AUTHORS:
Sunday O. Oladejo[1]
Liam R. Watson[1,2]
Bruce W. Watson[1]
Kanshukan Rajaratnam[1]
Maritha J. Kotze[3]
Douglas B. Kell[4,5,6]
Etheresia Pretorius[4,6]

AFFILIATIONS:
[1]School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa
[2]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
[3]Division of Chemical Pathology, Department of Pathology, National Health Laboratory Service, Tygerberg Hospital & Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[4]Department of Biochemistry and Systems Biology, Faculty of Health and Life Sciences, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK
[5]The Novo Nordisk Foundation Centre for Biosustainability, Technical University of Denmark, Lyngby, Denmark
[6]Department of Physiological Sciences, Faculty of Science, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Sunday Oladejo

EMAIL:
sunday@sun.ac.za

DATES:
Received: 08 Sep. 2022
Revised: 15 Feb. 2023
Accepted: 27 Mar. 2023
Published: 30 May 2023

HOW TO CITE:
Oladejo SO, Watson LR, Watson BW, Rajaratnam K, Kotze MJ, Kell DB, et al. Data sharing: A Long COVID perspective, challenges, and road map for the future. S Afr J Sci. 2023;119(5/6), Art. #14719. https://doi.org/10.17159/sajs.2023/14719

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☐ On request from author(s)
☐ Not available
☒ Not applicable

EDITOR:
Pascal Bessong

KEYWORDS:
Long COVID, data sharing, data science

FUNDING:
None

Data sharing: A Long COVID perspective, challenges, and road map for the future

'Long COVID' is the term used to describe the phenomenon in which patients who have survived a COVID-19 infection continue to experience prolonged SARS-CoV-2 symptoms. Millions of people across the globe are affected by Long COVID. Solving the Long COVID conundrum will require drawing upon the lessons of the COVID-19 pandemic, during which thousands of experts across diverse disciplines such as epidemiology, genomics, medicine, data science, and computer science collaborated, sharing data and pooling resources to attack the problem from multiple angles. Thus far, there has been no global consensus on the definition, diagnosis, and most effective treatment of Long COVID. In this work, we examine the possible applications of data sharing and data science in general with a view to, ultimately, understand Long COVID in greater detail and hasten relief for the millions of people experiencing it. We examine the literature and investigate the current state, challenges, and opportunities of data sharing in Long COVID research.

Significance:

Although millions of people across the globe have been diagnosed with Long COVID, there still exist many research gaps in our understanding of the condition and its underlying causes. This work aims to elevate the discussion surrounding data sharing and data science in the research community and to engage data sharing as an enabler to fast-track the process of finding effective treatment for Long COVID.

Introduction

Post-acute sequelae of COVID-19 (PASC), otherwise known as 'Long COVID', is a health crisis resulting from the COVID-19 pandemic. In essence, Long COVID is the long-term reoccurrence of the symptoms and health challenges associated with a COVID-19 infection.[1-3]

Although the definition of Long COVID has initiated many complex conversations globally[4,5], major Long COVID symptoms and complications agreed upon in the literature include: chest pain; heart palpitations; constant tiredness; muscular and joint pain; breathing difficulties (including low oxygen levels and shortness of breath); anosmia; difficulty concentrating; forgetfulness and brain fog; kidney problems; and digestive problems[3,6-8] (Figure 1). COVID-19 survivors who still experience these persistent symptoms are called 'Long haulers'.[9,10]

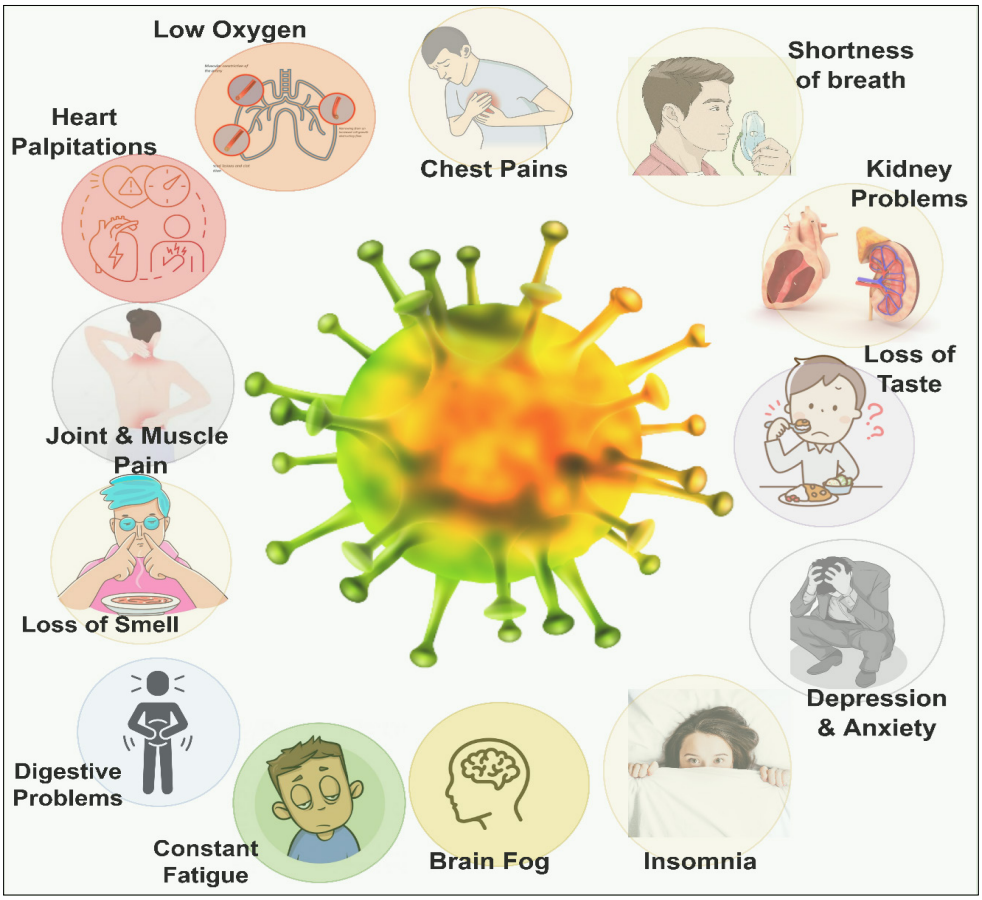

Figure 1: Illustration of the common Long COVID symptoms and complications reported in the literature.

