



Big data and AI in health sciences research in
sub-Saharan Africa

GUEST EDITORS

Keymanthri Moodley 
Department of Medicine,
Stellenbosch University, South Africa

Stuart Rennie 
Department of Social Medicine,
University of North Carolina, Chapel
Hill, USA

EDITOR-IN-CHIEF

Leslie Swartz 
Academy of Science of South Africa

MANAGING EDITOR

Linda Fick 
Academy of Science of South Africa

ONLINE PUBLISHING SYSTEMS ADMINISTRATOR

Nadia Grobler 
Academy of Science of South Africa

MARKETING & COMMUNICATION

Henriette Wagener
Academy of Science of South Africa

ASSOCIATE EDITORS

Priscilla Baker 
Department of Chemistry, University
of the Western Cape, South Africa

Pascal Bessong 
HIV/AIDS & Global Health Research
Programme, University of Venda,
South Africa

Floretta Boonzaier 
Department of Psychology, University
of Cape Town, South Africa


Chrissie Boughey 
Centre for Postgraduate Studies,
Rhodes University, South Africa

Teresa Coutinho 
Department of Microbiology and
Plant Pathology, University of Pretoria,
South Africa

Jemma Finch 
School of Agricultural, Earth and
Environmental Sciences, University
of KwaZulu-Natal, South Africa

Jennifer Fitchett 
School of Geography, Archaeology
and Environmental Studies, University
of the Witwatersrand, South Africa


Michael Inggs 
Department of Electrical Engineering,
University of Cape Town, South Africa

Ebrahim Momoniat 
Department of Mathematics and
Applied Mathematics, University of
Johannesburg, South Africa

Sydney Moyo 
Department of Biology, Rhodes
College, Memphis, TN, USA

ASSOCIATE EDITOR

MENTEES

Nkosinathi Madondo 
Academic Literacy and Language
Unit, Mangosuthu University of
Technology, South Africa

Amanda-Lee Manicum 
Department of Chemistry, Tshwane
University of Technology, South Africa

Adriaan van der Walt 
Department of Geography, University
of the Free State, South Africa



South African Journal of Science

eISSN: 1996-7489

Guest Leader

The many faces of the big data revolution in health for sub-Saharan Africa

Keymanthri Moodley, Stuart Rennie..... 1

Book Review

Biased and sex-aggregated data: The forgotten half of the population

Nezerith Cengiz, Siti M. Kabanda..... 4

Perspectives

Global health and big data: The WHO's artificial intelligence guidance

Kenneth W. Goodman, Sergio G. Litewka, Rohit Malpani, Sameer Pujari, Andreas A. Reis..... 6

Responsible application of artificial intelligence in health care

Adetayo E. Obasa, Andrea C. Palk..... 10

Managing and assembling population-scale data streams, tools and workflows
to plan for future pandemics within the INFORM-Africa Consortium

*Jenicca Poongavanan, Joicymara Xavier, Marcel Dunaiski, Houriiyah Tegally,
Sunday O. Oladejo, Olawole Ayorinde, Eduan Wilkinson, Cheryl Baxter, Tulio de Oliveira*..... 13

The role of an ethics advisory committee in data science research in sub-
Saharan Africa

Sharon Kling, Shenuka Singh, Theresa L. Burgess, Gonasagrie Nair..... 17

Revisiting community engagement methods in the context of data science
research and big data use in South Africa

Gonasagrie Nair, Theresa L. Burgess, Adetayo E. Obasa, Sharon Kling, Shenuka Singh..... 20

Commentaries

Data science research in sub-Saharan Africa: Ethical considerations in
crowdsourcing for community engagement

Suzanne Day, Stuart Rennie..... 23

Setting up data science research in Africa and engagement of stakeholders

*Fati Murtala-Ibrahim, Jibreel Jumare, Manhattan Charurat, Chenfeng Xiong,
Vivek Naranbhai, Patrick Dakum, Shirley Collie, Waasila Jassat, Gambo Aliyu,
Adetifa Ifedayo, Alash'le Abimiku*..... 27

Research Articles

Revealing human mobility trends during the SARS-CoV-2 pandemic in Nigeria
via a data-driven approach

*Weiye Luo, Chenfeng Xiong, Jiajun Wan, Ziteng Feng, Olawole Ayorinde, Natalia
Blanco, Man Charurat, Vivek Naranbhai, Christina Riley, Anna Winters, Fati Murtala-Ibrahim*..... 29

EDITORIAL ADVISORY BOARD

Stephanie Burton 
Professor of Biochemistry and
Professor at Future Africa, University
of Pretoria, South Africa

Felix Dakora 
Department of Chemistry, Tshwane
University of Technology, South Africa

Saul Dubow
Smuts Professor of Commonwealth
History, University of Cambridge, UK

Pumla Gobodo-Madikizela 
Trauma Studies in Historical Trauma
and Transformation, Stellenbosch
University, South Africa

Robert Morrell 
School of Education, University of
Cape Town, South Africa

Catherine Ngila 
Deputy Vice Chancellor – Academic
Affairs, Riara University, Nairobi, Kenya


Lungiswa Nkonki 
Department of Global Health,
Stellenbosch University, South Africa

Daya Reddy 
South African Research Chair –
Computational Mechanics, University
of Cape Town, South Africa

Brigitte Senut
Natural History Museum, Paris, France

Benjamin Smith 
Centre for Rock Art Research and
Management, University of Western
Australia, Perth, Australia

Himla Soodyall 
Academy of Science of South Africa,
South Africa

Lyn Wadley 
School of Geography, Archaeology
and Environmental Studies,
University of the Witwatersrand,
South Africa

Published by
the Academy of Science of
South Africa (www.assaf.org.za)
with financial assistance from the
Department of Science & Innovation.

Design and layout
Elzahn Swarts
E: swarts.elzahn@gmail.com

**Correspondence and
enquiries**
sajs@assaf.org.za

Copyright
All articles are published under a
Creative Commons Attribution Licence.
Copyright is retained by the authors.

Disclaimer
The publisher and editors accept no
responsibility for statements made
by the authors.

Submissions
Submissions should be made at
www.sajs.co.za

Public health research using cell phone derived mobility data in sub-Saharan Africa: Ethical issues <i>Stuart Rennie, Caesar Atuire, Tiwonge Mtande, Walter Jaoko, Sergio Litewka, Eric Juengst, Keymanthri Moodley</i>	38
What constitutes adequate legal protection for the collection, use and sharing of mobility and location data in health care in South Africa? <i>Dirk Brand, Annelize G. Nienaber McKay, Nezerith Cengiz</i>	45
Exploring perspectives of research ethics committee members on the governance of big data in sub-Saharan Africa <i>Nezerith Cengiz, Siti M. Kabanda, Tonya M. Esterhuizen, Keymanthri Moodley</i>	52
Data sharing and data governance in sub-Saharan Africa: Perspectives from researchers and scientists engaged in data-intensive research <i>Siti M. Kabanda, Nezerith Cengiz, Kanshukan Rajaratnam, Bruce W. Watson, Qunita Brown, Tonya M. Esterhuizen, Keymanthri Moodley</i>	61
Data sharing: A Long COVID perspective, challenges, and road map for the future <i>Sunday O. Oladejo, Liam R. Watson, Bruce W. Watson, Kanshukan Rajaratnam, Maritha J. Kotze, Douglas B. Kell, Etheresia Pretorius</i>	73
Regulating scientific and technological uncertainty: The precautionary principle in the context of human genomics and AI <i>Marietjie Botes</i>	81

Cover caption

This special issue on 'Big data and AI in health sciences research in sub-Saharan Africa' introduces a broad range of scientific, ethical, legal and social concerns in the realm of data-intensive research and artificial intelligence (AI) in sub-Saharan Africa. Technology brings enormous benefit, but comes at a price, and with potential harms. Articles in this special issue explore the benefits of data science; the importance of data management, quality and integrity; challenges of engaging communities and stakeholders in data science; ethical and legal issues raised by the gathering and use of mobile phone data; the direction of AI governance in the sub-Saharan African context; and voices from scientists and research ethics committee members.



GUEST LEADER



AUTHORS:

Keymanthri Moodley^{1,2}

Stuart Rennie^{2,3}

AFFILIATIONS:

¹Centre for Medical Ethics and Law, Department of Medicine, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

²Department of Social Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

³UNC Center for Bioethics, Department of Social Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

HOW TO CITE:

Moodley K, Rennie S. The many faces of the big data revolution in health for sub-Saharan Africa. *S Afr J Sci.* 2023;119(5/6), Art. #16158. <https://doi.org/10.17159/sajs.2023/16158>

The many faces of the big data revolution in health for sub-Saharan Africa

This special issue of the *South African Journal of Science* on 'Big data and AI in health sciences research in sub-Saharan Africa' comes from within a large-scale initiative, sponsored by the US National Institutes of Health, to promote research use of 'big data' for health promotion in Africa. As stated on its website (<https://dsi-africa.org>), the Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative aims to leverage data science technologies to transform biomedical and behavioural research and develop solutions that would lead to improved health for individuals and populations. Started in 2021, DS-I Africa has the ambitious goals of creating pan-African scientific networks; developing data science centres of excellence; creating new data collection and analytic systems, applications and tools; facilitating data resource access to the global scientific community; and advancing policies in Africa related to ethical issues raised by data science. A notable structural feature of DS-I Africa is the intentional pairing of specific scientific projects (or 'data hubs') with projects focusing on the ethical, legal and social implications (or ELSI) of data science. While this embedding of ELSI projects within large scientific initiatives in Africa is by no means new – it was also a feature of the H3Africa initiative (<https://h3africa.org>) – it does raise some complex questions about the relationships between social science, ethics, law and the scientific pursuit of knowledge through digital technologies in the context of global, regional and domestic inequities.

Africa is, albeit unevenly in some regions, undergoing an accelerated process of data digitisation. Increased access to and use of the Internet, personal computers and mobile devices in Africa, as well as advances in data storage and transfer capacity, means that individuals, communities and environments are becoming more 'visible' to researchers, and with this new visibility comes the potential for improved understanding and more effective health interventions. In principle, this digital (r)evolution should be warmly welcomed by adherents to evidence-based medicine and public health. For decades, there have been complaints about a 'data vacuum' in Africa, which has hampered efforts to provide effective clinical care, conduct rigorous scientific research, strengthen fragile health systems and tackle emerging public health threats. The pendulum, it seems, is starting to move in the opposite direction, with massive volumes of health-related data in sub-Saharan Africa being collected, analysed, stored, shared and utilised by numerous stakeholders. But while scarcity of data constituted a problem, so too does an abundance.

