Translating Data in a Pandemic 2

Data capture and sharing in the COVID-19 pandemic: a cause for concern

Louis Dron, Vinusha Kalatharan, Alind Gupta, Jonas Haggstrom, Nevine Zariffa, Andrew D Morris, Paul Arora, Jay Park

Routine health care and research have been profoundly influenced by digital-health technologies. These technologies range from primary data collection in electronic health records (EHRs) and administrative claims to web-based artificial-intelligence-driven analyses. There has been increased use of such health technologies during the COVID-19 pandemic, driven in part by the availability of these data. In some cases, this has resulted in profound and potentially long-lasting positive effects on medical research and routine health-care delivery. In other cases, high profile shortcomings have been evident, potentially attenuating the effect of—or representing a decreased appetite for— digital-health transformation. In this Series paper, we provide an overview of how facets of health technologies in routinely collected medical data (including EHRs and digital data sharing) have been used for COVID-19 research and tracking, and how these technologies might influence future pandemics and health-care research. We explore the strengths and weaknesses of digital-health research during the COVID-19 pandemic and discuss how learnings from COVID-19 might translate into new approaches in a post-pandemic era.

Introduction

The need for timely and high-quality evidence to inform decision making for policy makers during the COVID-19 pandemic has catalysed the development of innovative digital-health technologies, which have had variable effects and benefits.[1] Timely communication of crucial information requires the acquisition, access, and analysis of large volumes of data. For example, analysis of electronic health records (EHRs) that capture real-time patient records of routine clinical care have resulted in better disease surveillance and produced evidence to inform public-health decisions. EHRs can be supplemented with clinical trial data or administrative claims to monitor the effect of novel treatment strategies and vaccines for COVID-19 in almost real time.[2,3] Although the use of EHRs, observational datasets, public epidemiological data, and clinical trial data to inform public-health policies sounds promising, there has been mixed success during the COVID-19 pandemic. Challenges with data capture (including non-standardised data collection, heterogeneity in data terminologies, and the enduring structural, cultural, and political barriers to data sharing) have led to data gaps and attenuated the potential research effect of these data. To overcome the limitations of individual datasets, and the challenges of larger integrated datasets, several research groups have undertaken data mining for potentially valuable insights into disease processes.[4,5] These challenges in data access and dissemination are not unique to COVID-19 research, although the scale of data, the necessity of timely generation and dissemination, and the concurrent global research focus represent an entirely unique context for these topics.

There have been numerous reviews on the impact of the COVID-19 pandemic on topics related to digital health, including on publication practices,[6] public perception of science,[7] and research funding overall.[8] In this Series paper, we specifically consider the effect of the pandemic on primary data collection through digital-health technologies, subsequent data sharing for secondary use, and the associated challenges for collaborative international scientific endeavours uncovered by the COVID-19 pandemic.

Routinely collected health data: real-world evidence

Both before and during the COVID-19 pandemic, research using routinely collected health data has had a substantial focus on the concept of real-world evidence. The US Food and Drug Administration defines this concept as the translation of routinely collected medical data via EHRs, claims and billing activities, disease registries, and other mobile-health disease-monitoring data into actionable and meaningful evidence sets.[9] A distinction is made between real-world evidence and real-world data, which pertains to the primary data capture associated with medical care. The use of real-world evidence in research programmes has been varied, covering disparate topics (such as core epidemiological estimates of disease burden), being integrated into regulatory control datasets for comparative efficacy, and used to identify routine study procedures for pragmatic clinical trials.[10,11] The uptake of these data has been driven by timely access and standardisation,[12] methodological developments,[13] and regulatory guidance.[14]

Existing momentum and expertise enabled rapid engagement and dissemination of real-world evidence for COVID-19 research, as illustrated by the sheer volume of output. By July 7, 2022, there had been 13 395 publications of real-world evidence on COVID-19. Restricting this to the first 600 days of the pandemic (September 24, 2021), there were 5951 peer-reviewed

Real World & Advanced Analytics, Cytel Health, Vancouver, BC, Canada (L Dron MSc, J Haggstrom PhD, P Arora PhD); **Department of Epidemiology and Biostatistics, Western University, London, ON, Canada** (V Kalatharan PhD); **Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada** (A Gupta PhD, P Arora); **The International COVID-19 Data Alliance (ICODA), Health Data Research UK, London, UK** (J Haggstrom, N Zariffa MMath, A D Morris MD); **NMD Group, LLC, Bala Cynwyd, PA, USA** (N Zariffa); **Department of Experimental Medicine, Department of Medicine, University of British Columbia, Vancouver, BC, Canada** (J Park PhD); **Department of Health Research Methods, Evidence and Impact, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada** (J Park)

Correspondence to:
Mr Louis Dron, Real World & Advanced Analytics, Cytel Health, Vancouver, BC V5Z 4J7, Canada
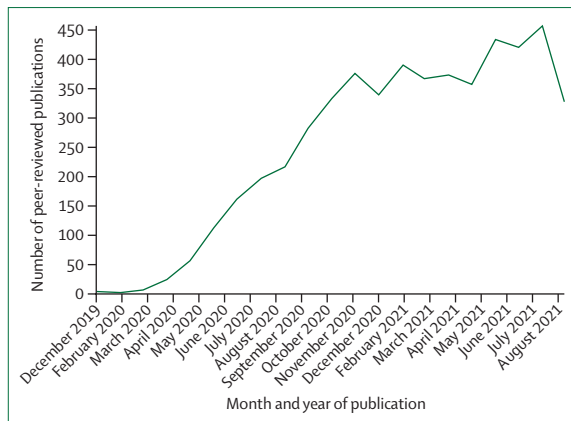**louis.dron@cytel.com**