The severity and rate of occurrence of Long COVID symptoms in Long haulers differs depending on the patient's health status prior to contracting COVID-19 and during treatment.[11] Because of this, there remains considerable debate among medical professionals regarding how to make Long COVID diagnoses and what optimal treatment plans should look like.[12] Disagreements and uncertainty often also result from the ways in which Long COVID data – post and prior to diagnosis (and treatment) – are collected, interpreted and reported.[13,14] Data collection can be affected by the way that questions are phrased, the types of surveys used, and the potential biases of participants.[15] Interpretation of the data can be affected by the way that they are presented, the types of analyses used, and the potential biases of the researchers. Reporting of the data can be affected by the way that data are summarised and the types of media outlets that are used, which can lead to miscommunication or confusion. As such, it is important to ensure that data collection, interpretation, and reporting are done in a transparent, unbiased manner in order to minimise disagreements and uncertainty. To this end, the processes involved in creating electronic health data and records must be more efficiently scrutinised and understood to avoid further muddying the waters.[11,14,16-18] A single platform is required for data processing extending from sample/information collection to report generation.

The lack of a consistent definition for Long COVID has resulted in diverse data sets, with the further consequence of ambiguity in defining patients' conditions and categorising based on patients' conditions.[11] Policies that define Long COVID can be improved in a variety of ways to better support Long COVID patients. First, there is a need to consider whether a new policy should be written, or rather be provided through an existing and appropriate form of management document. This would help healthcare providers to create standardised data collection and reporting systems that track Long COVID patient symptoms and health outcomes over time. These data could be aggregated and analysed to create a better understanding of the impact of Long COVID on patients, and to inform decisions about which treatments and interventions are most effective. The person responsible for keeping the data management plan or policy up to date must ensure that clear guidelines are provided for access and use in order to enforce adherence to the requirements. The lack of a standardised definition of Long COVID may also lead to unnecessary suffering on the individual level and exacerbates the existing strain on an already fragile global healthcare infrastructure and systems.

To establish effective and efficient management of Long COVID in patients, a standardised data capturing framework is therefore essential. A holistic data management framework would entail a wide-ranging collaboration across different specialities, drawing on research and expertise from a variety of sectors.[19] In this paper, we examine the present challenges of applying data science and artificial intelligence (AI) to the problem, together with a consideration of other multidisciplinary approaches to solving the Long COVID conundrum.

Data-driven frameworks in Long COVID management

Globally, healthcare organisations have accumulated several corpora of data from processes such as clinical workflows, drug trials, and patient medical records. These organisations are still, for the most part, utilising traditional approaches to recordkeeping and management. Traditional approaches to recordkeeping typically involve a paper-based system. This system includes the patient's medical records, research data, and trial forms being entered into paper-based forms, notebooks, and logbooks. This system is often labour-intensive, but it is an effective method for collecting and organising data in a clinical trial. However, it can lead to inefficiencies in operations, such as poor patient admission and treatment and an overall sub-optimal management of and preparedness for epidemics and pandemics.[20,21]

A data-driven approach to healthcare management will improve on the efficiencies, agility, and robustness of healthcare institutions, enabling them to meet the intersecting challenges of increasingly complex patient needs and navigate the potential of ever-evolving medical technology in a dynamic global society. To achieve this goal, data science, AI, and information technology will play vital roles.[22-24]

Data-driven systems can also play a vital role in the management of Long COVID. Figure 2 illustrates some of the benefits of data-driven Long COVID management. However, there is a paucity of open big data sets for Long COVID management, which may be attributed to the novelty of the disease.[25] Open big data sets are required by governments, healthcare institutions and policymakers across the world in designing capable healthcare systems to address the looming Long COVID crisis.[25]

The global move towards open science is largely seen as a positive development in the scientific community. Open science encourages the sharing of data, ideas, and methods, enabling researchers to collaborate more easily and efficiently. This promotes faster and more effective research and encourages the development of new approaches to research. Open science also allows for greater transparency and public engagement, as well as improved data accuracy and reproducibility. Ultimately, open science will help to ensure that scientific findings are as accurate and reliable as possible.[26,27]

In relation to Long COVID, the open science movement will be beneficial in helping researchers to collaborate and share data, which can be used to better understand the long-term effects of COVID-19. Open science can also provide a platform for patients to share their experiences and data, which can be used to inform further research. Furthermore, open data can be used to evaluate the effectiveness of treatments and develop new approaches to managing Long COVID. Ultimately, open science has the potential to advance our understanding of Long COVID and help to develop better strategies for prevention, diagnosis, and treatment.[13]

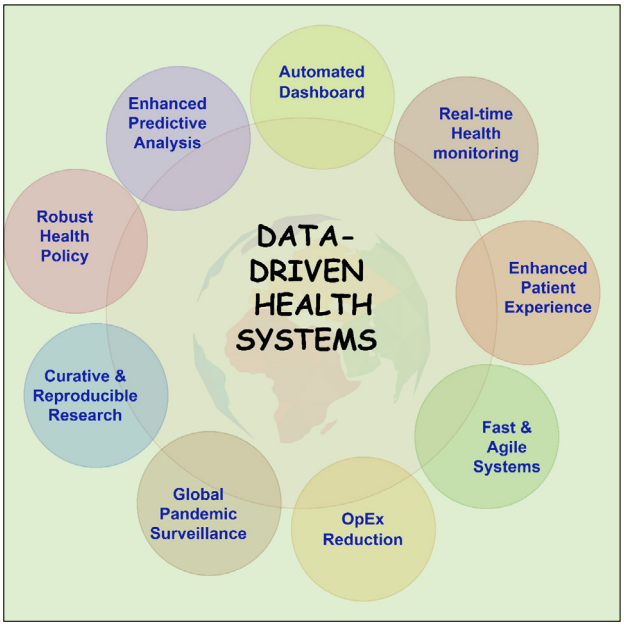

Figure 2: Benefits of adopting a data-driven framework for Long COVID management and healthcare systems in general.

Open big data sets for Long COVID

Data are a critical part of scientific research and the implementation of solutions proffered by researchers. Generally, data are also a major output in research endeavours, including clinical trials. Scientific data sets can be categorised as open sourced or closed source. Open-source data sets are available to everyone across the world without restriction. Open data sets support reproducible and collaborative research; enhance trust in research outcomes; and enforce best practices.[28] Closed-source data sets are not made available to the public to protect intellectual property rights and privacy. Closed-source data sets include government-classified and privately owned data. Researchers who engage in restricting access to their data sets often do not share the base codes, methods, or techniques with the research community.

Data-driven systems and AI run on large data sets that are typically sourced from multiple sources and, hence, include open data sets but not exclusively so. Data science and AI played an important role in surveillance, treatment, and vaccination in the COVID-19 era, which was made possible due to data sharing among researchers and professionals globally.

However, the story is not the same for Long COVID, as there are only a few open-source data sets available on Long COVID surveys, clinical trials, and research. We carried out a text and meta search for Long COVID data sets online and in related published works, and found a total of 12 related data sets. Table 1 presents the outcome of our findings.