Whether having an abundance of data (and tools that make use of it) is a cause for celebration depends on a number of conditions, including how the data were gathered, how they are shared, who stands to benefit from the data, who may be burdened by the data, and in general how the data are likely to impact the health and well-being of populations in need. As the old saying goes, 'bigger' is not necessarily 'better'. At the same time that the use of 'big data' is being promoted in Africa, warnings can be heard coming from the industrialised North about the downsides of digital technologies. In March of this year, more than a thousand technology leaders wrote an *open letter* urging artificial intelligence (AI) labs to pause development of the most sophisticated systems, because they present "profound risks to society and humanity". Words of caution and calls for reflection about the use of digital technologies are clearly nothing new. 'Critical data studies' is a field devoted to the economic, political, ethical and legal issues concerning (big) data, including questions about social justice.¹ However, a case can be made that Africa finds itself at a moment of particular vulnerability in this context. For one thing, critical data studies have been disproportionately focused on concerns in high-income countries; African critical data scholarship is relatively nascent. Secondly, public awareness in Africa about data science and potential concerns associated with it appears to be very low. While this is an area for empirical research, citizens in high-income countries (with longer experience with digital technology and critical discourses surrounding it) may have a stronger awareness that what they do on the Internet or with their phones – or in interactions with their medical provider – is being collected/shared for purposes largely beyond their knowledge or control. Thirdly, the generation of voluminous data about Africa and Africans cannot be disentangled from history, and especially colonial history. Africans live with the consequences of the plunder of their natural resources that started during the colonial era. When data are described as the 'new gold' or the 'new oil', worries about exploitation naturally arise. Even the language of 'data sharing' in this context may raise some skepticism: what does 'sharing' involve? This means that projects in large (and externally funded) data science initiatives such as DS-I Africa may have to work to earn community trust, no matter how well-intentioned and scientifically rigorous their studies are.

This special issue presents work from authors involved in the DS-I Africa initiative. More specifically, the authors are drawn from two DS-I Africa projects that have been paired with one other: Role of Data Streams in Informing Infection Dynamics in Africa (INFORM-Africa) and Research for Ethical Data Science in Southern Africa (REDSSA). The overarching goal of INFORM-Africa is to make effective use of big data to address pressing public health needs (including COVID-19 and HIV) as well as to develop population-scale data streams (from public and private sources) to support future pandemic preparedness. Focusing on Nigeria and South Africa, the project aims to develop geospatial tools for the purpose of pandemic surveillance by governments, support data science pilot projects, and work with policymakers to promote open access to the project's high-quality data and tools. As an ELSI project, REDSSA has the overall aims of producing new knowledge about the ethical, legal and social implications of conducting data science, using empirical research and scholarship to help develop evidence-based and context-specific guidance for data science initiatives, and to contribute to the strengthening of the responsible conduct of data science in sub-Saharan Africa.

For all involved, the DS-I Africa initiative is a journey into largely uncharted territory. Even if the urgency of the COVID-19 pandemic recedes, the use of data science for health promotion remains highly relevant for Africa, given its many other pressing public health challenges and the growing threats posed by climate change. The data

tools developed may come to play roles different from their original purposes. The social, ethical and legal implications of data science, and the changes it will bring about in Africa, will also likely evolve and only become clearer as time goes on.

In this sense, this special issue is a snapshot of perspectives and findings that offer some glimpses into the future. A number of common themes in the issue are discernible: an indication of the potential benefits of data science; the importance of data management, quality and integrity; challenges of engaging communities and stakeholders in data science; ethical and legal issues raised by the gathering and use of mobile phone data; the direction of AI governance in the African context; and voices from scientists and research ethics committee members. A brief sample of these themes, with reference to the authors, is presented below.

For those of us who work in ELSI projects, challenges raised by new technologies can sometimes obscure appreciation of their potential benefits. It is therefore important to be reminded of what (social) good new approaches could possibly do. The Research Article by Oladejo et al. focuses on a health issue of global importance – Long COVID – which will occupy clinicians and public health professionals for years to come. Medical information on Long COVID collected during the pandemic has been fragmented; centralising, sharing and analysing data could reveal patterns that could improve our understanding of this condition and open up new directions for scientific inquiry. Similarly, the research findings reported by Luo et al. reveal that important public health information can be learned by collecting and analysing mobile phone data, particularly in the domain of public health policy. Improving techniques to quantify human mobility patterns and relating these patterns to other data in order to answer specific public health related questions, means that the potential health benefits of this research approach for Africa may extend far beyond the context of COVID-19.

However, that data science activities will be beneficial is not a given. As with any scientific enterprise, much depends on how the research is designed, how and what data are collected, and especially how the collected data are processed and managed. A central part of INFORM-Africa's mission is the establishment and maintenance of its Data Management and Analysis Core (DMAC) and its Next Generation Sequencing Core (NGS). In this issue, Poongavanan et al. provide a window into the inner workings of INFORM-Africa's data infrastructure, which could potentially serve as a model for health organisations in sub-Saharan Africa wanting to enter into the data science space. The importance of maintaining high data quality, as well as being reflective about how data are 'constructed', is also underlined in the Book Review offered by Cengiz and Kabanda in this special issue. In their reading of Caroline Perez's *Invisible Women: Exposing Data Bias in a World Designed for Men*, they note how gender bias can permeate the construction of data at all stages of the process: from lack of data about women in sources used, to bias towards men in algorithms, to the baking of gender biases into AI programs. There is a real threat of women becoming (more) 'invisible' in sub-Saharan Africa by creating data science tools and outputs that magnify existing gender inequities. This shows that data management is not just about having accurate or reliable data, but also data that do not perpetuate social harm through bias.

A number of the contributions in this special issue touch upon, or are devoted to, issues related to mobile phone data. There are some good reasons for this. Mobile phone use in sub-Saharan Africa has increased dramatically over the last decade, and particularly as smart phones have become more common, human activities related to mobile phone use (such as apps) are generating massive amounts of data, in real time. As noted above in reference to the study by Luo et al., such data can be highly valuable for public health researchers, to help tackle all sorts of health research questions. However, as Brand et al. note in their Research Article, mobile phone data also raises a number of pressing legal questions about privacy, consent, liability and accountability. To some extent, similar legal questions have been raised (and to some extent, addressed) in high-income countries. An important question is how to legally address these emerging concerns when national laws (often legacies from colonial times) are not keeping pace with technological advances. The authors note that the paradigmatic mechanism for protecting individuals in health research – informed consent – falls short

in this context when mobile phone users (and particularly those with low levels of literacy) are typically unaware that their phone data are used for research purposes. The Research Article by Rennie et al. includes this concern about the limits of informed consent, while examining other ethical issues raised by the research use of mobile phone data in the sub-Saharan African context. These issues include concerns about group privacy, function creep, power dynamics among stakeholders and how mobility analyses are 'translated' into health policy by government authorities. As the authors note, if individuals do not provide valid informed consent for researchers to track their phone activities, then community awareness and input will be crucial to maintain public trust in this kind of research.

In the history of HIV research, a well-known slogan in community advocacy was: 'nothing about us without us'. This was a call for robust community engagement in research. When it comes to data science, however, a lot is collected about us – from our mobile phones and many other sources – without us knowing. It is easy to say that engagement and awareness should be increased. In the case of data science, perhaps even more than with HIV clinical trials, the question is how, when the activities and outputs of data science are often highly technical. This is not just a challenge for ordinary citizens, but also for other stakeholders who are not themselves experts in data science. The Commentary by Murtala-Ibrahim et al. offers experiences of INFORM-Africa data science investigators engaging with stakeholders in South Africa and Nigeria. Their account suggests that it is important to include a broad range of stakeholders and involve them in the initial design of projects, even if their understanding of the technical aspects of the projects are a matter of degree. Stakeholders like government agencies, health data custodians (such as clinic managers), community gatekeepers, and leaders in the scientific community have interests in and/or are affected by data science projects, and these relationships are as fundamental to the success of these projects as the technical infrastructure and scientific expertise are. But what about the community at large, i.e. ordinary citizens? The Commentary by Day and Rennie maps out the strengths, limitations and ethical considerations raised by using crowdsourcing to engage communities in data science. The process of creating a contest about data science, encouraging entries from participants, and disseminating contest results can to some extent send a message of awareness about the existence and nature of data science into communities. While crowdsourcing is only one approach towards community engagement, a number of studies have indicated that it can be impactful, and it could be a promising approach in sub-Saharan Africa. The REDSSA project is in fact currently conducting a crowdsourcing project that focuses on how best to engage communities in data science. The Perspective by Nair et al. points out that existing and familiar practices – such as community advisory boards, flexible forms of consent, and research ethics committees – still have important roles to play in the big data era in Africa, although these practices will require some adaptation and need to be conjoined with educational initiatives. In addition, in this special issue, Kling et al. suggest that we can also leverage a less traditional community engagement mechanism, in the form of Ethics Advisory Committees – a structure that complements the work of Research Ethics Committees and Clinical Ethics Committees. Ethical Advisory Committees would comprise diverse members who genuinely represent community interests and concerns and could help steer data science projects in a mutually satisfactory direction. No doubt community engagement in data science will require a multitude of approaches, including innovative ones yet to be conceived.

As mentioned, AI receives substantial attention, both positive and negative. The worldwide rise of ChatGPT has suggested that the gap between AI and human intelligence is rapidly narrowing, and also that the use of this technology could cause a great deal of disruption and harm. The idea that AI needs to be regulated is nothing new, but its regulation within the domain of data science in the sub-Saharan African context to some extent is. As Goodman et al. point out in their Perspective, the World Health Organization (WHO) has invested a concerted effort in organising stakeholder meetings and developing thoughtful guidance on the ethics and governance of AI for health. As far as general ethical principles about AI are concerned, there is no need to reinvent the wheel. The ethical principles endorsed by the WHO are

meant to be applicable anywhere, although their application in different country settings (including incorporation into policy and law) will be the work of governments, programmers, companies, civil society, and inter-governmental organisations. The contributions in this special issue by Botes, and Obasa and Palk, offer some complications and nuances in regard to 'translating' general principles into in-country practices. As Botes points out, the use of AI may give rise to additional risks depending on for what it is used, such as human genomic research. Due to these additional risks, Botes argues for the precautionary principle to be incorporated into South African legislation governing AI, as it can cover a wide range of consequences when the effects of technologies are uncertain. In their account of ethical considerations surrounding AI in the South African context, Obasa and Palk note that the *Protection of Personal Information Act* (POPIA) does not accommodate for the potential for reidentification of individuals when AI-driven algorithms are run in health data repositories. In addition, while WHO guidance rightly advocates for transparency in AI as a general ethical principle, Obasa and Palk point out that certain machine learning programs used in clinical contexts operate as 'black boxes', whose inner processes producing the outcomes may be literally impossible for humans to understand. This raises the question of whether such programs should be used at all, even in a supportive role, in clinical or research contexts.² Clearly there is a lot of future work to be done in AI governance in Africa.

Lastly, social science has much to contribute to our understanding of data science as it is unfolding in Africa. As the Research Article by Kabanda et al. reports, the REDSSA project has conducted a survey with 160 researchers and scientists representing 43 different sub-Saharan African countries to investigate their views on data use, data sharing and data governance. Some of the results speak to the gaps in research infrastructure – a reminder that projects in large-scale initiatives such as DS-I Africa are still

working under conditions of general resource constraint. Finally, Cengiz et al. present REDSSA project survey results from another key stakeholder group, research ethics committee members, which identifies inadequacies in regulations relative to data science and inexperience in dealing with data-intensive research protocols. Clearly, capacities in these areas need to be strengthened – and quickly! – to ensure the responsible conduct of data science in sub-Saharan Africa.