Figure 1: Number of peer-reviewed publications related to real-world data and COVID-19 published during the first year and a half of the COVID-19 pandemic

For the **International COVID-19 Data Alliance see** https://icoda-research.org/

publications related to real-world data and COVID-19 (figure 1), and a substantial number of research hypotheses were tested using real-world data. We categorised the available studies by assigning broad research topics to studies on the basis of their principal stated research question, according to prespecified categories of study designs based on consensus by two reviewers (LD and VK). Studies were classified as being related to safety and efficacy if they reported on clinical outcome measures, clinical biomarkers of disease, or any safety-related outcomes from interventions to treat or prevent COVID-19. Research with a primary objective of describing the influence of COVID-19 in a population of patients with existing and known disease status was classified as at-risk population research, whereas studies describing outcomes and their variance (such as time in hospital) among general COVID-19 populations was classified as research on identifying outcomes and trajectories. Reasearch was classified as risk-factor studies if they explored the quantitative relationship between a clinical parameter and associated COVID-19 outcomes.

The most frequent research questions pertained to the safety and efficacy of potential treatment for COVID-19 (21%), followed by specific at-risk population outcome research (16%), research identifying outcomes and trajectories of patients affected with COVID-19 (15%), and research identifying risk factors for COVID-19 and COVID-19 induced outcomes (11%). Our search might not have captured the full extent of research in this space, because studies incorporating EHRs might not have used the terms observational or real-world or their synonyms within their abstract or title. The publications from our search were used to inform our discussions and identify specific cases mentioned herein.

The COVID-19 pandemic has fostered many international collaborative efforts for real-world evidence research. One example is the international Consortium for Clinical Characterization of COVID-19 by EHR

(4CE).[15] The consortium gathered data of 27 584 patients diagnosed with COVID-19 between Jan 1, 2020, and April 11, 2020, as well as 187 802 laboratory tests from 96 hospitals across five countries within 3 weeks. This initiative provided early data on the disease progression of people with COVID-19. Similarly, descriptive statistics harnessing large-volume EHRs were instrumental in the rapid identification of at-risk populations, such as the disproportionate morbidity and mortality of Black and Asian people in the UK being reported as early as March, 2020.[16] This work was subsequently validated via the OpenSAFELY review of over 17 million primary care records in England, which similarly found that Black and South Asian people with COVID-19 had significantly higher mortality rates than white people did, even after adjustment for other sociodemographic details.[17]

Other international endeavours include the International COVID-19 Data Alliance, which aims to build an open and trustworthy international research partnership to support a rapid response to COVID-19, and a long-term alliance for making data accessible to health researchers and scientists worldwide. Collaborative approaches to data, technology, public engagement, and governance infrastructures are supporting 12 research projects, enabling data sharing across up to 42 countries.[18] Sponsored projects include population-driven estimates of clinical complications across different COVID-19 waves in South Africa[19] and assessing the efficacy of vaccines in marginalised communities within Brazil.[20]

Separate to these initiatives were developments by groups traditionally involved in offering commercial real-world data services. Multiple organisations offer these services, typically partnering with either health networks or insurance providers depending on the availability of data.[21] These organisations frequently work with regulatory and reimbursement agencies to evaluate treatment patterns in non-randomised controlled settings.[22] These existing networks pivoted towards COVID-19, with commercial vendors of real-world data rapidly advertising services relating to COVID-19 patient data from their health-care networks predominantly based in North America; there has been little coverage of commercial COVID-19 EHR data in other regions of the world.[23]

Differing health record systems across organisations and jurisdictions and localised processes still posed challenges in using these data for research during the COVID-19 pandemic. Case definition and case ascertainment were initially evolving concepts with little access to internationally agreed-upon criteria. During the early days of the pandemic, EHR systems did not have a systematic way to identify patients with COVID-19.[24] To overcome this challenge, an International Classification for Diseases (ICD)-10 medical code for COVID-19 was developed by April 1, 2020.[25] By December, 2020, updated ICD-10 codes for encounters related to COVID-19 (such

as encounters for screening for COVID-19, or pneumonia due to COVID-19) were developed.[26] These advances made identifying patients with COVID-19, and conducting COVID-19-related research using routine health-care data, less confounded by poor data access and heterogeneity in disease definitions. Despite these advances, most patients were still cared for outside health-care systems with digitised medical coding, and health-care system burden alongside a rapid introduction of new coding might have contributed to incomplete data capture, particularly at the earlier stages of the pandemic and for marginalised populations.[27]

There have been high-profile examples of routinely collected data generating high-value research outputs using combinations of innovative methodologies and data sources. Much attention has been focused on the use of these large-scale data-ready formats to assess vaccine effectiveness for COVID-19. One such example used causal inference methodologies applied to centrally provided electronic health data on 596 618 individuals who had received a COVID-19 vaccine to determine vaccine effectiveness in Israel.[28] This study was one of the first large-scale studies for vaccine effectiveness in real-world settings. Similarly, a study used national linked data, based on routinely collected EHR data from different national databases, on 5·4 million residents of Scotland (approximately 99% of the Scottish population) to prospectively evaluate the real-world effectiveness of vaccination efforts in health-care personnel, using multiple vaccine types and timepoints.[29] With this dataset, data on the differential risk of mortality from the delta variant was available within 6 weeks of its identification.[30] These methods provide important insights to vaccination efforts for COVID-19 not achievable by conventional trials, owing to the particularly large sample sizes that are achievable with these methods.