Data sharing strategies

To foster data sharing for Long COVID research, establishing effective data sharing strategies is important. In data sharing, for Long COVID and other health-related research, there are two broad storage strategies: (1) the centralised approach and (2) the federated approach. In the centralised repository approach, each respective research hub, community, or institution hosts and curates its data sets in one central data warehouse or storage facility, which connects to all other research hubs. Simply put, all research hubs store their data sets in the same data warehouse or repository. This architecture or approach is well suited for research purposes and research-generated data sets. In the federated approach, each respective research hub has its own data warehouse for data storage and other research hubs can only access the data sets via a web server. In the federated approach, restrictions can be enforced by the data sets' owners due to data regulatory constraints and intellectual property rights. Each research hub is saddled with the responsibility of ensuring data privacy, security, and quality. The federated approach is well suited for electronic health data and records. Figure 3 illustrates the two approaches described above.

Potential challenges in data sharing for Long COVID research

Data availability and limitations

Owing to the novelty of Long COVID, there are few or, in some cases, no available data sets for researchers globally to compare notes. Moreover, the negligible quality of the available data sets may slow the process of finding appropriate solutions to Long COVID. The quality of a data set may, for instance, be undermined by the quality of available genomic sequences, unlabelled medical images, or low pixel resolution of medical images such as fluorescence microscopy and micrographs. Moreover, the population sizes of patients administered by a research community may also affect the generalisations and conclusions drawn from such studies.

Table 1: Related Long COVID data sets in the literature

Study	Country of study/participants	Number of participants	Mode of data sourcing	Duration of study	Data availability
Patient-led Research Collaborative[29]	56 Countries	3762	Online survey	6 Sep 2020 – 25 Nov 2020	On request
SA Long COVID[6,30]	South Africa	845	Online survey		–
Long COVID Support Group[31]	United Kingdom	114	Physical interview and focus group	May 2020 – Sep 2020	–
Schools Infection Survey Long COVID[32]	England	3779 Primary 2961 Secondary	Questionnaire	15 Mar 2022 – 1 Apr 2022	Available
Hiroshima Prefecture Survey[33]	Hiroshima, Japan	140	Self-administered questionnaire	25 Aug 2020 – 15 Mar 2021	On reasonable request
ZOE COVID-19 Tracker[34]	United Kingdom, USA, Sweden	4182	Phone app (self)	24 Mar 2020 – 2 Sep 2020	–
Symptom Burden Question for Long COVID (SBQ-LC)[35]	United Kingdom	274	Remote data collection and social media channels	14 Apr 2021 – 1 Aug 2021	–
DATCOV Post COVID Condition[36]	South Africa	1873	–	1 Dec 2020 – 23 Aug 2021	–
Long COVID Dataverse[37]	United Kingdom, Lesotho, Angola, Israel, USA	1131	–	Mar 2022	Available
Self-Reported Long COVID after Omicron[38]	United Kingdom	–	–	18 Jul 2022 – 6 May 2022	Available
Prevalence of Ongoing COVID19 Symptoms[39]	United Kingdom	–	–	1 Apr 2021 – 7 Jul 2022	Available
Kenya, Malawi, Long COVID effect survey[40]	Kenya Malawi	806 Kenya 885 Malawi		6 Sep 2021 – 2 Oct 2021	Available
American Academy of Physical Medicine (AAPM&R)[41]	USA	–	–	From July 2021	–

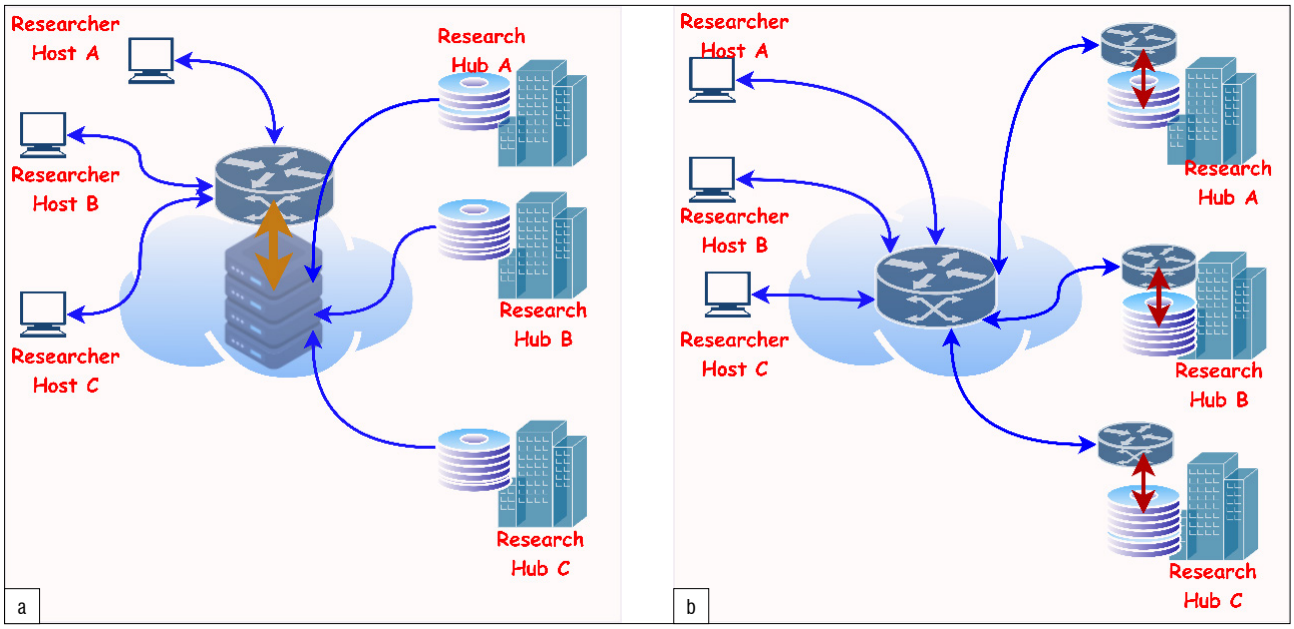

Figure 3: Illustration of the two main data-sharing strategies: (a) the centralised architecture and (b) the federated architecture.

The generalisation of AI-based medical systems is heavily reliant on the size and quality of the data used to train the system. With small data sets, it can be difficult to create an AI system that can generalise due to sample size issues, especially to new, unseen data. This is because small data sets can lead to a lack of diversity and a lack of statistical power, which can lead to overfitting and poor generalisation. Furthermore, small data sets can lack the necessary complexity to accurately capture the nuances of a medical problem. Therefore, when using an AI-based medical system, it is important to ensure that the data set used to train the system is large enough and of high enough quality to support accurate generalisation. Quality of data and data sets refers to a standardised definition of variables, and data sets that are difficult to harmonise. Moreover, creating AI models from data sets sourced from several research hubs or communities may be a daunting task, owing to different naming, file saving, and meta nomenclature, which could create serious problems when federating the data.