Overall, this special issue introduces a broad range of scientific, ethical, legal and social concerns in the realm of data-intensive research and AI in sub-Saharan Africa. These transdisciplinary challenges were once in their infancy but the exponential voluminous growth in digital technology, the speed of early adoption, and the contentious debates that are emerging make engagement with the digital world a responsibility of African scientists and civil society alike. The widespread production, storage and processing of large volumes of data – the "oxygen on which AI depends"³ – causes collateral environmental damage, using up limited supplies of water and energy and accelerating climate change. Technology brings enormous benefit, but comes at a price, and with potential harms. Responsible governance is required to ensure that the price we pay and the harms sustained do not outweigh the overall scientific benefit to humanity.

References

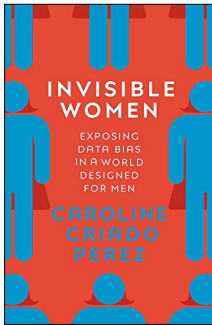
1. Richterich A. The big data agenda: Data ethics and critical data studies. London: University of Westminster Press; 2018. <https://doi.org/10.16997/book14>
2. Duran JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics*. 2021;47:329–335. <https://doi.org/10.1136/medethics-2020-106820>
3. Petrozino C. Who pays for ethical debt in AI? *AI Ethics*. 2021;1:205–208. <https://doi.org/10.1007/s43681-020-00030-3>



Check for updates

BOOK TITLE:

Invisible women: Exposing data bias in a world designed for men



AUTHOR:

Caroline Criado Perez

ISBN:

9781784706289 (paperback, 411 pp)


PUBLISHER:

Chatto & Windus, London, UK; GBP9.99

PUBLISHED:

2019

REVIEWERS:

Nezerith Cengiz¹ 
Siti M. Kabanda¹

AFFILIATION:

¹Centre for Medical Ethics and Law, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

CORRESPONDING EMAIL:

ncengiz@sun.ac.za

HOW TO CITE:

Cengiz N, Kabanda SM. Biased and sex-aggregated data: The forgotten half of the population. *S Afr J Sci.* 2023;119(5/6), Art. #14722. <https://doi.org/10.17159/sajs.2023/14722>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

PUBLISHED:

30 May 2023

Biased and sex-aggregated data: The forgotten half of the population

Big data, data science and artificial intelligence (AI) in research hold enormous potential to improve health and the quality of life for all. However, when harnessing benefit from new digital technologies, a major concern is to minimise harm. Inherent bias in historical data presents a major threat to the veracity of data and the impact thereof on algorithms, machine learning and, ultimately, AI.

In this respect, *Invisible Women: Exposing Data Bias in a World Designed for Men* by Caroline Criado Perez makes a timely contribution by exploring the gender data representation gap in health care, education, economic development, and public policy. Criado Perez is a feminist author, journalist and social activist who has successfully campaigned for empowerment and representation of women in a multitude of different contexts, including the media presence of female experts and social media, in particular.

The preface clarifies an important concept – namely the distinction that must be drawn between sex and gender in biological and social sciences. Criado Perez refers to sex as the “biological characteristics that determine whether an individual is male or female” (XX or XY). By ‘gender’ she refers to the social framing of the biological distinction or the way in which women are treated based on perceptions of them being female. In her book, the term “gender data gap” is used throughout because Criado Perez says “sex is not the reason women are excluded from data. Gender is.”

Although many women are aware of the gender inequality and bias they face on a daily basis, there are many spheres of a woman’s life that are negatively affected by exclusionary practices which may be less obvious. These range from car design (seatbelts, headrests and airbags), travel data, politics, sanitation and employment conditions, to clinical trials and health care that are all geared towards serving men, first and foremost. In some of these aspects, the bias goes beyond discrimination and places women’s lives at risk, such as where road safety and health are concerned.

Invisible Women is structured into six parts with each focusing on various aspects of the many daily struggles that women face simply because the data fail to account for them and exclude them from overall planning and decision-making. The author justifies her arguments with extensive research, statistics, and case studies, and provides vast endnotes for further reading. Each intriguing chapter of this book appears to arrive at the same inevitable conclusion: the gender data gap is “both a cause and a consequence of the type of unthinking that conceives of humanity as almost exclusively male”. Disregarding gender data bias fosters the placement of women as subordinate in our society, where what gets measured, weighed, and made statistically visible is prejudicially determined by the stereotypes that portray distinct social roles and the related spheres of activity for men and women. Generating high-quality data relies on eliminating gender bias at all phases. This in turn requires cognisance of data biases in order to take preventative measures and to make better decisions using data.

Criado Perez essentially provides all the gender data gaps that have occurred through her compilation of studies conducted, statistics reported, and data collected centuries ago that have been used to inform resource allocation, decision-making and policy development. The biggest societal issue seems to be that there is a collective lack of understanding amongst both men and women on the potential for bias in data that are collected. Discriminatory practices raised in the book are not intentional – but because things have been done a certain way for so long that people fail to identify the bias therein. *Invisible Women* unpacks unconscious gender bias and the unconscious way in which things have always been done, which cannot simply be removed mathematically or made ‘unbiased’ on a data level.

Chapters 10 and 11 will be of particular interest to basic scientists, researchers and health professionals. Here the bias inherent in medical research starts with preclinical research, specifically in animal studies in which female animals are not included in investigations on diseases that affect female humans predominantly. Criado Perez quotes Yoon et al.’s¹ review of high-impact surgical journals which found that up to 22% of articles did not specify the sex of the animal studied and, when reported, 80% of studies used only male animals.

It is astonishing that women are not equally included in clinical research and that most male-biased research findings and conclusions are generalised to represent the whole population. Non-representative data negatively affect the quality and validity of results and inferences. Criado Perez reveals that the same applies to drug trials investigating drug responses: results from male participants have been unconsciously considered valid for female individuals because of the assumption that sex did not affect outcomes, yet men and women are known to manifest different symptoms and reactions to various treatments. The chapters ‘The Drugs Don’t Work’ and ‘Yentl Syndrome’ emphasise that these issues can prove fatal for women who are misdiagnosed or medically mistreated unless their clinical indicators or illnesses conform to those of men. Criado Perez highlights that 80% of prescription drugs withdrawn from the US market at the beginning of the 21st century were shown to cause higher adverse drug reactions in women.

Criado Perez shows that the exclusion of women could only have occurred in a culture that conceives men as the default human and women as a niche aberration. Although true, it is also important to consider the role that different cultural, religious, and geographical contexts play.

In the chapter ‘A Cost-less Resource to Exploit’, the author writes: “...the unpaid work that women do isn’t simply a matter of ‘choice’. It is built into the system we have created...”. This statement is potentially correct when

© 2023. The Author(s). Published under a Creative Commons Attribution Licence.



considering patriarchal cultural societies in which women are subjugated and seen as subordinates or second-class citizens with no autonomy to make judgements or decisions. However, some of the 'unpaid work' done by women is done by choice to maintain a healthy family structure. Not all women see 'unpaid work' as burdensome.

Contexts like developing countries in Africa have not been explored in as much depth as settings in the Global North, leading to some claims made on flawed inferences from uncontextualised aggregated data. Readers of *Invisible Women* would have benefitted from inclusion of more cases or examples from Africa to better balance different perspectives. It would also have been valuable to present more optimism about women rather than repeatedly depicting them as victims.

The afterword of the book summarises its key message well:

The solution to the sex and gender data gap is clear: we have to close the female representation gap. When women are involved in decision-making... women do not get forgotten... This is to the benefit of women everywhere... and it is often to the benefit of humanity as a whole.

This book is a valuable resource for both men and women in the current century and may stimulate ideas to close the gender representation gap. Unbiased and sex-disaggregated data collection are essential to guide problem-solving and improve decision-making that impacts populations at large. Most importantly, evidence-based medicine depends strongly on accurate, objective, high-quality data. Gender and ethnic biases seriously undermine this approach and erode trust in science and the health profession. In the era of data and AI-driven technology, eliminating bias is an ethical imperative. This book is essential reading for students, researchers and professionals, especially those working in data science as well as those in the biological and health sciences.

Acknowledgement

We thank Dr Aneeka Domingo for perspectives in earlier stages of reviewing this book.

Reference

1. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. *Surgery*. 2014;156(3):508–516. <https://doi.org/10.1016/j.surg.2014.07.001>



Check for updates

AUTHORS:

Kenneth W. Goodman¹
Sergio G. Litewka¹
Rohit Malpani²
Sameer Pujari²
Andreas A. Reis²

AFFILIATIONS:

¹Institute for Bioethics and Health Policy, Miller School of Medicine, University of Miami, Miami, Florida, USA
²World Health Organization, Geneva, Switzerland

CORRESPONDENCE TO:

Kenneth Goodman

EMAIL:

KGoodman@med.miami.edu

HOW TO CITE:

Goodman KW, Litewka SG, Malpani R, Pujari S, Reis AA. Global health and big data: The WHO's artificial intelligence guidance. *S Afr J Sci.* 2023;119(5/6), Art. #14725. <https://doi.org/10.17159/sajs.2023/14725>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

KEYWORDS:

artificial intelligence, big data, bioethics, global health, governance

FUNDING:

US National Institutes of Health (U01MH127704)

PUBLISHED:

30 May 2023



Global health and big data: The WHO's artificial intelligence guidance

Significance:

The growth and adoption of artificial intelligence tools and systems has the potential to transform health and wellness, even as this expansion raises challenging ethical issues, including data and privacy protection, appropriate uses and users, human rights concerns, and inequitable access. The WHO in 2020 committed to an 18-month process of guideline development, leading to the 2021 publication of the WHO's Guidance on *Ethics and Governance of Artificial Intelligence for Health*. The document identifies salient ethical principles, assesses a range of ethical issues and challenges, addresses governance strategies, and offers recommendations; it is apparently the first to offer global guidance.

Introduction

Rarely in the history of science has a new tool or technology engendered the excitement, concern, and interest as artificial intelligence and machine learning in health and medicine. Although the Human Genome Project is a noteworthy antecedent in this regard, more lives will likely be touched by health information technology, including artificial intelligence (AI), than genetics – at least for the foreseeable future.

The world's bioethics community has risen to the occasion with extraordinary thoughtfulness and, indeed, rapidity, as it seeks to keep pace with the ever-expanding uses of AI for health. Scholars on nearly every continent have turned or refocused their attention to challenges raised by the use of intelligent machines in clinical practice, public health, and biomedical research. This has led to a significant increase in the literature on AI and big data ethics over the past several years, including recommendations for appropriate use and users of a burgeoning technology.

