delivery partners, Aridhia Informatics, Certara, Cytel, MMS Holdings, PA Consulting, Preva Group, and SAIL Databank for their contributions. The ICODA initiative was shaped and supported by three key governance bodies: the Scientific and Strategic Advisory Council, the Ethics Advisory Council, and the Executive Leadership Team. We thank everyone involved for so generously giving their time and expertise to guide the initiative to success. This work was financially supported by ICODA, an initiative funded by the Gates Foundation (INV-017293), the Minderoo Foundation, supported by Microsoft's AI for Good Research Laboratory, and convened by Health Data Research UK. Aridhia Informatics was funded by the Gates Foundation (INV-021793) to provision the Workbench for the driver projects. The views and opinions of the authors expressed herein do not necessarily reflect those of EDCTP. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

References

1 Science Academies of the Group of Seven (G7). Data for international health emergencies: governance, operations and skills. March 31, 2021. https://royalsociety.org/-/media/about-us/international/g-science-statements/G7-data-for-international-health-emergencies-31-03-2021.pdf (accessed July 25, 2023).

2 Moorthy V, Henao Restrepo AM, Preziosi MP, Swaminathan S. Data sharing for novel coronavirus (COVID-19). *Bull World Health Organ* 2020; **98:** 150.

3 Wellcome. Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. Jan 31, 2020. https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak (accessed July 25, 2023).

4 Baynes G, Hahnel M. Research practices in the wake of COVID-19: busting open the myths around open data. https://www.springernature.com/gp/advancing-discovery/springboard/blog/blogposts-open-research/research-practices-in-the-wake-of-covid/18256280 (accessed July 25, 2023).

5 Näher AF, Vorisek CN, Klopfenstein SAI, et al. Secondary data for global health digitalisation. *Lancet Digit Health* 2023; **5:** e93–101.

6 Benchimol EI, Smeeth L, Guttmann A, et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015; **12:** e1001885.

7 Bernier A, Raven-Adams M, Zaccagnini D, Knoppers BM. Recording the ethical provenance of data and automating data stewardship. *Big Data Soc* 2023; **10:** 20539517231163174.

8 Sigfrid L, Maskell K, Bannister PG, et al. Addressing challenges for clinical research responses to emerging epidemics and pandemics: a scoping review. *BMC Med* 2020; **18:** 190.

9 Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. 2016. https://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf (accessed July 25, 2023).

10 International COVID-19 Data Alliance. Research review process. October, 2022. https://doi.org/10.57775/mhbw-2v62 (accessed Aug 7, 2023).

11 International COVID-19 Data Alliance. Generic output research review process for research projects. October, 2022. https://doi.org/10.57775/6jph-dt22 (accessed Aug 7, 2023).

12 Health Data Research (UK) Innovation Gateway. International COVID-19 data alliance datasets. June, 2023. https://web.www.healthdatagateway.org/collection/29513503379534136 (accessed Aug 7, 2023).

13 Dron L, Kalatharan V, Gupta A, et al. Data capture and sharing in the COVID-19 pandemic: a cause for concern. *Lancet Digit Health* 2022; **4:** e748–56.

14 Calvert C, Brockway MM, Zoega H, et al. Changes in preterm birth and stillbirth during COVID-19 lockdowns in 26 countries. *Nat Hum Behav* 2023; **7:** 529–44.

15 International COVID-19 Data Alliance. Publications and outputs. 2022. https://icoda-research.org/research/publications (accessed Jan 9, 2024).

16 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3:** 160018.

17 Maxwell L, Shreedhar P, Dauga D, et al. FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries. *Lancet Digit Health* 2023; **5:** e712–36.

18 International COVID-19 Data Alliance. ICODA metadata instructions. 2024. https://icoda-research.org/wp-content/uploads/2024/01/ICODA-metadata-instructions.pdf (accessed Jan 26, 2024).

19 Wang Z, Liu XF, Du Z, et al. Epidemiologic information discovery from open-access COVID-19 case reports via pretrained language model. *iScience* 2022; **25:** 105079.

20 Arsenault C, Gage A, Kim MK, et al. COVID-19 and resilience of healthcare systems in ten countries. *Nat Med* 2022; **28:** 1314–24.

21 Liu XF, Xu XK, Wu Y. Mobility, exposure, and epidemiological timelines of COVID-19 infections in China outside Hubei province. *Sci Data* 2021; **8:** 54.

22 Garcia-Gallo E, Merson L, Kennon K, et al. ISARIC-COVID-19 dataset: a prospective, standardized, global dataset of patients hospitalized with COVID-19. *Sci Data* 2022; **9:** 454.

23 Pereira FAC, Filho FMHS, de Azevedo AR, et al. Profile of COVID-19 in Brazil-risk factors and socioeconomic vulnerability associated with disease outcome: retrospective analysis of population-based registers. *BMJ Glob Health* 2022; **7:** e009489.

24 Boonpatcharanon S, Heffernan JM, Jankowski H. Estimating the basic reproduction number at the beginning of an outbreak. *PLoS One* 2022; **17:** e0269306.

25 Wegehaupt O, Endo A, Vassall A. Superspreading, overdispersion and their implications in the SARS-CoV-2 (COVID-19) pandemic: a systematic review and meta-analysis of the literature. *BMC Public Health* 2023; **23:** 1003.

26 Marincowitz C, Sbaffi L, Hodkinson P, et al. Prognostic accuracy of triage tools for adults with suspected COVID-19 in a middle-income setting. *Emerg Med J* 2022; **39:** A976–77.

27 Ranzani OT, Silva AAB, Peres IT, et al. Vaccine effectiveness of ChAdOx1 nCoV-19 against COVID-19 in a socially vulnerable community in Rio de Janeiro, Brazil: a test-negative design study. *Clin Microbiol Infect* 2022; **28:** 736.e1–4.

28 dos Anjos Fonseca A, Pimenta DM, de Almeida MRS, Lima RT, Barreto ML, Ichihara MYT. Public involvement & engagement in health inequalities research on COVID-19 pandemic: a case study of CIDACS/FIOCRUZ BAHIA. *Int J Popul Data Sci* 2023; **5:** 2133.

29 International COVID-19 Data Alliance. Incidence and risk factors for COVID-19 amongst pregnant and lactating women and their infants in Uganda (DP-IROC). 2023. https://icoda-research.org/project/dp-iroc/ (accessed Aug 7, 2023).

30 Davies MA, Morden E, Rousseau P, et al. Outcomes of laboratory-confirmed SARS-CoV-2 infection during resurgence driven by omicron lineages BA.4 and BA.5 compared with previous waves in the Western Cape Province, South Africa. *Int J Infect Dis* 2023; **127:** 63–68.