Ethics, privacy, and security

Ethics play a critical role in health sciences and medical professionals' ability to provide safe and effective diagnoses and treatment for patients. Clinical trials should always adhere to best practices.[42] COVID-19 and rising cases of Long COVID have initiated an intense discussion[12] over how to find a compromise between the undeniable urgency of a globally accepted treatment, and the necessity of maintaining global best practices and ethics. In finding and achieving the desired balance, the quality of data sets from processes such as clinical trials in finding effective Long COVID treatment should not be compromised. Scientific rigour is essential for patient safety. Moreover, a data scientist must also adhere to AI ethics[43], as illustrated in Figure4. In Figure 4, 'explication', also known as interpretability or explainability, is the transparency and the ability to understand how AI systems make decisions. For instance, an AI-powered medical diagnostic system that is opaque and not explainable could lead to mistrust among patients and healthcare providers. 'Non-maleficence' is closely related to the concept of safety in AI, in which AI-driven systems should not cause harm to humans or animals. For example, if an AI-powered medical diagnostic system misdiagnoses a patient, the patient could be harmed by receiving the wrong treatment. 'Autonomy' refers to the idea that individuals, communities, groups, and societies should have control over the use of AI systems that affect their lives. This principle is important to consider in AI development and deployment, as AI systems have the potential to make decisions that affect people's lives in many ways, such as employment, health care, and criminal justice. Moreover, AI systems should be fair and not perpetuate or exacerbate existing inequalities; for example, an AI-powered criminal justice system that has been trained on biased data could lead to discrimination against certain groups of people. In order to ensure that the system is fair and does not make decisions that perpetuate existing inequalities, it is imperative that the data and data sets generated and studied do not possess or reproduce racial, gender, age, sexuality, religious, or disability-based biases. Likewise, the AI models developed from the data sharing effort must be devoid of biases.

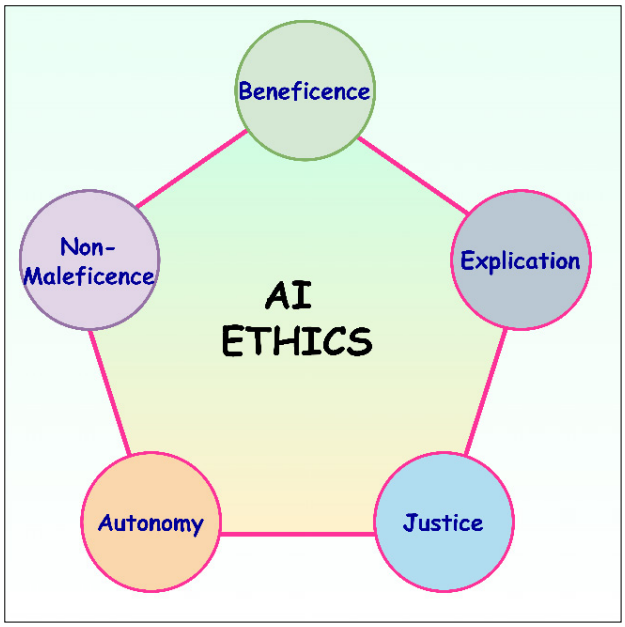

Figure 4: Pillars of artificial intelligence (AI) ethics.[43]

Sanctions and embargos on sharing information

Sanctions and embargos should not be placed on researchers and their respective home countries for sharing privacy-preserving Long COVID data sets, as this is both unreasonable and counterproductive. Such was infamously experienced by South African researchers as a consequence of their acting in the international community's best interests by sharing their data on the SARS-CoV-2 Omicron variant.[44-46] Travel restrictions put in place by the United Kingdom and other countries caused further damage to developing countries' struggling economies while also

worsening international relations. This incident generated discussions in research communities on the clear need to ensure that open science is not threatened. Long COVID researchers should be encouraged to look beyond narrow national interests and cultivate a global perspective in confronting Long COVID head on. Additionally, policymakers should consider long-term benefits of data sharing over narrow or irrational action which may result in short-term political benefits but hamper scientific discoveries and innovations. To illustrate this, globally, we now have two case studies to compare the consequences of sharing and not sharing data. In 2002, the Chinese government withheld SARS data and was severely criticised. However, travel bans were not enacted. This resulted in inadequate measures to prevent the virus spreading across borders.[47,48] On the other hand, the South African government's policy of open and transparent data sharing resulted in travel bans and restriction on freedom of movement.[47] The latter had a negative impact on the economy and an adverse effect on import of much-needed medical products, resulting in further suffering. The negative reaction to South Africa's sharing of data disincentivises countries from sharing data that may result in consequences for the global health system.[47]

Open science, virtual research collaborations, massive use of open access repositories, and agile research publication models should be encouraged, even in closed-border or travel-restricted situations.[49-52] Open access publishing models should be encouraged to ensure that research results are accessible to all, regardless of geographical location.[51]

Geopolitics of inclusivity and transparency

The geopolitics of global health have been a major determinant of whether people, nations, and continents have access to vaccines, patent waivers, and knowledge technology.[53-55] As Long COVID patients are found across all countries, there is an urgent need for the discussions on diagnostic criteria, clinical trials, and treatment to be all-inclusive. To forestall the COVID-19 pandemic vaccine-hoarding phenomenon, developing countries should have their voices heard in the global conversation surrounding COVID-19 and be allowed to contribute their wealth of research and data. This will help to improve the accuracy and usefulness of models generated. Moreover, the developing world should not be treated as a monolith by wealthier nations. Surveys, clinical trials, and data-capturing processes should consider developing countries' unique cultural, geographical, and political characteristics and how these might influence research at a micro and macro level.

National and regional data regulatory frameworks

Ideally, national and regional regulatory frameworks should foster ethical data sharing and multinational collaboration. This is not usually the case, as data regulatory institutions and bodies enforce data protection laws which do not encourage data sharing. Concerning health-related issues, regulatory bodies are even stricter.[56] There are technologies that allow for privacy-preserving sharing of data, which also protect to a large-extent the reverse engineering of such data sets to identify individuals or groups of individuals.[57] Removing these barriers to privacy-preserving data-sharing would greatly encourage collaborative research for Long COVID.[58-60]

Road map for the future: Health-related data sharing

The road map for health-related data sharing includes building health data science capacity, paradigm change in infrastructure, interoperability, and new governance and data ownership models.

Health data science capacity building

To improve health-related data sharing among researchers and institutions health, the data science capacity of these researchers and institutions would need to be expanded.[61] With health-related researchers and experts armed with the knowledge and importance of health data science, the culture of ethical data sharing and health data science would be embedded in the policies, operations, and processes such as clinical trials. To achieve this, the two other critical domains (i.e. computer science and mathematics/statistics) would need to be tailored to health-related professions in the health sciences curriculum globally. Moreover, all stakeholders, like health science educational standardisation institutions, would need to be engaged to see the importance of data science in uncovering insights into health-related diseases such as Long COVID and yet-to-happen pandemics. Additionally, health and medical practitioners should be encouraged (and mandated where/when necessary) to attend health data science trainings.[60,62-65] Consequently, in the long term, data sharing and data science knowledge and skill sets would be imbibed in the medical and health sciences.