Against this background, the World Health Organization (WHO), which for some two decades has supported a Global Network of Collaborating Centres for Bioethics¹, organised a first WHO meeting on Ethics, Big Data and AI in 2017. This consultation, hosted by the University of Miami Miller School of Medicine's Institute for Bioethics and Health Policy, was to identify the scope and range of ethical issues and questions related to big data and AI in health, in order to inform the work of WHO and to develop future principles and guidance for stakeholders.²

In 2020, after more than two years of consultations with Member States and many other stakeholders, the 73rd World Health Assembly adopted the 'Global strategy on digital health 2020–2025'.³ The vision of the global strategy is to improve health for everyone, everywhere, by accelerating the development and adoption of appropriate, accessible, affordable, scalable, and sustainable person-centric digital health solutions to prevent, detect, and respond to epidemics and pandemics; to develop infrastructure and applications that enable countries to use health data to promote health and well-being; and to achieve the health-related Sustainable Development Goals and the 'Triple Billion' targets of WHO's Thirteenth General Programme of Work, 2019–2023. The strategy is built on four strategic objectives:

1. To promote global collaboration and advance the transfer of knowledge on digital health.
2. To advance the implementation of national digital health strategies.
3. To strengthen governance for digital health at global, regional, and national levels.
4. To advocate people-centred health systems that are enabled by digital health.

These strategies are intended to provide guidance and coordination on global digital health transformation and to strengthen synergies between initiatives and stakeholders to improve health outcomes and mitigate associated risks at all levels.

Based on this previous work, WHO in 2020 committed to an 18-month process of guideline development, analysis of prior work and a comprehensive synthesis, leading to the 2021 publication of the WHO's Guidance on *Ethics and Governance of Artificial Intelligence for Health*.⁴ The guidance was based upon the collective knowledge and insights of an international and multidisciplinary expert group from academia, government, industry, law, and non-governmental organisations, including human rights organisations, and represented all WHO regions. The report declares that⁴:

...for AI to have a beneficial impact on public health and medicine, ethical considerations and human rights must be placed at the centre of the design, development, and deployment of AI technologies for health. For AI to be used effectively for health, existing biases in healthcare services and systems based on race, ethnicity, age, and gender, that are encoded in data used to train algorithms, must be overcome. Governments will need to eliminate a pre-existing digital divide (or the uneven distribution of access) to the use of information and communication technologies. Such a digital divide not only limits use of AI in low- and middle-income countries but can also lead to the exclusion of populations in rich countries, whether based on gender, geography, culture, religion, language, or age.

The document reviews a variety of AI applications; salient laws, policies and principles; key ethical principles; ethical challenges; guidance for "building an ethical approach" to health AI; "liability regimes"; and several areas of



governance – for example of data, intellectual property, and the private sector – that can assure that ethical principles can be effectively applied.

Global data and AI context

Data are the fuel of artificial intelligence. Data from a vast range of sources are collected, stored, shared, and then analysed by AI systems, which are tuned or trained on very large data sets. There are, moreover, many data and information sources applicable to the use of AI for health, and they range across varied domains: mobile use and user location, clinical care, public health repositories and registries, biomedical research – as well as data which, while not explicitly about health, bears on people's well-being. From finance and food to transportation and other social determinants of health, these and other domains all constitute and shape a vast digital ecosystem. Artificial Intelligence programs run on such records.

'Data' and 'information' are often and sometimes wrongly used synonymously. 'Data' has come, in many contexts, to refer to machine-readable or processable representations of facts. The binary code for 'kidney disease', for instance, is 01101011 01101001 01100100 01101110 01100101 01111001 00100000 01100100 01101001 01110011 01100101 01100001 01110011 01100101. Data can become information when rendered as facts humans can understand. A database might contain ones and zeroes, diagnostic references, or natural language expressions, for instance. In principle, all of these can be coded and so 'de-identified' or 'pseudonymised', or scrambled without a code and likely anonymised.

The ability to link or aggregate disparate data sets offers profound scientific opportunities, from improving diagnoses to guiding public health interventions to enhancing biomedical research. It also raises equally profound ethical issues. AI, or 'knowledge discovery in databases', mines these data sets in search of patterns. Such patterns could help clinicians prevent and treat disease but also, depending on the adequacy of security protocols and legal protections, expose individuals to confidentiality breaches. These patterns can help public health scientists identify disease trajectories and shape interventions to limit, say, pandemics – and they can foster stigma against some populations or population subgroups. In the opposite direction, to the extent that AI tools can improve clinical care and the health of populations, those individuals and populations without access to care and devices to improve it (those who exemplify the 'digital divide') are unlikely to benefit from the new technology. Generally, data applied to AI is biased towards the majority and may place a minority population – whether on the basis of race, gender, or age – at a disadvantage, with such biases enshrined in the AI.

Moreover, AI software can be difficult to explain and understand, and is sometimes or often not fully transparent; it is often biased; and it is frequently unclear who or what is responsible for oversight, maintaining standards, or ensuring safe use. This is in part the challenge of governance, some credible form of which is widely recognised as necessary if AI applications are to be trustworthy, trusted, and successfully used.

Against this background, the WHO guidance development group grappled with competing values, conflicting duties, and diverse stakeholder interests. It was essential to identify a set of core values that would undergird the final report and guide its conclusions and recommendations.

Ethical principles

The WHO report reflects the trade-offs that should be considered to ensure that potential benefits of AI application to clinical practice, public health, or biomedical research do not outweigh the technology's risks, while also assuring that certain core values and rights are fully protected. Most generally, it is uncontroversial to require that AI in health (and, indeed, in many other domains) be used fairly, avoid bias and discrimination, and promote equitable access. Healthcare systems can help achieve these ends by decreasing cost, ensuring diagnostic accuracy, and "storing and managing data [and] data collection via electronic health records, and exponential consumer data generation [creating] a data rich healthcare ecosystem"⁵.

Principles that should govern the development and use of big data and AI had already been enunciated by various organisations and countries. In fact, an analysis published in 2020 at the outset of the WHO guidance development process identified 36 sets of principles which either applied to the whole range of applications of AI or specific stakeholders/end-users (private sector, intergovernmental organisations, civil society, government, and multistakeholders).⁶ That and other initiatives point more broadly to the extraordinary amount of work devoted to establishing foundations for the ethically optimised use of AI tools. These initiatives may be regarded as a kind of international ethics "crowdsourcing", the best antecedent for which is perhaps that of the Ethical, Legal, and Social Implications project that helped guide the Human Genome Project more than 30 years ago.⁷

The principles identified and agreed to by the WHO international expert group are the first specifically geared toward AI in health with international scope. The six principles endorsed by WHO are:

- protecting human autonomy
- promoting human well-being and safety and the public interest
- ensuring transparency, explainability and intelligibility
- fostering responsibility and accountability
- ensuring inclusiveness and equity
- promoting AI that is responsive and sustainable

The WHO's experts intended these principles to be used as a basis for governments, programmers, companies, civil society, and inter-governmental organisations to adopt ethical approaches to guide appropriate use of AI for health. To be sure, any individual organisation might want to adapt or augment this or any set of principles and, indeed, the process of doing so should be regarded as an important exercise in ethics analysis, professional development, and community engagement.

Ethical challenges

Principles alone do not provide guidance. They 'govern' conceptually and should inform debate surrounding practical questions and challenges. The first of these addressed by WHO was fundamental: should AI systems be used in the first place? Navigating between eager promotion and hyperbolic caution, the WHO report states that the benefits of AI systems can be realised only if they are unbiased, transparent, safe, and,

Even after an AI technology has been introduced into a health-care system, its impact should be evaluated continuously during its real-world use, as should the performance of an algorithm if it learns from data that are different from its training data. Impact assessments can also guide a decision on use of AI in an area of health before and after its introduction.⁸

Ethical challenges addressed by WHO's work include¹:

- Digital divide – It was clear that the growth and update of AI tools should not worsen disparities shaped by limited access, and that technology providers "should be required to provide infrastructure, services and programs that are interoperable" as countries narrow the divide.
- Data collection and use – From privacy to "function creep" and the commercialisation of personal data and information, the team debated the scope and limits of "appropriate use" and "appropriate users".
- Data colonialism – At ground here, for instance, is the concern that high-income countries with "strict regulatory frameworks and data protection laws" might collect data from low- and middle-income countries that lack parallel data-protection laws.
- Accountability and responsibility – Basic ethical obligations related to standards, safety, and quality of AI systems rely on system

developers, vendors, users, and their institutions to make plain and adhere to processes for ensuring best practices.

- Autonomous decision-making – The questions whether and to what extent an AI tool may operate without human control continue to be among the most interesting and challenging at the intersection of ethics and intelligent systems. Moreover, institutions must address the related questions of whether and to what extent patients and communities ought to be informed if self-governing machines are making medical or public health decisions.
- Bias and discrimination – That training sets introduce racial and other biases into AI systems remains a source of deep disquiet among scholars and advocates. Awareness and a plenary attention to mitigation is essential if future AI tools are to enjoy the trust of the communities they purport to serve and not exacerbate existing biases that undermine healthcare provision and patient outcomes.
- Safety and cybersecurity – Among key findings here is that safety and security issues might arise even after a thorough review before a system's implementation. This underscores the need for ongoing vigilance.
- Labour and employment – AI adoption might have a deleterious effect on clinicians' professional development and engender skill degradation and, indeed, good systems might even replace traditional humans through various forms of automation.
- Commercialisation – Although markets can drive innovation, they can also corrupt the environments they shape. A concern raised by the expert team: "When most data, health analytics and algorithms are managed by large technology companies, it will be increasingly likely that those companies will govern decisions that should be taken by individuals, societies and governments..."
- Climate change – Some AI applications generate non-trivial emissions of greenhouse gases and have other effects on the environment. The WHO working group calls for "stringent oversight by governments and good governance".

The process to develop the guidance document revealed the rich scope of AI ethical issues and challenges faced by the world's health community, as well, significantly, as the extraordinary effort by the informatics and ethics scholars to address them. Indeed, the task of analysing and synthesising the many previous and ongoing efforts to foster ethical and trustworthy AI – and doing so for an international community – was an opportunity to identify the most compelling arguments for good practice, as well as those approaches most likely to succeed. An overarching goal was to encourage consensus in a complex and fraught environment.

Governance

Good governance requires more than carefully vetted and balanced values. In parallel to the appropriate and adequate oversight of AI systems, the WHO working group addressed issues of data control and sharing, data sovereignty, transparency, valid consent and its breadth or scope, benefit sharing, and the potential role of federated data. An exemplary governance scheme must also encompass accountability and responsibility. Two overarching governance questions need to be addressed: what exactly should be governed and who or what should do it? According to the guidance document⁴:

Governance in health covers a range of steering and rule-making functions of governments and other decision-makers, including international health agencies, for the achievement of national health policy objectives conducive to universal health coverage. Governance is also a political process that involves balancing competing influences and demands.

The rapid and broad growth of AI research, development, and adoption embeds numerous points and processes to monitor and influence.

The software development lifecycle is already in many cases vetted for reliability and quality, albeit not explicitly for ethics. Likewise, the creation, maintenance, and use of databases used for training AI algorithms. The question of which points and processes to oversee or scrutinise will likely be best answered after a thorough review of which oversight strategies are found most effective in achieving the goal of fair and trustworthy systems. This is in part an empirical question.