Major advancements in using routinely collected health data for research have been made over the past 15 years, and this process has continued during the pandemic. However, there is a need for a comprehensive checklist to critically appraise research that uses real-world data to better promote high-quality research. The regulatory framework for the use of real-world data in clinical research is distinct from clinical trials;[14] therefore, the associated assumptions on quality assurance for clinical trials might not be directly translated to real-world evidence research. Checklists, such as the 4CE checklist for EHR research in COVID-19,[31] have been proposed and other broader tools, such as the STaRT-RWE guidelines[32] and ISPOR-ISPE recommendations,[33] which were developed to critically appraise studies that use real-world evidence, have been suggested to critically appraise studies that use EHR data. These concepts are not necessarily new. Indeed, the RECORD statement[34] in 2015 has been used extensively for studies using observational evidence from routinely collected health data.

Whether these criteria receive the international endorsement of other critical appraisal tools, such as the CONSORT criteria for clinical trials, is unclear.[35] Currently, there is little endorsement by medical journals on standards for publication of research harnessing EHRs or other real-world data sources.[31] Standards for publication could improve reporting of key study characteristics, as noted in other clinical research areas, and this has often been achieved together with groups of journals and their editors.[36] Whether such standards can be applied to EHR-related research remains to be seen, owing to the multifaceted analytical and research questions addressed using RWD.

Advancements in routinely collected health-care data for individual research projects and opportunities have occurred,[28,29] but little success has been noted with their integration into clinical trial activities. Pragmatic clinical trials, a design that harnesses the existing capture of patients' health-care data[37] to generate testable hypotheses on health-care interventions, were argued to be uniquely applicable to the emergent nature of COVID-19.[38] However, few examples of published pragmatic trials for COVID-19 management are noted.[39] There remains uncertainty as to whether the small number of pragmatic trials on COVID-19 is owed to an incomplete existing infrastructure, insufficient familiarity with the design-related challenges of such trials, or challenges in coordinating such trials.[38] Complications might exist in a pandemic context, with substantial non-capture of medical services administered in systems without electronic data capture, and for contexts in which the health-care system was overwhelmed and organisational structuring of data was deprioritised for patient care.[27]

Routinely collected medical data present a unique opportunity to address health inequities. As we have identified, multiple studies were able to rapidly identify disproportionate burdens of COVID-19-associated hospitalisation and mortality in minority populations.[16,17] Global health inequities are often widened by low visibility into these inequities, and little engagement and active recruitment of diverse clinical research groups into both active clinical trials and retrospective database reviews.[40] As such, the rapid and representative acquisition of routinely collected health data of patients with COVID-19[17] could account for one illustrative example of how these routinely collected data might result in more representative and equitable health research.

Other uses of routinely collected data to address COVID-19 health inequities have been focused on the possible role of EHRs to target populations or geographies with proportionately lower vaccination uptake.[41] These efforts must be counterbalanced against the potential for research based on real-world evidence to inadvertently worsen these inequities. Data sources used in real-world evidence research are subject to their own selection bias for some patient demographics, owing to disparities in health-care use, and these can result in unbalanced

	Solutions implemented during the COVID-19 pandemic	Solutions to improve future digital-health technology research
Routinely collected health-care data		
Inadequate standardisation, including administrative codes for data capture and comparability	Adoption of shared administrative codes for COVID-19 and related conditions[32]	Adoption of comprehensive checklists for research from these data by authors and scientific journals
Variability in uptake and availability for research use across different geographical settings	Screening of EHRs to identify populations at high-risk of COVID-19[18,28]	Adoption of community-driven solutions and collaboration with researchers and community leaders for EHR research
Problems with accessibility of technology platforms due to cost and insufficient technology infrastructure	Causal inference methods applied to real-world observational data[34]	Implementation of standardised data capture, dictionaries, and technology systems
Absence of unique patient identifiers resulting in potential duplicated patient records	Applications of real-time predictive analytics for in-hospital mortality[47]	Educational foundations to improve researcher and reader literacy in associated methods and limitations
Data sharing		
Absence of organisational support, staff, and incentives	Rapid dissemination of annotated imaging data[48,49]	Building good quality and accessible common data infrastructures for scientific communities
Infrequent audit and enforcement of data sharing by scientific journals and governing bodies	Rapid dissemination of disease models alongside associated codes and datasets[50]	Continued mandate and reinforcement of data reporting in trial registries and other forms of data sharing by regulators and scientific journals
Multiplicity in data-sharing avenues, increasing the burdens on data collectors	..	Establishment of quasi-automated data pipelines and review processes

EHRs=electronic health records. Concepts in the table were informed by the literature search, in conjunction with discussions with owners of trial data, research funding organisations, and data scientists.