Paradigm change in infrastructure

The global health industry sits on a vast amount of data such as electronic health data and records, genomic sequences, clinical trials, health surveys, and disease registries. To foster data sharing of health-related data sets, the mode and means of data set storage needs to be redesigned. Owing to the peculiarities of health-related data sets (such as privacy, security, and size), new technologies[66] including blockchain, cloud storage, and quantum computing, should be embedded in the healthcare systems of the future. Blockchain and quantum computing can both help protect data and increase privacy and security. Blockchain technology is used to create an immutable, distributed ledger system that is secure and transparent (where transparency refers to the existence of the blockchain, while the actual data may be kept private). This system can help protect data from tampering and unauthorised access, while enabling users to control who has access to their data.[67-69] Blockchain technology therefore enables privacy and security critical to health-related data sets. In addition, some aspects of quantum computing (specifically quantum information processing) can be used to secure data in two combinable ways. First, quantum key distribution (commonly known as QKD) uses quantum mechanics to create a secure and tamper-proof channel for data transmission, which is more secure than traditional encryption methods. Second, quantum-resilient cryptography (QRC, but also sometimes referred to as post-quantum cryptography, PQC) uses recently standardised algorithms – running on normal computers – that are practically impossible to crack, even with the help of the most powerful of computers.[67,70,71] For instance, blockchain technology would enable privacy and security critical to health-related data sets.[72,73] These technologies combined will play significantly critical roles in promoting data sharing and collaborative health-related research in future.

Soon, health-related research hubs and systems may outsource their data operations and management to technology-based corporations. This would allow health-related institutions and research hubs to leverage the computational and AI efficiencies of these specialised technology-savvy companies. To this end, the concept of health-data science/analytics as a service would dominate the discussions in the health industry.

Interoperability

Interoperability of data would play a critical role in sharing of health-related data. Interoperability, in this case, is the ability of stakeholders such as users, patients, their families, medical experts, and researchers to efficiently, securely, and timeously exchange health-related data with ease.[74] Technologies such as blockchain enable interoperability that secures and allows for timeous exchange of health-related data. These technologies achieve interoperability through six main characteristics as depicted in Figure 5, which illustrates the factors that contribute to the realisation of health data interoperability. Interoperability is one of the main enablers of real-time data sharing of health information and data sets. Additionally, clinical trials and treatment of Long COVID will benefit from the transparency fostered by the interoperability of data sharing. There is no doubt that interoperability will promote a nationwide, inter-nation, and global-wide data-sharing culture.[75]

New governance and data ownership models

The discussion around data ownership determines the ease with which, how, where, and what type of data are captured, stored, and shared. Currently, health institutions and research hubs believe that their own patients' data are in their custody.[76] On the contrary, patients are increasingly aware of their data rights and, consequently, demand consent before their data are used. New governance and owner models would greatly forestall legal bottlenecks to efficient data sharing that may

arise from data ownership. Data governance and ownership models (such as data sharing pools, data cooperatives, public data trusts, and personal data sovereignty) as a future road map for health data sharing have been discussed in the literature.[77-79] Fulfilling data regulations such as POPIA and GDPR, although onerous, require consent from patients and should be integrated in both existing and future systems.[80]

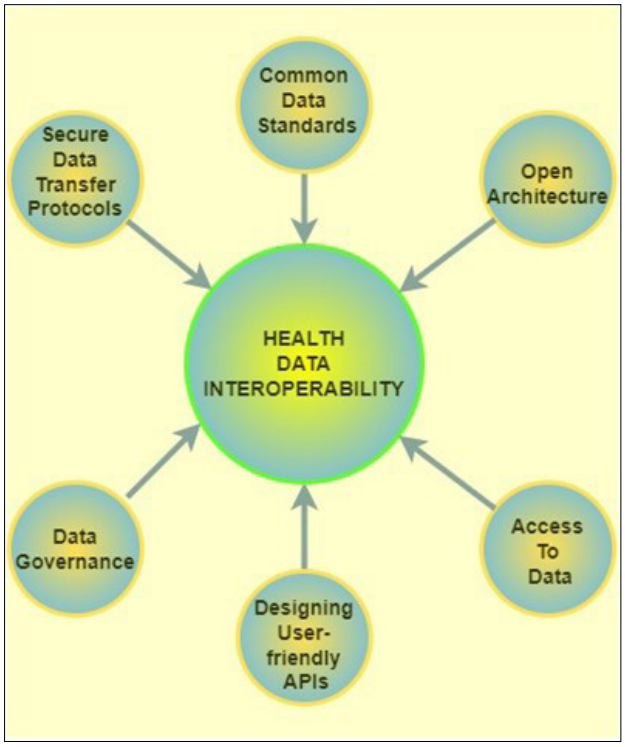

Figure 5: Key steps that contribute to the realisation of interoperability of health data.

Data sharing templates and agreements

Sharing medical and health-related data raises concerns about the ethical use of data sets. To forestall future legal issues and ensure the ethical use of data sets, a data sharing template and agreement should be used by the data custodians. Data sharing templates and agreements may help assuage the fears of data custodians who are not ready or willing to make their data sets open to the public by rethinking 'on reasonable request'. The data sharing template and agreements will provide a guide from scientific discovery to clinical application of our current knowledge about the pathogenesis of Long COVID. A readiness checklist including the requirement of a data sharing agreement for implementation of genomic medicine programmes involving return of research results at the intersection of research and service delivery is given by Jongeneel et al.[81] Although data sharing templates and agreements are not new in medical research, Long COVID research is relatively in its early stages. Data sharing templates and agreements designed for COVID, if invested in, would significantly help to foster data sharing among Long COVID researchers.

Clinical policymakers as gatekeepers

Data sharing should create value that benefits adopters[82], i.e. generators of the data. Clear benefits create incentives to move from few adopters to mainstream practices. We posit that clinical policymakers are the gatekeepers of information flow from clinical research to best practice policy in a patient setting. Given the incentive for clinical researchers to impact on patient treatment practices, clinical policymakers are in a position to create incentives for data sharing. Clinical policymakers may provide incentives within the requirements for successful research funding and grants to support clinical research, through recognition, and through the promotion of their research at the institutional or national level, as well as through academic recognition in the form of awards and publications. Additionally, clinical researchers may be incentivised by professional satisfaction when they see their research directly impacting patient care and clinical practice. Moreover, there are inherent advantages of data sharing to both clinical researchers and policymakers such as enhancing transparency and public trust. Clinical policymakers have the opportunity to increase diffusion of data-sharing practices among data-generating researchers by ensuring best practices with respect to data sharing are followed during the clinical research that results in patient treatment policies. These best practices can be ensured by: establishing clear policies and procedures for data sharing that outline the expectations; providing training and education for clinical researchers on data sharing best practices; monitoring and auditing (including periodic reviews of) data sharing activities; encouraging collaboration among clinical researchers; and utilising data sharing platforms and services that provide secure and efficient ways to store and share data. This is analogous to mortgage lenders being the gatekeepers to encourage uptake of energy-efficient homes.[83]

Conclusion

Despite millions of people across the world having been diagnosed with Long COVID, and the detrimental impact on the health and wealth of individuals and economies, there have been few global concerted efforts to encourage data sharing and data science to uncover insights into this disease. In this paper, we examined the benefits of data-driven frameworks, in particular open big data sets, for Long COVID. Moreover, a review of the research data set and the current state of data sharing was carried out on Long COVID research in Africa and the world in general. To encourage data sharing and collaborative Long COVID research, we examined potential challenges and also discussed the road map for the future of health data sharing.