As to the question of what entities should exercise a governance function, the most apt approach will be multifaceted. This means that there might be a role for software developers themselves; their commercial, academic, and government employers; institutions that use the systems; professional societies; perhaps even a kind of lay oversight, a regulatory version of 'citizen science'. There is also a need for legislative action to compel testing, evaluation, and adherence to best practices, and a regulatory apparatus that can put such laws into good practice. WHO is currently developing a separate guidance document that examines regulatory considerations that governments may adopt. There are already ample precedents for such regulatory supervision in the form of data privacy laws in individual states (e.g. South Africa's *Protection of Personal Information Act* and the *Health Insurance Portability and Accountability Act* in the USA) and their federations (the European Union's General Data Protection Regulation). As is the case in many other areas of health care, civil society, patients, and communities that are most directly affected by the deployment of such technologies must have adequate means to influence the development and use of AI. Thus, the WHO guidance document recommends⁴:

Patients, community organizations and civil society should be able to hold governments and companies to account, to participate in the design of technologies and rules, to develop new standards and approaches and to demand and seek transparency to meet their own needs as well as those of their communities and health systems.

Conclusion

WHO and several other organisations have issued normative frameworks on the ethical development and use of AI for health. Now more efforts are needed to ensure that these international norms are taken up by the various stakeholders (from governments to industry) and implemented in daily practice. Specific tools need to be developed (for programmers to actually implement 'ethics by design' in their work; for governments to address the ethical challenges in their laws and regulations; etc.). Technology and knowledge transfer need to be promoted alongside investments to overcome an enduring digital divide. The effort to forge the first global guidelines to meet ethical challenges raised by this exciting new technology is both an affirmation of shared values and an opportunity to ensure appropriate use of this technology.

Acknowledgements

Research reported in this publication was supported in part by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Competing interests



We have no competing interests to declare. Prof. Goodman and Dr Litewka are leaders of a WHO Collaborating Centre in Ethics and Global Health Policy (for which they receive no funding), and the other authors are employees of the WHO. The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.



References

1. World Health Organization. Global network of WHO collaborating centres for bioethics [webpage on the Internet]. c2023 [cited 2023 Apr 28]. Available from: <https://www.who.int/groups/global-network-of-who-collaborating-centres-for-bioethics>
 2. World Health Organization. Big data and artificial intelligence for achieving universal health coverage: An international consultation on ethics [webpage on the Internet]. c2018 [cited 2023 Apr 28]. Available from: <https://www.who.int/publications/i/item/WHO-HMM-IER-REK-2018-2>
 3. World Health Organization. Global strategy on digital health 2020-2025 [document on the Internet]. c2021 [cited 2023 Apr 28]. Available from: <https://apps.who.int/iris/bitstream/handle/10665/344249/9789240020924-eng.pdf>
 4. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. c2021 [cited 2023 Apr 28]. Available from: <https://www.who.int/publications/i/item/9789240029200>
 5. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: A report from the National Academy of Medicine. *JAMA*. 2020;323(6):509–510. <https://doi.org/10.1001/jama.2019.21579>
 6. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication no. 2020-1 [document on the Internet]. c2020 [cited 2023 Apr 28]. Available from: <https://doi.org/10.2139/ssrn.3518482>
 7. National Human Genome Research Institute. ELSI planning and evaluation history [webpage on the Internet]. c2005 [cited 2023 Apr 28]. Available from: <https://www.genome.gov/10001754/elsi-planning-and-evaluation-history>
 8. London AJ. Groundhog day for medical artificial intelligence. *Hastings Cent Rep*. 2018;48(3). <https://doi.org/10.1002/hast.842>
-

**AUTHORS:**

Adetayo E. Obasa¹ 
 Andrea C. Palk² 

AFFILIATIONS:

¹Centre for Medical Ethics and Law, WHO Bioethics Collaborating Centre, Department of Medicine, Stellenbosch University, Cape Town, South Africa

²Unit for Bioethics, Centre for Applied Ethics, Philosophy Department, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:

Adetayo Obasa

EMAIL:

obasa@sun.ac.za

HOW TO CITE:

Obasa AE, Palk AC. Responsible application of artificial intelligence in health care. *S Afr J Sci*. 2023;119(5/6), Art. #14889. <https://doi.org/10.17159/sajs.2023/14889>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

KEYWORDS:

artificial intelligence, responsible application, algorithm biases, ethical governance and regulation

FUNDING:

US National Institutes of Health (U01MH127704)



Responsible application of artificial intelligence in health care

Significance:

The responsible application of artificial intelligence (AI) in health care is crucial as it has the potential to revolutionise medical practices. AI technologies can analyse medical data, identify patterns, and generate insights that can inform clinical decision-making, improve patient outcomes, and reduce healthcare costs. However, the ethical, legal, and social implications of AI in health care must be considered to ensure that its implementation is safe, transparent, and equitable. It is essential to prioritise the responsible application of these technologies to maximise their benefits and minimise potential harm. As AI continues to advance, its responsible application will play a vital role in shaping the future of health care.

Introduction

Artificial intelligence (AI) can broadly be defined as the computational simulation of complex intellectual processes associated with intelligent human behaviour, such as learning, decision-making, problem-solving, executing tasks and self-correction.¹⁻³ While the application of AI has widespread potential, its possibilities in health care are particularly significant, with research findings indicating that these technologies can already outperform humans in key healthcare tasks. For example, AI-powered machines are assisting radiologists in timeously identifying malignant tumours.⁴ The introduction of AI in the healthcare sector is primarily aimed at supporting the move towards precision medicine, including ensuring more efficient and accurate diagnoses and treatment plans. This will also have the benefit of relieving clinicians from the burden of mundane tasks. In this regard, AI technologies were successfully used during the COVID-19 pandemic to assist decision-making about prioritisation and allocation of scarce resources.⁵ While the introduction of AI in the healthcare sector is primarily aimed at improving service delivery within the industry⁶, the impact it will have on the healthcare sector as a whole, and on patient well-being in particular, will depend on how AI is developed, applied and regulated. Related to these are several ethical concerns that require urgent and continued attention.

First, to perform a given task with precision and efficiency, AI systems require access to extensive data sets. Within the healthcare context, these data sets are patient health information that would have been obtained from private and public hospitals, including government entities. This raises privacy concerns relating to data security as well as to ensuring that the appropriate consent to use data has been sought. Second, given human involvement in the initial training and learning of these systems, there are concerns that existing human prejudices and biases may inadvertently be introduced, leading to algorithmic, and consequently, decision-making biases. This has implications for health equity. Third, AI systems might make errors as part of the process of learning and becoming more efficient. If such systems improve to the extent that they can operate autonomously, we may have to reconfigure our models of responsibility and liability to accommodate such errors. These concerns regarding AI in health care are by no means exhaustive, but we regard them as particularly salient. Moreover, they imply the need for responsible and effective governance and regulation informed by a multidisciplinary and collaborative approach that considers the full array of ethical, legal, social, and economic implications of the use of AI technologies.⁷ In this Perspective, we discuss each of these concerns and provide some suggestions for ensuring responsible AI in health care.

Data security, privacy and appropriate consent

Ethical AI includes respecting privacy as a fundamental value and right which in turn requires data security and protection.⁸ In South Africa, the *Protection of Personal Information Act* (POPIA) balances the right to protection of privacy, access to information and freedom of expression.⁹ This is pertinent given that for AI to function optimally in the healthcare sector, it requires access to extensive personal biometric information and data. However, POPIA does not accommodate all the specificities and challenges posed by the use of AI in health care. With the new reality of big data, mass quantities of patient data and personal data would be required by big tech companies to train and build algorithms. Although the data would be de-identified, the risk of reidentification remains plausible. Recent studies have shown how computational strategies can be used to reidentify individuals in health data repositories managed by both public and private institutions.^{10,11} One such study found that an algorithm could be used to reidentify 85.6% of adults and 69.8% of children in a physical activity cohort study “despite data aggregation and removal of protected health information”¹². Insofar as the possibility of reidentification poses a significant obstacle to privacy, there is a need for new and improved data regulations that bolster this value and right. With the rapid pace of technology development, there are gaps in regulation and oversight that should be addressed through an innovative and multidisciplinary approach.

A related concern is how to ensure that appropriate models of consent have been used to obtain permission for the use of personal patient data, given that AI systems require access to vast data sets. The challenge here is ensuring that individual patients understand how their data might be used and the risk of reidentification, both requirements for meaningful consent. Moreover, as AI systems develop further, and are able to perform increasingly complex procedures, securing consent may prove challenging. While a sufficiently informative explanation of AI-enabled procedures would be necessary to ensure meaningful consent, the possibility of mistrust or fear of such technologies would require consideration. This implies that more studies are needed, in contexts in which such systems might be used, in order to ascertain optimum ways of communicating information and risks regarding these complex technologies.

Algorithm biases and health equity

As mentioned above, access to vast data sets is crucial for the optimum functionality of AI, and for the process of machine learning and algorithm development, in particular. Therefore, if the data set itself is biased, this bias is transferred to the model that learns from the data. There is evidence that algorithm bias has already found its way into some AI devices; for example, pulse oximeters which have lower accuracy for populations with non-European ancestry due to the associated algorithms drawing on data sets comprised predominantly from populations of European ancestry.¹³ This raises distinct concerns about equity in health care. Biases fall into three main categories. First, bias could occur when skewed or misrepresentative data are fed as training data into an algorithm, for instance, data sets that exclude or underrepresent vulnerable populations, as is the case in the above example. Second, bias could occur due to malfunction or faulty algorithms. Third, bias could be introduced due to human prejudice informed by erroneous assumptions. In Africa, limited high-quality electronic data due to non-uniform or incomplete data sets could undermine data-oriented technologies and further exacerbate bias. Concerns about algorithmic inclusivity and the perpetuation of such biases are particularly urgent given that populations with African ancestry, across the globe, and in Africa, in particular, continue to be negatively impacted and harmed by ongoing prejudice. In clinical contexts in which AI is involved in diagnoses or providing predictions about the best possible treatment outcomes, biases in algorithmic processes could lead to serious harms related to misdiagnoses or inappropriate treatment. The responsible use of AI requires that its deployment in health care must be free from bias, and data ethics governance should be established to oversee software and algorithm development.¹⁴

'Black box' AI systems, trust and responsibility

Machine learning refers to the system of coded algorithms by which engineers inform artificial intelligence systems what to learn, what rules to apply to the learning process and the fundamental principles to apply. However, in the case of certain kinds of machine learning, these rules are not always fixed, they can be changed by the machine itself.¹⁵ Machine learning is commonly used in precision medicine to predict what treatment protocols will succeed based on various patient attributes and the treatment context.¹⁶ More complex forms of machine learning involve deep learning or neural network models with several layers of features and variables that predict outcomes. For example, a typical application of deep learning in health care is the recognition of potentially cancerous lesions in radiology images.