Table: Summary of the barriers to, and challenges of, research using digital health technology, the solutions implemented during the COVID-19 pandemic and future, long-term solutions

representations of patient disease burden.[42] It is important to emphasise that the use of medical services itself might not be the source of these disparities, rather access to services that are expressed as use are likely influential.[43] Further, some authors contend that the use of race as an identifier in medical research is itself problematic, owing to its frequent use as a poor surrogate for socioeconomic health determinants.[44] Indeed, this problem could be further compounded in routinely collected health data, for which recording of patients' ethnicity and race has been noted to be suboptimal and non-representative.[45] Accordingly, developing recording structures and reporting standards that acknowledge and articulate health inequities and racial biases in medical research might be beneficial in contextualising the health research generated for marginalised populations.[46]

The generation of these data has proven an essential element of research insights in the COVID-19 pandemic, yet their capture, reporting, and conversion to meaningful and interpretable outputs are conditional on two crucial aspects: the ability for insights to be shared both with the general public and other researchers, and the tools through which they are analysed. As such, we describe the importance of data sharing and its role in the COVID-19 pandemic.

Data sharing for COVID-19 research

Medical research data is multifaceted in the way that it is captured, covering data such as from qualitative surveys or wearable medical devices, efficacy and safety data from clinical trials, routinely collected health data, and audits. Accordingly, the ways in which data are shared are highly variable, ranging from direct transfer of raw medical data on individual patients to printed summaries of clinical trials published in medical journals or press releases. In these ways, the data are shared to enable and facilitate further research, as well as to provide quality assurance if required. Data, regardless of capture and primary type, are shared and assessed in a multitude of formats, ranging from imaging data through to highly ordered clinical codes and numerical outputs of laboratory assessments. Even within these subformats, varying specifications exist with regards to data standards and the software required to read and use the data. Accordingly, challenges exist both in the ability to share and use core medical data (table).

Although the most detailed form of data sharing would involve individual patient-level data, the most frequent form of data sharing occurs in publications as aggregate summary-level data. The framework for sharing of observational data, such as EHRs, is not as clear as the framework for sharing data of clinical trials, with varying formats being proposed.[15] Several coordinated efforts have been made to improve data flows and sharing over the past decades. Some efforts include the adoption of FAIR (findable, accessible, interoperable, and reusable) principles;[51] these principles represent a framework to improve data sharing across scientific disciplines, focusing on four domains of data properties. Separately there have been improvements to user interfaces and project-request pipelines from existing data sharing platforms. In 2017, the international committee of medical journal editors mandated data sharing of clinical trial data,[52] and many funding bodies for medical research around the world have necessitated data sharing from any funded clinical research.[53] Research participants could also be advocates for data sharing, with previous research indicating that a majority are willing to share their data with both the public and private sectors.[53] Although many might wish for better data sharing, this is not invariably true for all individuals involved in research, and gaps exists across geographies, age groups, and other sociodemographic factors.[54] Regardless of these broad characteristics, individuals might not trust the existing safeguards or research communities using this data. As such, without appropriate engagement of patient groups to ensure individuals are more informed about how their data will be used for research, the potential exists to exacerbate these existing gaps and perpetuate health inequities.

In particular, challenges exist for medical data shared at the individual patient level. In interviews with clinical trialists', Rathi and colleagues[55] identified clinicians and

researchers involved in clinical trials concerned with potentially misleading secondary analyses, implications for future planned research activities on the primary data, financial and workload burdens of data sharing, and concerns over ethics approval or regulatory implications of data sharing. These concerns have been echoed in systematic reviews on the topic of sharing individual patient data, additionally noting concerns over reidentification of research participants.[56] These challenges are particularly intensified when considering global health inequities in data-sharing exercises, for which access to resources that facilitate transfers might be more scarce (ie, in low-income and middle-income countries), alongside concerns over extractive research practices.[57] As such, there exists an intrinsic balance to be struck between the capacity to share individual patient data to facilitate novel and meaningful research activities and the individual burdens on principal investigators and research participants. This balance becomes increasingly challenging in the context of a global pandemic, as the relative benefit of novel research might be offset by the rapid pace of data generation and its subsequent pathways to data sharing agreements.[58]

Despite the significant support received for improving data sharing for the COVID-19 pandemic, sharing clinical trial data was still challenging.[59] Although international registries have seen close to 3000 interventional clinical trials registered for COVID-19, there have been only 121 published articles on drug treatments for COVID-19 as of April 25, 2021.[60] Despite large trial registration numbers, less than 4% of trials have been published as of April 25, 2021, and less than 2% of trials registered had results available within associated trial registries by this date.[61] Of these, an unknown percentage have data-sharing agreements established, although the low percentage of available evidence indicates that this might only be a small part of the total evidence established to date for COVID-19 trials. Of 132 545 studies registered on the WHO International Clinical Trials Registry Platform between January, 2019, and December, 2020, 11·2% stated that individual patient data would be shared. Studies of COVID-19 on this platform had similar rates of data agreements to studies of other non-pandemic diseases in 2020 (13·7%).[62]