Competing interests

We have no competing interests to declare.

Authors' contributions

S.O.O.: Wrote the paper, edited the paper, corresponding author, study leader. L.R.W.: Contributed to the scientific context. K.R.: Contributed to the scientific context, writing and editing of the paper. B.W.W., M.J.K., D.B.K. and E.P.: Contributed to the scientific context and edited the paper.

References

1. Tran VT, Porcher R, Pane I, Ravaud P. Course of post COVID-19 disease symptoms over time in the ComPaRe Long COVID prospective e-cohort. Nat Commun. 2022;13(1), Art. #1812. https://doi.org/10.1038/s41467-022-29513-z
2. Sugiyama A, Miwata K, Kitahara Y, Okimoto M, Abe K, Ouoba S, et al. Long COVID occurrence in COVID-19 survivors. Sci Rep. 2022;12(1), Art. #6039. https://doi.org/10.1038/s41598-022-10051-z
3. Xie Y, Xu E, Bowe B, Al-Aly Z. Long-term cardiovascular outcomes of COVID-19. Nat Med. 2022;28(3):583–590. https://doi.org/10.1038/s41591-022-01689-3
4. Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV, Group WC. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect Dis. 2022;22(4):102–107. https://doi.org/10.1016/S1473-3099(21)00703-9
5. Diaz JV, Herridge M, Bertagnolio S, Davis HE, Dua T, Kaushic C, et al. Towards a universal understanding of post COVID-19 condition. Bull World Health Organ. 2021;99(12):901–903. https://doi.org/10.2471/BLT.21.286249
6. Pretorius E, Venter C, Laubscher GJ, Kotze MJ, Oladejo SO, Watson LR, et al. Prevalence of symptoms, comorbidities, fibrin amyloid microclots and platelet pathology in individuals with Long COVID/Post-Acute Sequelae of COVID-19 (PASC). Cardiovasc Diabetol. 2022;21(1), Art. #148. https://doi.org/10.1186/s12933-022-01579-5
7. Lopez-Leon S, Wegman-Ostrosky T, Perelman C, Sepulveda R, Rebolledo PA, Cuapio A, et al. More than 50 long-term effects of COVID-19: A systematic review and meta-analysis. Sci Rep. 2021;11(1), Art. #16144. https://doi.org/10.1038/s41598-021-95565-8
8. Kell DB, Laubscher GJ, Pretorius E. A central role for amyloid fibrin microclots in long COVID/PASC: Origins and therapeutic implications. Biochem J. 2022;479(4):537–559. https://doi.org/10.1042/BCJ20220016

9. Rubin R. As their numbers grow, COVID-19 "long haulers" stump experts. JAMA. 2020;324(14):1381–1383. https://doi.org/10.1001/jama.2020.17709

10. Marshall M. The lasting misery of coronavirus long-haulers. Nature. 2020;585(7825):339–342. https://doi.org/10.1038/d41586-020-02598-6

11. Rando HM, Bennett TD, Byrd JB, Bramante C, Callahan TJ, Chute CG, et al. Challenges in defining Long COVID: Striking differences across literature, Electronic Health Records, and patient-reported information. MedRxiv. 2021. https://doi.org/10.1101/2021.03.20.21253896

12. Willyard C. Could tiny blood clots cause long COVID's puzzling symptoms? Nature. 2022;608:662–664. https://doi.org/10.1038/d41586-022-02286-7

13. Patrucco AS, Trabucchi D, Frattini F, Lynch J. The impact of Covid-19 on innovation policies promoting Open Innovation. R D Manag. 2022;52(2):273–293. https://doi.org/10.1111/radm.12495

14. Galaitsi SE, Cegan JC, Volk K, Joyner M, Trump BD, Linkov I. The challenges of data usage for the United States' COVID-19 response. Int J Inform Manage. 2021;59, Art. # 102352. https://doi.org/10.1016/j.ijinfomgt.2021.102352

15. Sheng J, Amankwah-Amoah J, Khan Z, Wang X. COVID-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions. Br J Manag. 2021;32(4):1164–1183. https://doi.org/10.1111/1467-8551.12441

16. Aiyegbusi OL, Hughes SE, Turner G, Rivera SC, McMullan C, Chandan JS, et al. Symptoms, complications and management of long COVID: A review. J R Soc Med. 2021;114(9):428–442. https://doi.org/10.1177/01410768211032850

17. Wang L, Foer D, MacPhaul E, Lo YC, Bates DW, Zhou L. PASCLex: A comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. J Biomed Inform. 2022;125, Art.# 103951. https://doi.org/10.1016/j.jbi.2021.103951

18. Sarri G, Bennett D, Debray T, Deruaz-Luyet A, Soriano Gabarró M, Largent JA, et al. ISPE-endorsed guidance in using electronic health records for comparative effectiveness research in COVID-19: Opportunities and trade-offs. Clin Pharmacol Ther. 2022;112(5):990–999. https://doi.org/10.1002/cpt.2560

19. Gaber T. Assessment and management of post-COVID fatigue. Prog Neurol Psychiatry. 2021;25(1):36–39. https://doi.org/10.1002/pnp.698

20. Rahman MA, Zaman N, Asyhari AT, Al-Turjman F, Bhuiyan MZ, Zolkipli MF. Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices. Sustain Cities Soc. 2020;62, Art. #102372. https://doi.org/10.1016/j.scs.2020.102372

21. Ros F, Kush R, Friedman C, Gil Zorzo E, Rivero Corte P, Rubin JC, et al. Addressing the Covid-19 pandemic and future public health challenges through global collaboration and a data-driven systems approach. Learn Health Syst. 2021;5(1), e10253. https://doi.org/10.1002/lrh2.10253

22. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Identifying who has long COVID in the USA: A machine learning approach using N3C data. Lancet Digital Health. 2022;4(7):532–541. https://doi.org/10.1016/S2589-7500(22)00048-6

23. Vinod DN, Prabaharan SR. Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19. Chaos, Solitons Fractals. 2020;140, Art. #110182. https://doi.org/10.1016/j.chaos.2020.110182

24. Harrison TM, Pardo TA. Data, politics and public health: COVID-19 data-driven decision making in public discourse. Digital Gov Res Pract. 2020;2(1), Art. #11. https://doi.org/10.1145/3428123

25. Crook H, Raza S, Nowell J, Young M, Edison P. Long Covid-mechanisms, risk factors, and management. BMJ. 2021;374, Art. #1648. https://doi.org/10.1136/bmj.n1648