In clinical contexts there are concerns about the more complex forms of machine learning techniques, particularly the so-called 'black box' systems. The concern here is that black box systems are characterised by "opacity, complexity, and unpredictability" with the result that it is not possible to ascertain the process by which these systems deliver their output.¹⁷ While such systems are highly efficient, the possibility of errors is also a precondition of part of the learning process, in the same way that human beings learn more effectively through the allowance of error.¹⁵ Black box systems raise numerous ethical concerns, including explicability and accuracy, patient-clinician trust and broader questions regarding responsibility and liability in the case of errors or decisions that produce harmful consequences. In terms of the former, trade-offs might be required between increasing accuracy (at the cost of explainability) and enhancing a system's explainability (which may reduce its accuracy).¹⁸ However, the degree of necessary explicability depends on the context and the risk involved. When there is a high risk of harm or negative outcomes associated with the decisions of such systems, we should be able to ascertain a full understanding of the decision-making process of the system. This implies that black box systems, for which such an explanation is not possible, should not be used with procedures that carry such high risk.

Currently, AI technologies support clinicians in decision-making, rather than operating autonomously; however, insofar as these systems improve and are able to operate independently, the transfer of decision-making from human agents to AI will elicit considerable ethical and legal concerns. Given that the law is configured in terms of the rights and

obligations of human persons, an argument can be made that these rights should not be solely subjected to automated devices, especially when their decisions could have dire consequences.^{19,20} In South Africa, the da Vinci Xi fourth-generation system, one of the most advanced surgical robots in the world, is currently used by surgeons to perform robotic-assisted minimally invasive surgery in two public hospitals and several private hospitals.²¹ This system has been built drawing on knowledge gained over the past two decades, ensuring substantial improvements in design and performance; its precision and accuracy cannot be overemphasised. While da Vinci is not fully autonomous, there is a possibility that future iterations might be deemed capable of independently performing specific tasks, carrying out decision-making processes, and proposing and validating strategies. Various ethical challenges will need to be addressed by regulatory bodies before this possibility is realised. As mentioned above, these include informed consent related challenges but also possibly a need to reconfigure our frameworks of responsibility to account for such autonomous systems as well as our legal frameworks in terms of liability for errors that might be made during procedures or associated harms.

Moreover, to foster trust and transparency, these systems might require the capacity to be sensitive to both ethical and social values in various multicultural contexts, and to justify their output, not only in the case of errors but in general. This would of course depend on the nature and purpose of the system. Trust is fundamental to the clinician-patient relationship insofar as the success of most medical interventions depends on it. As evidenced by previous abuses of trust in clinical and research contexts, this relationship is tenuous. While doctor-patient trust could be conferred to AI systems, any small failure in AI could significantly erode public confidence in health care. Once again, these challenges indicate a need for a regulatory framework that protects the safety of end users and ensures that the development of these devices is informed by a concern for fundamental human principles and values.

Ethical governance and regulation

The report on *Ethics & Governance of Artificial Intelligence for Health* published by the World Health Organization in 2021 offers an excellent and practical resource for responsible development, design, use and regulation of AI.²² The guiding principles suggested in the report emphasise that the use, governance and regulation of AI should promote autonomy, well-being, trust, accountability, and equity, whilst being sustainable.²²

In the context of considering ethical AI in health care, the notion of responsibility is fundamental. This includes both retrospective responsibility and prospective responsibility. The former is relevant in the case of dealing with errors that might be made by such systems, implying accountability or the need to be able to understand and explain the decisions of such systems, including any errors. In cases where harm is caused by an AI system in healthcare contexts, we should ensure that human beings are meaningfully involved in a way that we can identify parties who can be held accountable and responsible. However, the implication here is that completely autonomous AI systems that employ black box processes should not be used in certain healthcare contexts, given that such systems are not appropriate targets of our ascriptions of responsibility and accountability. Prospective responsibility requires that all stakeholders assume the duty to ensure the ethical roll out of AI. Responsible AI also underscores the significant role that educational interventions can play to ensure widespread knowledge and awareness and promote public acceptability and participation. Developers and manufacturers of these devices must also be accountable to regulatory bodies and the public. Furthermore, there is a need for a regulatory framework mechanism to ensure that algorithm processes involved in AI systems meet declared ethical standards and expectations, such as the World Health Organization's guidelines.²²

Conclusion

Given the enormous potential of AI to improve health care and enhance health outcomes in other areas, there will undoubtedly be an increase in the use of such systems over the next few decades. Addressing the above concerns will require ongoing ethical discussion, good governance



and robust regulation. As argued by Jonas²³, the development and application of science and technology should be grounded in recognition of the responsibility we bear to future generations. In the case of AI, we must govern and regulate it with awareness of the impact of our decisions on the well-being not only of all human beings who currently live, but also of those in the future.

Acknowledgement

Research reported in this publication was supported by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

We have no competing interests to declare.

References

1. Jha S, Topol EJ. Adapting to artificial intelligence. *JAMA*. 2016;316:2353–2354. <https://doi.org/10.1001/jama.2016.17438>
2. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25:44–56. <https://doi.org/10.1038/s41591-018-0300-7>
3. Korteling JE (Hans), Van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR. Human- versus artificial intelligence. *Front Artif Intell*. 2021;4. <https://doi.org/10.3389/frai.2021.622364>
4. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, et al. Deep learning in radiology. *Acad Radiol*. 2018;25:1472–1480. <https://doi.org/10.1016/j.acra.2018.02.018>
5. Singh JA, Moodley K. Critical care triaging in the shadow of COVID-19: Ethics considerations. *S Afr Med J*. 2020;110:355–359. <https://doi.org/10.7196/SAMJ.2020v110i6.14842>
6. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719–731. <https://doi.org/10.1038/s41551-018-0305-z>
7. Dubber MD, Pasquale F, Das S, editors. *The Oxford handbook of ethics of AI*. New York: Oxford University Press; 2020. <https://doi.org/10.1093/oxfordhb/9780190067397.001.0001>
8. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1:389–399. <https://doi.org/10.1038/s42256-019-0088-2>
9. Protection of Personal Information Act (POPI Act). Available from: <https://popia.co.za/>.
10. Check Hayden E. Privacy loophole found in genetic databases. *Nature*. 2013. <https://doi.org/10.1038/nature.2013.12237>
11. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339:321–324. <https://doi.org/10.1126/science.1229566>
12. Na L, Yang C, Lo CC, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA Netw Open*. 2018;1, e186040. <https://doi.org/10.1001/jamanetworkopen.2018.6040>
13. Tobin MJ, Jubran A. Pulse oximetry, racial bias and statistical bias. *Ann Intensive Care*. 2022;12:2. <https://doi.org/10.1186/s13613-021-00974-7>
14. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight – reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383:874–882. <https://doi.org/10.1056/NEJMms2004740>
15. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf Technol*. 2004;6:175–183. <https://doi.org/10.1007/s10676-004-3422-1>
16. Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun*. 2018;9:42. <https://doi.org/10.1038/s41467-017-02465-5>
17. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philos Technol*. 2021;34:1057–1084. <https://doi.org/10.1007/s13347-021-00450-x>
18. London AJ. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent Rep*. 2019;49:15–21. <https://doi.org/10.1002/hast.973>
19. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci Rep*. 2019;9:1879. <https://doi.org/10.1038/s41598-019-38491-0>
20. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Med*. 1996;22:707–710. <https://doi.org/10.1007/BF01709751>
21. Solomons L. Medical renaissance: Tygerberg becomes first public hospital to use da Vinci robot for surgery. *News24*. 23 February 2022. Available from: <https://www.news24.com/news24/southafrica/news/medical-renaissance-tygerberg-becomes-first-public-hospital-to-use-da-vinci-robot-for-surgery-20220223>
22. World Health Organization (WHO). Health workforce. Geneva: WHO; 2021 [cited 2022 Sep 12]. Available from: https://www.who.int/health-topics/health-workforce#tab=tab_1
23. Jonas H. *The imperative of responsibility: In search of an ethics for technological age*. Chicago, IL: University of Chicago Press; 1984.



Check for updates



Managing and assembling population-scale data streams, tools and workflows to plan for future pandemics within the INFORM-Africa Consortium

AUTHORS:

Jenicca Poongavanan¹
Joicymara Xavier^{2,3,4}
Marcel Dunaiski⁵
Houriyah Tegally^{1,6}
Sunday O. Oladejo¹
Olawole Ayorinde⁷
Eduan Wilkinson¹
Cheryl Baxter^{1,8}
Tulio de Oliveira^{1,6,8,9}

AFFILIATIONS:

¹Centre for Epidemic Response and Innovation (CERI), School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa

²Institute of Agricultural Sciences, Federal University of Vales do Jequitinhonha e Mucuri, Unai, Brazil

³René Rachou Institute, Oswaldo Cruz Foundation, Belo Horizonte, Brazil

⁴Institute of Biological Science, Federal University of Minas Gerais, Belo Horizonte, Brazil

⁵Department of Computer Science, School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa

⁶KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa

⁷Institute of Human Virology, Abuja, Nigeria

⁸Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

⁹Department of Global Health, University of Washington, Seattle, WA, USA

CORRESPONDENCE TO:

Jenicca Poongavanan

EMAIL:

jenicca1193@gmail.com

HOW TO CITE:

Poongavanan J, Xavier J, Dunaiski M, Tegally H, Oladejo SO, Ayorinde O, et al. Managing and assembling population-scale data streams, tools and workflows to plan for future pandemics within the INFORM-Africa Consortium. *S Afr J Sci.* 2023;119(5/6), Art. #14569. <https://doi.org/10.17159/sajs.2023/14659>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

KEYWORDS:

INFORM-Africa data management, collaboration, SARS-CoV-2, HIV

FUNDING:

US National Institutes of Health through INFORM-Africa (U54 TW012041), eLwazi Open Data Science Platform and Coordinating Center (U2CEB032224)

PUBLISHED:

30 May 2023



Significance:

The INFORM-Africa Consortium, a research hub of the NIH-funded DS-I Africa, will leverage the Data Management and Analysis Core (DMAC) and Next Generation Sequencing (NGS) Core to ensure effective data management and analysis. The DMAC will capture and analyse data, making it accessible to collaborators across multiple African countries and future research hubs. The aim is to increase access to high-quality, reproducible data that can be used to engage policymakers and better prepare for future pandemics, while also removing barriers to data sharing and integration across institutions. Ultimately, this goal will facilitate data-driven decision-making and advance public health initiatives.

Introduction

The SARS-CoV-2 virus has caused over 12 million recorded cases of COVID-19 in Africa with over 256 000 lives claimed.¹ The rapid growth of COVID-19 to pandemic proportions in Africa occurred against a backdrop of existing epidemics of HIV, tuberculosis and malaria and a rising burden of non-communicable diseases, which placed additional demands on already strained healthcare systems. In 2020, an initial survey by the World Health Organization (WHO), using clinical and epidemiological data predominantly from South Africa, suggested that people living with HIV were 30% more likely to die from COVID-19 among those hospitalised with the disease.^{2,3} On the other hand, some reports indicate that HIV infection itself does not appear to be a risk factor for severe COVID-19.^{4,5} Individually and collectively, these studies do not provide sufficient data, due largely to their limited sample sizes, to understand the relationship between SARS-CoV-2 infection and HIV. In order to create a core capacity for governments across Africa to better respond to current and future epidemics, it is crucial to understand the synergies at work between the two diseases at a population scale.