Within the context of the COVID-19 pandemic, there remains an absence of centralised, sponsor-driven, primary data sharing. As of July 7, 2022, there were 8012 studies registered on ClinicalTrials.gov, mixed between observational and interventional trials, identified as being associated with COVID-19 (figure 2). Across all registered COVID-19 research studies, a total of 259 (3·2%) had their results posted on the platform. This is despite 4239 (52·9%) studies having completion dates listed as Dec 31, 2021, or earlier (figure 2). It is important to note that there are multifaceted explanations as to why data are not shared in this way, despite the legal requirement in the USA for data sharing via this
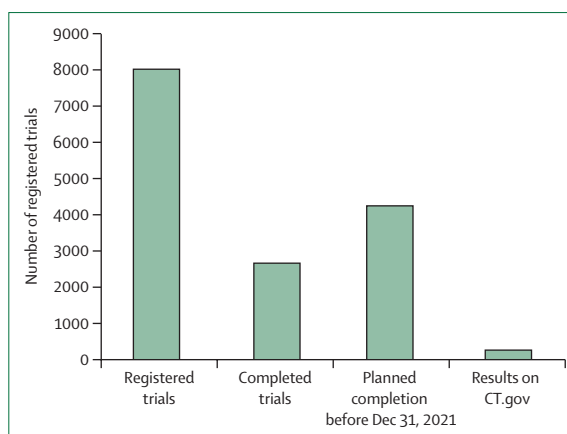

Figure 2: **Data sharing of registered clinical trials investigating COVID-19 from CT.gov**
CT.gov=ClinicalTrials.gov

platform,[63] including (but not limited to) an unawareness the legal requirements, poorer than anticipated trial results, trial misconduct, statistically important under-recruitment, and language barriers.

For researchers involved in clinical research integrating data discussed in this Series paper, the scarcity of data sharing during the COVID-19 pandemic is perhaps not a surprise, but it is a disappointment. Data sharing requires resources and investment. Data sharing and governance agreements must be made between organisations involved in the data transfer from one organisation to another. When international research is conducted there are potentially multiple legal frameworks that must be adhered to, creating challenges when synthesising data assets. Further, data transfer can involve many branches, from the originating source to specific external groups, and then from those groups to others, depending on the data being organised. Organising data for sharing purposes, as opposed to the originally defined research question, requires dedicated personnel time, both on behalf of the body providing data and the organisation receiving and disseminating data.[64] Simultaneously, there are recognised sociopolitical barriers to data sharing both within and between institutions,[65] and although these are potentially one of the most substantial barriers to data sharing, they are also among the most challenging to quantify and eradicate, even within the context of a global pandemic.

Although there are many metaregistries, a type of clinical data registry that houses or links the data from multiple individually unique clinical data sources, the absence of a common data dictionary across different individual data sources can restrict data linkage and accordingly creates the need for staff time and resources to minimise discordant definitions. Additionally, individual researchers could be concerned that reanalysis of their data for alternative research questions might lead to contradictory conclusions from their original statistical

analyses, and hence might be less willing to share their data.[53] Finally, there might be commercial or financial reasons to restrict access to data, including proprietary information or data which could be subsequently monetised by groups incorporating data into their analyses.

Several data sharing initiatives, however, have been organised through existing bodies and new entities in attempts to facilitate data sharing.[66–68] For example, F1000Research is an example of an open research publishing platform that offers peer review, data deposition, and sharing of open research following the FAIR principles. Vivli is another example of a data-sharing initiative for clinical research data. This global clinical research data-sharing platform offers 21 COVID-19 studies eligible for review as of April 22, 2021; however, this represents only 0·74% of the trials registered on ClinicalTrials.gov.[61,66] This is despite widespread acknowledgment of the importance of data sharing, particularly in the context of emerging diseases with high unmet needs. Unlike for clinical trials, data sharing for EHRs and other non-trial data are less well established. Many of the existing data-sharing platforms are exclusive to clinical trial data. The data format of these data-sharing platforms is usually not in the Observational and Medical Outcomes Partnership format that could be more compatible with the clinical trial data standards set by the Clinical Data Interchange Standards Consortium.[69] Despite this drawback, data sharing of non-clinical trial data in COVID-19 has been promoted and distributed by several pre-existing and new non-commercial groups.[15,70,71]

Reproducibility is a fundamental concern in clinical research, particularly for digital-health technologies.[72] In the context of COVID-19, high-profile instances of fraudulent research activities[73,74] and increased public engagement with emerging research means that the importance of timely and independently conducted reproducible analytics is greater than ever. This issue is particularly pressing when integrating digital-health technologies into clinical research that harnesses and adapts to continually updated data, such as artificial intelligence and machine learning methodologies. In particular, the very research challenges that lend themselves to artificial-intelligence and machine-learning methodologies and data capture facilitate greater opportunities for data sharing than other primary data generation activities (such as in-vitro assays or psychological research). These opportunities are due to the requirements of these methods to use high volumes of data, which in turn can drive collaborative efforts to optimise the time invested in data acquisition and curation.[75] By reducing traditional barriers to data-sharing access for the purposes of reproducibility, the capacity for iterative and robust research harnessing digital-health technologies can simultaneously become more timely and reliable.[72]

Outside of reproducibility, timely data transfer of both aggregate data and individual patient data is another issue that needs to be addressed. Traditional timelines for data sharing do not facilitate timely data transfer within the context of a health crisis such as a pandemic. There is no singular timeline for data sharing, with the timeline conditional on the data type shared (ie, individual patient *vs* aggregate), the data types handled (ie, summary group statistics *vs* original imaging data), and existing tools to share data.[64,65] Research communities can improve responsiveness for future pandemics and medical research by modifying technological and legal frameworks with regular audits and better enforcement of data sharing. Auditing and enforcing data sharing are not without historical precedent. On April 28, 2021, the US Food and Drug Administration issued a notice of non-compliance to a trial sponsor who had failed to submit summary results within the federally mandated reporting period.[63] Although this pertains to an example of aggregate data sharing on a particular platform, similar frameworks could be envisioned for other data types, particularly when pressing health crises necessitate expedited timelines.