26. Banks GC, Field JG, Oswald FL, O'Boyle EH, Landis RS, Rupp DE, et al. Answers to 18 questions about open science practices. J Bus Psychol. 2019;3(4):257–270. https://doi.org/10.1007/s10869-018-9547-8

27. Bloemraad I, Menjívar C. Precarious times, professional tensions: The ethics of migration research and the drive for scientific accountability. Int Migr Rev. 2022;56(1):4–32. https://doi.org/10.1177/01979183211014455

28. Frazer JS, Shard A, Herdman J. Involvement of the open-source community in combating the worldwide COVID-19 pandemic: A review. J Med Eng Technol. 2020;44(4):169–176. https://doi.org/10.1080/03091902.2020.1757772

29. Davis HE, Assaf GS, McCorkell L, Wei H, Low RJ, Re'em Y, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. EClinicalMedicine. 2021;38, Art. #101019. https://doi.org/10.1016/j.eclinm.2021.101019

30. Pretorius E, Vlok M, Venter C, Bezuidenhout JA, Laubscher GJ, Steenkamp J, et al. Persistent clotting protein pathology in Long COVID/Post-Acute Sequelae of COVID-19 (PASC) is accompanied by increased levels of antiplasmin. Cardiovasc Diabetol. 2021;20(1), Art. #172. https://doi.org/10.1186/s12933-021-01359-7

31. Ladds E, Rushforth A, Wieringa S, Taylor S, Rayner C, Husain L, et al. Persistent symptoms after Covid-19: A qualitative study of 114 "long Covid" patients and draft quality principles for services. BMC Health Serv Res. 2020;20(1), Art. #1144. https://doi.org/10.1186/s12913-020-06001-y

32. UK Office for National Statistics. COVID-19 Schools Infection Survey, England: Long COVID and mental health [data set on the Internet]. c2022 [cited 2022 Jul 20]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/covid19schoolsinfectionsurveyquestionnairedataengland

33. Sugiyama A, Miwata K, Kitahara Y, Okimoto M, Abe K, Ouoba S, et al. Long COVID occurrence in COVID-19 survivors. Sci Rep. 2022;12(1), Art. #6039. https://doi.org/10.1038/s41598-022-10051-z

34. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of long COVID. Nat Med. 2021;27(4):626–631. https://doi.org/10.1038/s41591-021-01292-y

35. Hughes SE, Haroon S, Subramanian A, McMullan C, Aiyegbusi OL, Turner GM, et al. Development and validation of the symptom burden questionnaire for long Covid (SBQ-LC): Rasch analysis. BMJ. 2022;377, e070230. https://doi.org/10.1136/bmj-2022-070230

36. Dryden M, Mudara C, Vika C, Blumberg L, Mayet N, Cohen C, et al. Post COVID-19 condition in South Africa: 3-month follow-up after hospitalisation with SARS-CoV-2. medRxiv. 2022;1–22. https://doi.org/10.1101/2022.03.06.22270594

37. Kuodi P. Long Covid Data Set. Harvard Dataverse, V2. c2022 [cited 2022 Aug 15]. https://doi.org/10.7910/DVN/N5I10C

38. UK Office for National Statistics. Self-reported Long COVID after infection with Omicron variant in the UK [data set on the Internet]. c2022 [cited 2022 Jul 20]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/selfreportedlongcovidafterinfectionwiththeomicronvariantintheuk

39. UK Office for National Statistics. Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 7 July 2022 [document on the Internet]. c2022 [cited 2022 July 20]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/7july2022#:~:text=An%20estimated%202.0%20million%20people,2022%20(see%20Figure%201)

40. Humanitarian Data Exchange (Humdata-UNOCHA). Kenya, Malawi, Long Covid-19 effects survey dataset [data set on the Internet]. c2020 [cited 2022 Jul 20]. Available from: https://data.humdata.org/dataset/long-covid researchagenda

41. American Academy of Physical Medicine and Rehabilitation (AAPM&R). PASC Dashboard [webpage on the Internet]. c2022 [cited 2022 Jul 20]. Available from: https://pascdashboard.aapmr.org/

42. Bierer BE, White SA, Barnes JM, Gelinas L. Ethical challenges in clinical research during the COVID-19 pandemic. J Bioeth Inq. 2020;17(4):717–722. https://doi.org/10.1007/s11673-020-10045-4

43. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach. 2018;28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

44. Mallapaty S. Omicron-variant border bans ignore the evidence, say scientists. Nature. 2021;600:199. https://doi.org/10.1038/d41586-021-03608-x

45. Schermerhorn J, Case A, Graeden E, Kerr J, Moore M, Robinson-Marshall S, et al. Fifteen days in December: Capture and analysis of Omicron-related travel restrictions. BMJ Global Health. 2022;7(3), e008642. https://doi.org/10.1136/bmjgh-2022-008642

46. Singhal T. The emergence of Omicron: Challenging times are here again! Indian J Paediatr. 2022;89:490–496.

47. Mendelson M, Venter F, Moshabela M, Gray G, Blumberg L, de Oliveira T, et al. The political theatre of the UK's travel ban on South Africa. The Lancet. 2021;398(10318):2211–2213. https://doi.org/10.1016/S0140-6736(21)02752-5

48. Huang Y. The SARS epidemic and its aftermath in China: A political perspective. In: Institute of Medicine (US) Forum on Microbial Threats; Knobler S, Mahmoud A, Lemon S, et al., editors. Learning from SARS: Preparing for the next disease outbreak. Washington DC: US National Academies Press; 2004. p. 116–136.

49. Lee JJ, Haupt JP. Scientific globalism during a global crisis: Research collaboration and open access publications on COVID-19. High Educ. 2021;81:949–966. https://doi.org/10.1007/s10734-020-00589-0

50. Homolak J, Kodvanj I, Virag D. Preliminary analysis of COVID-19 academic information patterns: A call for open science in the times of closed borders. Scientometrics. 2020;124:2687–2701. https://doi.org/10.1007/s11192-020-03587-2

51. Jamali D, Barkemeyer R, Leigh J, Samara G. Open access, open science, and coronavirus: Mega trends with historical proportions. Bus Ethics A Eur Rev. 2020;29(3):419–421. https://doi.org/10.1111/beer.12289

52. Besançon L, Peiffer-Smadja N, Segalas C, Jiang H, Masuzzo P, Smout C, et al. Open science saves lives: Lessons from the COVID-19 pandemic. BMC Med Res Methodol. 2021;21(1), Art. #117. https://doi.org/10.1186/s12874-021-01304-y

53. Cole J, Dodds K. Unhealthy geopolitics: Can the response to COVID-19 reform climate change policy? Bull World Health Organ. 2021;99(2):148–154. https://doi.org/10.2471/BLT.20.269068

54. Ndlovu-Gatsheni SJ. Geopolitics of power and knowledge in the COVID-19 pandemic: Decolonial reflections on a global crisis. J Dev Soc. 2020;36(4):366–389. https://doi.org/10.1177/0169796X20963252

55. Sturm T, Mercille J, Albrecht T, Cole J, Dodds K, Longhurst A. Interventions in critical health geopolitics: Borders, rights, and conspiracies in the COVID-19 pandemic. Polit Geogr. 2021;91, Art. #102445. https://doi.org/10.1016/j.polgeo.2021.102445

56. Tacconelli E, Gorska A, Carrara E, Davis RJ, Bonten M, Friedrich AW, et al. Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response. The Lancet Regional Health - Europe. 2022;21, Art. # 100467. https://doi.org/10.1016/j.lanepe.2022.100467

57. Jin H, Luo Y, Li P, Mathew J. A review of secure and privacy-preserving medical data sharing. IEEE Access. 2019;7:61656–61669.