To better address issues around public health, it is important to develop the capacity to effectively generate, collect, store, clean, annotate, link, and share data from diverse sources. Furthermore, a research gap in epidemic modelling around the world, and specifically in Africa, is the lack of population-scale epidemiologic data sources, properly annotated and linked across health services. Between continent-wide technical and infrastructural resource limitations and fragile health systems, the need for population-scale epidemiological and frequently updated data, is even more urgent to inform interventions rapidly.^{6,7} Consequently, 'The Role of Data Streams in Informing Infection Dynamics in Africa' (INFORM-Africa) Research Hub was established; this Hub focuses on the effective use of big data from South Africa and Nigeria as a cornerstone of future pandemic preparedness.

The INFORM-Africa Hub consists of three main project groups: Project 1 focuses on how viral genomic variation, adoption of public health mitigation measures, and mobility patterns contribute to spatially and temporally explicit pathways of SARS-CoV-2 transmission at local and regional scales. Project 2 examines the effect of movement-based restrictions on mobility in Nigeria and South Africa, compares pre-pandemic to post-pandemic movement patterns using cell phone mobility data, and associates specific movement patterns with COVID-19 risk factors. This model incorporates state-of-the-art mobility analytics from the transportation sector, applying them to the African context, possibly for the first time. Project 3 studies the interplay between SARS-CoV-2 and HIV in the two countries most impacted by the syndemics in Africa. It investigates to what extent shared geospatial, mobility and demographic factors affect risk of both infections and how each infection affects the outcomes of the other, and whether the host genetic variation in Africa explains the COVID-19 outcomes in Africa.

During the COVID-19 pandemic, public health data has been employed to gain insight, track, and limit the spread of the virus. There are several institutions that have been collecting, managing and analysing clinical and epidemiological data in Africa, such as South Africa's National Institute for Communicable Diseases (NICD), which is a division of the National Health Laboratory Service (NHLS) in South Africa that conducts surveillance, outbreak investigations and research on communicable diseases. They collect and analyse epidemiological data to monitor the incidence and prevalence of diseases and to guide public health intervention. There is also the South African Medical Research Council (SAMRC) which has a wealth of information. They conduct research on a wide range of health issues, including infectious diseases, non-communicable diseases, and injury-related deaths. They also annually publish a report called the 'South African Burden of Disease Study' which provides detailed information on death and disability in South Africa. One of the SAMRC's key contributions during the pandemic was providing regular updates on the country's COVID-19 status. Stats-SA have also been collating data by conducting household surveys and collecting healthcare data, and at the same time providing statistical analysis.

Moreover, in order for policy decisions to be effective, we need to consider a holistic understanding of the epidemiological situation, including scientific data, as well as the broader context in which the disease is spreading. This context might include factors such as the availability of healthcare resources, the socio-economic impact of public health measures, and the political will to implement effective policies. In addition, the question of improving the reliability and quality of epidemiological data for morbidity, mortality and sero-prevalence in community and hospital settings as well as for understanding the impact of preventive measures, such as vaccination and other

© 2023. The Author(s). Published under a Creative Commons Attribution Licence.

measures through timely and targeted representative sampling methods, becomes crucial. While mobility and infectious agent genomics cannot influence policy alone, they are key factors that need to be integrated into a robust epidemiological data landscape to obtain broader understanding of transmission dynamics and inform effective policy decisions that can help control the spread of infectious diseases.

It is evident that data availability primes research and discovery in the sciences, but the global pandemic coverage has also propelled the engagement with public health data, and data in general, into the public discourse. Data ranging from genomic, patient management and mobility data are crucial for the respective projects to answer the questions they are investigating. However, key challenges include obtaining relevant genomic data and metadata together with patient data, integrating these data originating from multiple sources, applying efficient computational algorithms to cope with these large data sets, and establishing sampling frameworks to enable robust conclusions.

Data sharing amongst data custodians can be contentious and often involves navigating complex policy restrictions and political dynamics. The 2020 State of Open Data report identified trust (or the lack thereof) as a key barrier to data sharing.⁸ To help the INFORM-Africa Research Hub navigate through the ocean of multiple data streams, a Data Management and Analysis Core (DMAC) and Next Generation Sequencing (NGS) core was established. DMAC's responsibility is to address issues of trust, together with managing institutional policies on ethics, intellectual property rights and data ownership agreements – a challenge that requires innovative approaches on data access policies.

The DMAC and NGS Core

The DMAC and NGS core play a key role in assembling and managing the INFORM-Africa Research Hub's data and in providing seamless access to a set of tools and workflows as well as generating next generation sequencing data. The DMAC intends to empower the INFORM-Africa Research Hub by expanding data science research opportunities and capacity in Africa through the involvement of early-stage investigators and trainees, and the data science training and support provided within the INFORM-Africa and across the DS-I Africa Consortium.

The DMAC leverages state-of-the-art computing platforms and uses integrative data analysis frameworks to support the INFORM-Africa Research Hub. The core will accommodate multiple data types (ranging from existing population-scale individual-level clinical data and genomic data to geospatial and mobility data) and additional resources, such as standard operating procedures, protocols and training materials⁹ based on the FAIR principles for scientific data management and stewardship to improve the Findability, Accessibility, Interoperability and Reuse (reproducibility) of data. Guided by these principles, the overarching goal is to provide the Research Hub with a unified environment for data

management, computation, and technical support for collaborative work within and between projects in the INFORM-Africa Hub.

To date, our researchers have been using the traditional model of data analysis in which they are required to download their data from centralised data warehouses onto their local computers, install and maintain their suite of computational tools, and execute analyses using local computing resources. Following the traditional model, each project within the consortium would have to establish and maintain its own data centres, which would create major administrative inefficiencies such as the duplication of data and analysis tools that must be deployed and maintained separately within each centre. The data management tasks also become unsustainable given the unprecedented quantity of genomic and epidemiological data and the frequency at which these data are updated. Furthermore, many data analysis tasks using cutting-edge computational models are impracticable due to the scale of their data requirements and computational complexities when relying on the traditional computation paradigm.

The DMAC will progressively shift towards a more contemporary approach by moving to the cloud. Cloud computing as defined by the US National Institute of Standards and Technology is:

...a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.¹⁰

This definition goes hand in hand with the DMAC's goals and vision within the INFORM-Africa Research Hub.

With the massive volumes of data that we are expecting, integrating multiple genomic, epidemiological and patient data sources can be a complex process and requires techniques to resolve inconsistencies in temporal structure and encoding. To overcome these challenges, we have established a data lake architecture to guide an effective curation process. Data lakes have recently emerged as an enterprise solution to manage large amounts of heterogeneous data for modern data analytics. A data lake architecture can be described as a schema-free repository that allows users to store structured, unstructured, and unprocessed data at any scale, based on cloud computing.^{5,11,12} The main advantage of choosing a data lake architecture as a data management paradigm is the increased flexibility in terms of data type support, as well as the ability to more easily cater to the specific data needs of various users. This will allow the DMAC to continuously support and easily adapt to the requirements of the diverse projects that will be hosted on the DMAC and NGS platform. The DMAC's integrative data management and analysis strategy is summarised in Figure 1.

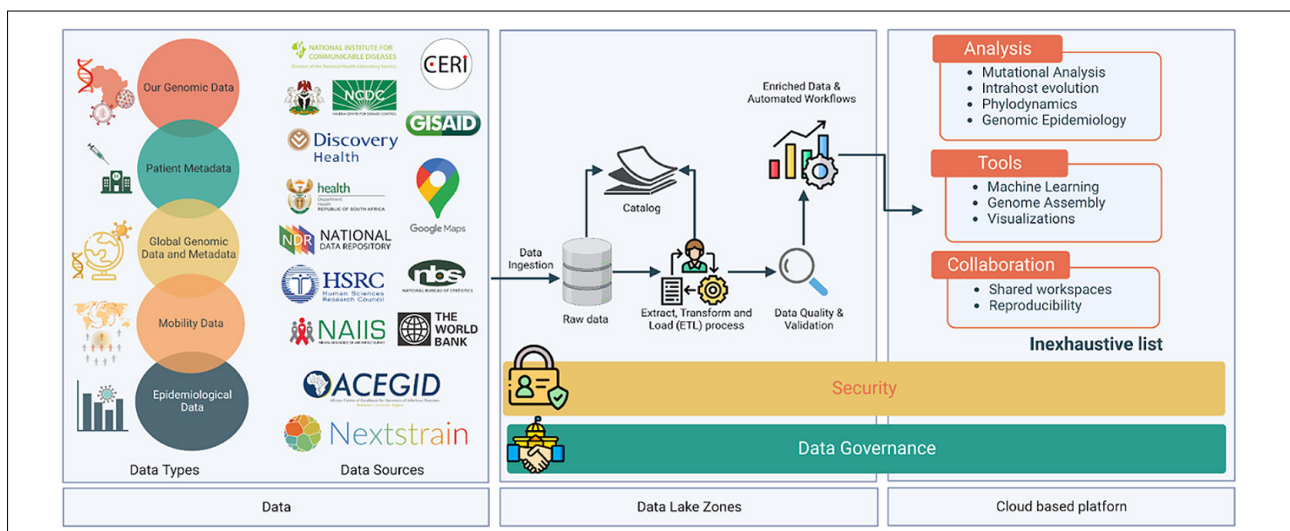


Figure 1: The Data Management and Analysis Core (DMAC) and Next Generation Sequencing (NGS) core architecture workflow.



The data panel in Figure 1 shows the diverse data types that the Hub necessitates and the different data sources that the DMAC and NGS core are responsible for integrating. All the data sources are described in detail in Table 1. The architecture will allow for efficient ingestion of any data types such as genomic files, epidemiological data, or GPS data, while supporting several data access types such as streamed data, batch file uploads, or API access. These are very different data types, each of which requires different standards, formats and storage.

In terms of data governance, access to data is made possible through signed data sharing agreements with public and private providers. We also work closely with the data providers to establish an efficient path for data transfers and updates to existing data sets. Several data sets have already been assembled from various data warehouse sources across Africa. The NGS core contributes toward generating genomic data and metadata that the Research Hub would require.

Once the raw data are acquired, they are deposited on a distributed file system before being curated by the DMAC team. All sensitive data are anonymised and, where necessary, encrypted for storage and transfer. The curation process encompasses extracting and transforming the data into a format that each project within the INFORM-Africa group can use to run their respective analyses. Data curation and quality control measures occur regularly, following a standard protocol for data monitoring and addressing any identified issues, by involving data providers and project investigators.