Guidance without enforcement or audit can be subject to significant variability and is more challenging to monitor. On the basis of the findings above, the current pace of reporting is insufficient. Deciding on what the appropriate timeframes are for data transfer is challenging, owing to the multifaceted nature of clinical research. Instead, a pragmatic system of active encouragement and ingrained funding opportunities to incentivise research groups and provide resources associated with data sharing might help in this task and improve data sharing agreements for future health emergencies.

Outside of clinical data, both data sharing and dissemination of aggregated health system metrics were often characterised by high-profile interest atypical of academic sources, particularly case, hospitalisation, and vaccination tracking dashboards, such as the John Hopkins University interactive web tracker.[76] Outside of academic impact and citations, the data were published in a shareable format via a GitHub repository, which itself has been used in the reporting of multiple media outlets. Pivotal to the success of this web tracker was the rapid publication (in February, 2020) and data being in a free, accessible, and standardised data format.

Similarly, there has been enormous success in the sharing of key molecular data at a rapid pace. For example, the full genomic characterisation of the SARS-CoV-2 virus was published on Jan 30, 2020.[77] This timely data sharing has been instrumental in preclinical characterisation of potential drug targets and vaccine development. Although clinical data and genomic data clearly have several differences, the framework of genomic data development and dissemination across large study teams could prove informative for subsequent

Search strategy and selection criteria

To identify relevant studies related to electronic health records (EHRs) and COVID-19, electronic bibliographical databases of published research on MEDLINE were searched on July 7, 2022, using the terms "((COVID 19 or COVID-19 or 2019-nCoV Infection$ or 2019 nCoV or coronavirus or COVID19) and (Observational Study/ or observational stud* or real world or real-world or real life or real-life))" for articles published after Jan 1, 2019. Full-text English language reports and reference lists were reviewed to gain insights into the different types of EHR research being conducted for the COVID-19 pandemic.

digital-health applications to improve data sharing. For example, sequencing of SARS-CoV-2 variants has been facilitated through several viral sequencing consortia including GISAID, Nextstrain, and Pango. These activities have provided valuable data for research focused on in silico work.[78]

Conclusion

With COVID-19 there has been growing anticipation of paradigm shifts and fundamental overhauls to medical research, particularly in the application of digital-health technologies. Enormous shifts to the routine research process are evidenced by the rapid development and implementation of diagnostic and prognostic criteria, interventions to minimise patient burden, and in global vaccination programmes. The topics of data and methods for data sharing covered in this Series paper are united in their long history before COVID-19, which were awaiting widespread acceptance and routine integration into health-care research programmes. In some cases, incremental gains over the preceding years have translated into tangible benefits in developing meaningful and actionable tools and data to help manage the pandemic. In other cases, existing barriers (technological, sociological, and operational) have reduced the effect of these digital-health technologies. A key question remains: if the imperative of the COVID-19 pandemic was insufficient to progress data collection and sharing efforts, are the issues identified intractable?

It is often stated that necessity is the mother of invention. However, uncertainty now exists on how these innovations can be sustained in the post-pandemic era when existing barriers might become more entrenched, or be considered insurmountable. As such, a unique and rare opportunity exists, as the lessons learned remain current, to potentially translate the successes and short-comings of digital tools in COVID-19 into meaningful, sustainable, and equitable tools of transformational change.

For **GISAID** see https://www.gisaid.org/

For **Nextstrain** see https://nextstrain.org/

For **Pango** see https://cov-lineages.org/

References
1 Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. *N Engl J Med* 2020; **382:** e82.
2 Simpson CR, Shi T, Vasileiou E, et al. First-dose ChAdOx1 and BNT162b2 COVID-19 vaccines and thrombocytopenic, thromboembolic and hemorrhagic events in Scotland. *Nat Med* 2021; **27:** 1290–97.
3 Cake C, Ogburn E, Pinches H, et al. Development and evaluation of rapid data-enabled access to routine clinical information to enhance early recruitment to the national clinical platform trial of COVID-19 community treatments. *Trials* 2022; **23:** 62.
4 Abd-Alrazaq A, Alajlani M, Alhuwail D, et al. artificial intelligence in the fight against COVID-19: scoping review. *J Med Internet Res* 2020; **22:** e20756.
5 Dhruva SS, Ross JS, Akar JG, et al. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *NPJ Digit Med* 2020; **3:** 60.
6 Palayew A, Norgaard O, Safreed-Harmon K, Andersen TH, Rasmussen LN, Lazarus JV. Pandemic publishing poses a new COVID-19 challenge. *Nat Hum Behav* 2020; **4:** 666–69.
7 Jensen EA, Kennedy EB, Greenwood E. Pandemic: public feeling more positive about science. *Nature* 2021; **591:** 34.
8 Harper L, Kalfa N, Beckers GMA, et al. The impact of COVID-19 on research. *J Pediatr Urol* 2020; **16:** 715–16.
9 US FDA. Framework for FDA's Real-World Evidence Program.2018. https://www.fda.gov/media/120060/download (accessed May 20, 2021).
10 Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *J Intern Med* 2013; **274:** 547–60.
11 Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and external controls in clinical trials—a primer for researchers. *Clin Epidemiol* 2020; **12:** 457–67.
12 Dolin RH, Alschuler L, Boyer S, et al. HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* 2006; **13:** 30–9.
13 Nordo AH, Levaux HP, Becnel LB, et al. Use of EHRs data for clinical research: historical progress and current applications. *Learn Health Syst* 2019; **3:** e10076.
14 Klonoff DC. The new FDA real-world evidence program to support development of drugs and biologics. *J Diabetes Sci Technol* 2020; **14:** 345–49.
15 Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med* 2020; **3:** 109.
16 ICNARC. ICNARC report on COVID-19 in critical care. London: ICNARC, 2020.