58. Yu K, Tan L, Shang X, Huang J, Srivastava G, Chatterjee P. Efficient and privacy-preserving medical research support platform against COVID-19: A blockchain-based approach. IEEE Consumer Electronics Magazine. 2020;10(2):111–120. https://doi.org/10.1109/MCE.2020.3035520

59. Ha YJ, Lee G, Yoo M, Jung S, Yoo S, Kim J. Feasibility study of multi-site split learning for privacy-preserving medical systems under data imbalance constraints in covid-19, x-ray, and cholesterol dataset. Sci Rep. 2022;12(1), Art. #1534. https://doi.org/10.1038/s41598-022-05615-y

60. Chen Y, Banerjee A. Improving the digital health of the workforce in the COVID-19 context: An opportunity to future-proof medical training. Future Healthc J. 2020;7(3):189–192. https://doi.org/10.7861/fhj.2020-0162

61. Beyene J, Harrar SW, Altaye M, Astatkie T, Awoke T, Shkedy Z, et al. A roadmap for building data science capacity for health discovery and innovation in Africa. Front Public Health. 2021;9. https://doi.org/10.3389/fpubh.2021.710961

62. Schull MJ, Azimaee M, Marra M, Cartagena RG, Vermeulen MJ, Ho M, et al. ICES: Data, discovery, better health. Int J Popul Data Sci. 2019;4(2), Art. #1135. https://doi.org/10.23889/ijpds.v4i2.1135

63. Wang Y, Kung L, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change. 2018;126:3–13. https://doi.org/10.1016/j.techfore.2015.12.019

64. Gutiérrez-Aguado A, Curioso WH, Machicao JC, Eguia H. Strengthening capacities of multidisciplinary professionals to apply data science in public health: Experience of an international graduate diploma program in Peru. Int J Med Inform. 2023;169, Art. #104913. https://doi.org/10.1016/j.ijmedinf.2022.104913

65. Lee O, Campbell T. What science and STEM teachers can learn from COVID-19: Harnessing data science and computer science through the convergence of multiple STEM subjects. J Sci Teach Educ. 2020;31(8):932–944. https://doi.org/10.1080/1046560X.2020.1814980

66. Kondylakis H, Koumakis L, Tsiknakis M, Kiefer S. Personally managed health data: Barriers, approaches and a roadmap for the future. J Biomed Inform. 2020;106, Art. #103440. https://doi.org/10.1016/j.jbi.2020.103440

67. Kaushik K, Kumar A. Demystifying quantum blockchain for healthcare. Secur Priv. 2022, e284. https://doi.org/10.1002/spy2.284

68. Attaran M. Blockchain technology in healthcare: Challenges and opportunities. Int J Healthc Manag. 2022;15(1):70–83. https://doi.org/10.1080/20479700.2020.1843887

69. Khezr S, Moniruzzaman M, Yassine A, Benlamri R. Blockchain technology in healthcare: A comprehensive review and directions for future research. Appl Sci. 2019;9(9):1736. https://doi.org/10.3390/app9091736

70. Gill SS, Kumar A, Singh H, Singh M, Kaur K, Usman M, et al. Quantum computing: A taxonomy, systematic review and future directions. Softw Pract Exper. 2022;52(1):66–114. https://doi.org/10.1002/spe.3039

71. Malviya R, Sundram S. Exploring potential of quantum computing in creating smart healthcare. Open Biol J. 2022;9(1):56–57. https://doi.org/10.2174/1874196702109010056

72. Mustafa M, Alshare M, Bhargava D, Neware R, Singh B, Ngulube P. Perceived security risk based on moderating factors for blockchain technology applications in cloud storage to achieve secure healthcare systems. Comput Math Methods Med. 2022;2022, Art. # 6112815. https://doi.org/10.1155/2022/6112815

73. Angraal S, Krumholz HM, Schulz WL. Blockchain technology: Applications in health care. Circ Cardiovasc Qual Outcomes. 2017;10(9), e003800. https://doi.org/10.1161/CIRCOUTCOMES.117.003800

74. Office of the National Coordinator for Health Information Technology. Connecting health and care for the nation: A shared nationwide interoperability roadmap [document on the Internet]. c2015 [cited 2022 Jul 12]. Available from: https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf

75. Satti FA, Ali T, Hussain J, Khan WA, Khattak AM, Lee S. Ubiquitous Health Profile (UHPr): A big data curation platform for supporting health data interoperability. Computing. 2020;102(11):2409–2444. https://doi.org/10.1007/s00607-020-00837-2

76. Hulsen T. Sharing is caring – data sharing initiatives in healthcare. Int J Environ Res Public Health. 2020;17(9), Art. #3046. https://doi.org/10.3390/ijerph17093046

77. Bak MA, Ploem MC, Tan HL, Blom MT, Willems DL. Towards trust-based governance of health data research. Med Health Care Philos. 2023:1–16. https://doi.org/10.1007/s11019-022-10134-8

78. Micheli M, Ponti M, Craglia M, Berti Suman A. Emerging models of data governance in the age of datafication. Big Data Soc. 2020;7(2). https://doi.org/10.1177/2053951720948087

79. Piasecki J, Cheah PY. Ownership of individual-level health data, data sharing, and data governance. BMC Medical Ethics. 2022;23(1), Art. #104. https://doi.org/10.1186/s12910-022-00848-y

80. Usynin D, Ziller A, Makowski M, Braren R, Rueckert D, Glocker B, et al. Adversarial interference and its mitigations in privacy-preserving collaborative machine learning. Nat Mach Intell. 2021;3(9):749–758.

81. Jongeneel CV, Kotze MJ, Bhaw-Luximon A, Fadlelmola FM, Fakim YJ, Hamdi Y, et al. A view on genomic medicine activities in Africa: Implications for policy. Front Genet. 2022;13. https://doi.org/10.3389/fgene.2022.769919

82. Greenhalgh C, Rogers M. Innovation, intellectual property, and economic growth. Princeton, NJ: Princeton University Press; 2010. https://doi.org/10.1515/9781400832231

83. Sanderford AR, Overstreet GA, Beling PA, Rajaratnam K. Energy-efficient homes and mortgage risk: Crossing the chasm at last?. Environ Syst Decis. 2015;35:157–168.