17 Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; **584:** 430–36.

18 Stock S, Zoega H, Brockway M, et al. The international Perinatal Outcomes in the Pandemic (iPOP) study: protocol *Wellcome Op Res* 2021; **6:** 21.

19 Davies MA, Kassanjee R, Rousseau P, et al. Outcomes of laboratory-confirmed SARS-CoV-2 infection in the Omicron-driven fourth wave compared with previous waves in the Western Cape Province, South Africa. *Trop Med Int Health* 2022; **27:** 564–573.

20 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.e1–736.e4

21 Baumann J. Pandemic Opens Door to More Real-World Data for Drug Decisions. New York, NY: Bloomberg Law, 2021.

22 HealthVerity. HealthVerity announces research collaboration with FDA to support COVID-19 clinical study and treatment opportunities. PRNewswire. June 9, 2020. https://www.prnewswire.com/news-releases/healthverity-announces-research-collaboration-with-fda-to-support-covid-19-clinical-study-and-treatment-opportunities-301072414.html (accessed May 20, 2021).

23 Akanbi MO, Ocheke AN, Agaba PA, et al. Use of electronic health records in sub-Saharan Africa: progress and challenges. *J Med Trop* 2012; **14:** 1–6.

24 Kadri SS, Gundrum J, Warner S, et al. Uptake and accuracy of the diagnosis code for COVID-19 among US hospitalizations. *JAMA* 2020; **324:** 2553–54.

25 FDA. ICD-10-CM Official Coding Guidelines.Feb 20, 2020. https://www.cdc.gov/nchs/data/icd/interim-coding-advice-coronavirus-March-2020-final.pdf (accessed April 30, 2021).

26 FDA. New ICD-10-CM code for the 2019 novel coronavirus (COVID-19). Dec 3, 2020. https://www.cdc.gov/nchs/data/icd/Announcement-New-ICD-code-for-coronavirus-19-508.pdf (accessed April 30, 2021).

27 Despins LA, Wakefield BJ. The role of the electronic medical record in the intensive care unit nurse's detection of patient deterioration: a qualitative study. *Comput Inform Nurs* 2018; **36:** 284–92.

28 Dagan N, Barda N, Kepten E, et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. *N Engl J Med* 2021; **384:** 1412–23.

29 Swift MD, Breeher LE, Tande AJ, et al. Effectiveness of messenger RNA coronavirus disease 2019 (COVID-19) vaccines against Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infection in a cohort of healthcare personnel. *Clin Infect Dis* 2021; **73:** e1376–79.

30 Sheikh A, Robertson C, Taylor B. BNT162b2 and ChAdOx1 nCoV-19 vaccine effectiveness against death from the delta variant. *N Engl J Med* 2021; **385:** 2195–97.

31 Kohane IS, Aronow BJ, Avillach P, et al. What every reader should know about studies using electronic health record data but may be afraid to ask. *J Med Internet Res* 2021; **23:** e22219.

32 Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021; **372:** m4856.

33 Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value Health* 2017; **20:** 1003–08.

34 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.

35 Schulz KF, Altman DG, Moher D, Fergusson D. CONSORT 2010 changes and testing blindness in RCTs. *Lancet* 2010; **375:** 1144–46.

36 Hopewell S, Altman DG, Moher D, Schulz KF. Endorsement of the CONSORT statement by high impact factor medical journals: a survey of journal editors and journal 'Instructions to Authors'. *Trials* 2008; **9:** 20.

37 Weinfurt KP, Hernandez AF, Coronado GD, et al. Pragmatic clinical trials embedded in healthcare systems: generalizable lessons from the NIH Collaboratory. *BMC Med Res Methodol* 2017; **17:** 144.

38 Almufleh A, Joseph J. The time is now: role of pragmatic clinical trials in guiding response to global pandemics. *Trials* 2021; **22:** 229.

39 Sholzberg M, Tang GH, Negri E, et al. Coagulopathy of hospitalised COVID-19: a pragmatic randomised controlled trial of therapeutic anticoagulation versus standard care as a rapid response to the COVID-19 pandemic (RAPID COVID COAG - RAPID Trial): a structured summary of a study protocol for a randomised controlled trial. *Trials* 2021; **22:** 202.

40 Cash-Gibson L, Rojas-Gualdrón DF, Pericàs JM, Benach J. Inequalities in global health inequalities research: a 50-year bibliometric analysis (1966–2015). *PloS One* 2018; **13:** e0191901.

41 Olusanya OA, Bednarczyk RA, Davis RL, Shaban-Nejad A. Addressing parental vaccine hesitancy and other barriers to childhood/adolescent vaccination uptake during the coronavirus (COVID-19) pandemic. *Front Immunol* 2021; **12:** 663074.

42 Bower JK, Patel S, Rudy JE, Felix AS. Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Curr Epidemiol Rep* 2017; **4:** 346–52.

43 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

44 Ioannidis JPA, Powe NR, Yancy C. Recalibrating the use of race in medical research. *JAMA* 2021; **325:** 623–24.

45 Knox S, Bhopal RS, Thomson CS, et al. The challenge of using routinely collected data to compare hospital admission rates by ethnic group: a demonstration project in Scotland. *J Public Health* 2020; **42:** 748–55.

46 Flanagin A, Frey T, Christiansen SL. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* 2021; **326:** 621–27.

47 Kim KM, Evans DS, Jacobson J, Jiang X, Browner W, Cummings SR. Rapid prediction of in-hospital mortality among adults with COVID-19 disease. *PLoS One* 2022; **17:** e0269813.

48 Ebadi A, Xi P, MacLean A, et al. COVIDx-US: an open-access benchmark dataset of ultrasound imaging data for AI-driven COVID-19 analytics. *Front Biosci (Landmark Ed)* 2022; **27:** 198.

49 Gunraj H, Wang L, Wong A. COVIDNet-CT: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images. *Front Med* 2020; **7:** 608525.

50 Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep* 2020; **10:** 19549.

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

52 Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the international committee of medical journal editors. *JAMA* 2017; **317:** 2491–92.

53 Terry RF, Littler K, Olliaro PL. Sharing health research data—the role of funders in improving the impact. *F1000 Res* 2018; **7:** 1641.

54 Anderson A, Borfitz D, Getz K. Global public attitudes about clinical research and patient experiences with clinical trials. *JAMA Netw Open* 2018; **1:** e182969.

55 Rathi V, Dzara K, Gross CP, et al. Sharing of clinical trial data among trialists: a cross sectional survey. *BMJ* 2012; **345:** e7570.

56 Bull S, Roberts N, Parker M. Views of ethical best practices in sharing individual-level data from medical and public health research: a systematic scoping review. *J Empir Res Hum Res Ethics* 2015; **10:** 225–38.

57 Pisani E, Whitworth J, Zaba B, Abou-Zahr C. Time for fair trade in research data. *Lancet* 2010; **375:** 703–05.

58 Mather N. How we accelerated clinical trials in the age of coronavirus. *Nature* 2020; **584:** 326.

59 Park JJH, Mogg R, Smith GE, et al. How COVID-19 has fundamentally changed clinical research in global health. *Lancet Glob Health* 2021; **9:** e711–20.

60 Thorlund K, Dron L, Park J, Hsu G, Forrest JI, Mills EJ. A real-time dashboard of clinical trials for COVID-19. *Lancet Digit Health* 2020; **2:** e286–87.

61 Dron L, Dillman A, Zoratti MJ, Haggstrom J, Mills EJ, Park JJH. Clinical trial data sharing for COVID-19–related research. *J Med Internet Res* 2021; **23:** e26718.

62   Merson L, Ndwandwe D, Malinga T, et al. Promotion of data sharing needs more than an emergency: an analysis of trends across clinical trials registered on the International Clinical Trials Registry Platform. *Wellcome Open Res* 2022; **7:** 101.

63   FDA. FDA takes action for failure to submit required clinical trial results information to ClinicalTrials.gov. Apr 28, 2021. https://www.fda.gov/news-events/press-announcements/fda-takes-action-failure-submit-required-clinical-trial-results-information-clinicaltrialsgov?utm_medium=email&utm_source=govdelivery (accessed April 30, 2021).

64   Miller J, Ross JS, Wilenzick M, Mello MM. Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices. *BMJ* 2019; **366:** l4217.

65   Butte AJ. Trials and tribulations—11 reasons why we need to promote clinical trials data sharing. *JAMA Netw Open* 2021; **4:** e2035043.

66   Bierer BE, Li R, Barnes M, Sim I. A Global, Neutral Platform for Sharing Trial Data. *N Engl J Med* 2016; **374:** 2411–13.

67   Transcelerate Biopharma. COVID-19 2021. https://www.transceleratebiopharmainc.com/covid-19/ (accessed April 22, 2021).

68   Dawex. The COVID-19 Data Exchange initiative. 2021. https://www.covid19-dataexchange.org/ (accessed April 22, 2021).

69   Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: transforming i2b2 data into the OMOP common data model. *PLoS One* 2019; **14:** e0212463.

70   Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Brief Bioinform* 2021; **22:** 812–22.

71   C3AI. C3 AI COVID-19 Data Lake. 2020. https://c3.ai/products/c3-ai-covid-19-data-lake/ (accessed April 22, 2021).

72   Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med* 2019; **2:** 2.

73   Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020: S0140–6736(20)31180–6.

74   Popp M, Stegemann M, Metzendorf MI, et al. Ivermectin for preventing and treating COVID-19. *Cochrane Database Syst Rev* 2021; **7:** CD015017.

75   Open Science Collaboration.Psychology: estimating the reproducibility of psychological science. *Science* 2015; **349:** aac4716.

76   Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; **20:** 533–34.

77   Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020; **395:** 565–74.

78   Shishir TA, Naser IB, Faruque SM. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. *PLoS One* 2021; **16:** e0245584.