Once transformed and validated, the data are stored and automatically shared on a data platform built in the cloud that allows extensive collaborative genomic research. The workspace will provide well-established tools to easily filter the data, perform and share analyses. By using a cloud-based data management platform, the DMAC's aim is to create workspaces dedicated to the INFORM-Africa consortium. Through these workspaces, we are able to enforce user access control protocols as required by the various projects.

The DMAC is also invested in sharing high-quality tools and workflows for use in the Research Hub. Dockstore (<https://dockstore.org/>) is a workflow and tool publishing platform that is widely used in the bioinformatics and genomics community. Dockstores can be leveraged both as a source of high-quality workflows and tools as well as a distribution platform for tools and workflows produced by the DMAC team. The goal is to share these workflows with researchers across the Hub so that they can easily use these workspaces.

The DMAC will provide training and support to help researchers develop the platform in line with their needs. The training and support will encompass a variety of topics including data quality assurance and quality control training, especially for early-stage investigators and trainees in the INFORM-Africa Hub. The DMAC and NGS core will also provide support for all new data collection, to ensure uniform data entry procedures and data quality across Hub partners. Additionally, after performing a needs analysis, we will implement an agile data science training programme that includes big data and bioinformatics analysis to train a broad range of stakeholders to manage, process, analyse and interpret biological data together with geospatial and other relevant data as needed.

Conclusion

In summary, the DMAC and NGS core are an essential link between the projects of the INFORM-Africa Research Hub. The DMAC and NGS core will facilitate the seamless integration and linking of various public and private data sets, access to new data and tools created by the projects and cores, and broader sharing, including with the Research Hub and the DS-I Africa Consortium. By using a cloud-based platform, our focus is to enable high-level and reproducible data analysis and cross-network projects between collaborators across the three projects to achieve the overall goal of the INFORM-Africa Research Hub. The platform will enable biological discovery from the big data that is available in Africa. Finally, the DMAC and NGS core will contribute toward the INFORM-Africa aims by expanding data science research opportunities and building capacity throughout Africa.

Table 1: A detailed list of data sets required for INFORM-Africa per topic and country

Topic	Databases from Nigeria	Databases from South Africa
SARS-Cov-2	COVID-19 Household Seroprevalence Survey (NCDC)	South Africa SARS-CoV-2 Seroprevalence Survey (HSRC)
	NCDC COVID-19 database (NCDC)	South Africa National Reference Laboratory Testing Database (NICD)
	ACAPS COVID-19 government measures data set (ACAPS)	Discovery Health (Discovery Health SA)
	Jurisdictional shapefiles across hierarchy	National DATCOV Hospital surveillance for COVID-19 (NICD)
	GRID3 Nigeria Settlement Extents, Version 01.01.	ACAPS COVID-19 government measures data set (ACAPS)
	GRID3 Nigeria - Total COVID risk	Jurisdictional shapefiles across hierarchy
	GRID3 Nigeria - Socioeconomic vulnerability	GRID3 South Africa Settlement Extents, Version 01.01
HIV		GRID3 South Africa Social Distancing Layers, Version 1.0
	Nigeria population-based AIDS impact study (NAIS) (NACA)	HSRC South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, (SABSSM V) (HSRC)
Mobility	National Data Repository (NDR) (NASCP)	South African Department of Health Electronic Patient Management System (EPMS) TIER.NET (SA Dept. of Health)
	Supplementary transportation-sector data (i-TRAFFIC)	Cell Phone Tracking Data (MTI)
	National Bureau of Statistics	Multimodal Transportation Network and point-of-interest information (HERE; OpenStreetMap (OSM); SANRAL)
	Multimodal Transportation Network and point-of-interest information (HERE; OpenStreetMap (OSM); SANRAL)	Supplementary transportation-sector data (i-TRAFFIC)
Survey	Cell Phone Tracking Data (MTI)	ACAPS COVID-19 government measures dataset (ACAPS)
	General Household Survey (World Bank)	General Household Survey (World Bank)
Genomics	The World Bank Open data (World Bank)	The World Bank Open data (World Bank)
	ACEGID database of viral genomic sequences	ACEGID database of viral genomic sequences
	Nextstrain genomic sequencing data sets	Nextstrain genomic sequencing data sets
	GESS genomic sequencing data sets	GESS genomic sequencing data sets
	GISAID genomic sequencing data sets	GISAID genomic sequencing data sets

Acknowledgements

We acknowledge support from the US National Institutes of Health through the INFORM-Africa project that is administered by IHVN (U54 TW012041) and the eLwazi Open Data Science Platform and Coordinating Center (U2CEB032224).

Competing interests

We have no competing interests to declare.



References

1. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. Our world in data: Coronavirus pandemic (COVID-19) [webpage on the Internet]. c2020 [cited 2022 Aug 10]. Available from: <https://ourworldindata.org/coronavirus>
2. Msomi N, Lessells R, Mlisana K, de Oliveira T. Africa: Tackle HIV and COVID-19 together. *Nature*. 2021;600(7887):33–36. <https://doi.org/10.1038/d41586-021-03546-8>
3. World Health Organization (WHO). Clinical features and prognostic factors of COVID-19 in people living with HIV hospitalized with suspected or confirmed SARS-CoV-2 infection, 15 July 2021 [data set]. <https://apps.who.int/iris/handle/10665/342697>
4. Brown LB, Spinelli MA, Gandhi M. The interplay between HIV and COVID-19: Summary of the data and responses to date. *Curr Opin HIV AIDS*. 2021;16(1):63–73. <https://doi.org/10.1097/COH.0000000000000659>
5. Eisinger RW, Lerner AM, Fauci AS. Human immunodeficiency virus/AIDs in the era of coronavirus disease 2019: A juxtaposition of 2 pandemics. *J Infect Dis*. 2021;224(9):1455–1461. <https://doi.org/10.1093/infdis/jiab114>
6. Kucharski AJ, Hodcroft EB, Kraemer MUG. Sharing, synthesis and sustainability of data analysis for epidemic preparedness in Europe. *Lancet Reg Health Eur*. 2021;9, Art. #100215. <https://doi.org/10.1016/j.lanepe.2021.100215>
7. Khan MS, Dar O, Erondu NA, Rahman-Shepherd A, Hollmann L, Ihekweazu C, et al. Using critical information to strengthen pandemic preparedness: The role of national public health agencies. *BMJ Glob Health*. 2020;5(9), e002830. <https://doi.org/10.1136/bmjgh-2020-002830>
8. Porter SJ, Hook DW. How COVID-19 is changing research culture. London: Digital Science; 2020.
9. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3, Art. #160018.
10. Mell P, Grance T. The NIST definition of cloud computing. Gaithersburg, MD: National Institute of Standards and Technology; 2011. <https://doi.org/10.6028/NIST.SP800-145>
11. Giebler C, Gröger C, Hoos E, Schwarz H, Mitschang B. Leveraging the data lake: Current state and challenges. In: Ordonez C, Song I-Y, Anderst-Kotsis G, Tjoa AM, Khalil I, editors. *Big data analytics and knowledge discovery: 21st International Conference, DaWaK 2019; 2019 August 26–29; Linz, Austria*. Cham: Springer International Publishing; 2019. p. 179–188. https://doi.org/10.1007/978-3-030-27520-4_13
12. Ordonez C, Song I-Y, Anderst-Kotsis G, Tjoa AM, Khalil I, editors. *Big data analytics and knowledge discovery: 21st International Conference, DaWaK 2019; 2019 August 26–29; Linz, Austria*. Cham: Springer International Publishing; 2019. <https://doi.org/10.1007/978-3-030-27520-4>



Check for updates

AUTHORS:

Sharon Kling^{1,2}
Shenuka Singh³
Theresa L. Burgess^{1,4}
Gonasagrie Nair¹

AFFILIATIONS:

¹Centre for Medical Ethics and Law, WHO Bioethics Collaborating Centre, Department of Medicine, Stellenbosch University, Cape Town, South Africa

²Department of Paediatrics and Child Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

³Discipline of Dentistry, School of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

⁴Division of Physiotherapy, Department of Health and Rehabilitation Sciences, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:

Sharon Kling

EMAIL:

sk@sun.ac.za

HOW TO CITE:

Kling S, Singh S, Burgess TL, Nair G. The role of an ethics advisory committee in data science research in sub-Saharan Africa. *S Afr J Sci.* 2023;119(5/6), Art. #14724. <https://doi.org/10.17159/sajs.2023/14724>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

KEYWORDS:

ethics, committee, data science, research, Africa

PUBLISHED:

30 May 2023



The role of an ethics advisory committee in data science research in sub-Saharan Africa

Significance:

Data science research involves large volumes of data, often derived from unconventional sources. Given the complex nature of big data research, there is a strong need for the development of ethically appropriate protocols that are sensitive to the complexities of data science and data sources. While reviews of health research by research ethics committees are necessary from an ethical and legal perspective, complementary advisory committees such as ethics advisory committees could be established to advise on ethics challenges more broadly. In this Perspective, we describe a multidisciplinary ethics advisory committee linked to a data science research hub in sub-Saharan Africa.

Data science is an interdisciplinary field in which scientific methods, processes, extremely large data sets, machine learning algorithms and information systems are used to extract knowledge and insights from structured and unstructured data.^{1,2} The United States National Institutes of Health (NIH)-funded programme, ‘Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa)’, was established to create a data science research and training network across Africa.³ The Research for Ethical Data Science in sub-Saharan Africa (REDSSA) project, also NIH-funded, is a unique project planned to complement the focus on data science and its emergence in Africa by exploring the ethical, legal, and social implications (ELSI) of this rapidly growing field. Two of the specific goals of the REDSSA ELSI project are to establish a consortium of sub-Saharan African bioethicists “to develop contextualised guidance in the ELSI of data science” and to establish a Data Science Ethics Advisory Committee (EAC) for the research hubs.² The REDSSA project is linked to the INFORM-Africa research hub, operating out of Nigeria. This research hub studies the interaction between SARS-CoV-2 and Human Immunodeficiency Virus (HIV) with the goal of using the data to improve pandemic preparedness.³

Ethics review of data science research

Research ethics committees (RECs) traditionally review health research protocols involving human participants with the aims of preventing harm and promoting benefit to research participants, while at the same time ensuring that the research is scientifically valid, and fair, and promotes respect for participants and the community.⁴ The focus of the review is on the protection of individuals or groups of participants.

Data science uses big data (large volumes of data), often derived from unconventional sources such as social media, cellular telephone mobility data, wearable technologies, or aggregated health data. Big data have been defined in terms of the three Vs: volume (very large data sets), variety (multiple data formats with structured and unstructured content), and velocity (“high rate of data inflow with non-homogenous structure”).⁵ Big data in the healthcare context are usually collected for reasons other than research. When accessed for research they are aggregated and deidentified.⁶ However, these big data sets are prone to inherent bias as the information is derived from existing data sets. The data analysis yields information relating to types of groups as well as individuals. “The massive scope of big health data coupled with hypothesis-generating interrogation approaches using artificial intelligence (AI) technologies such as machine learning (ML) yields a significant risk of spurious findings.”⁶ As an example, patients from certain racial groups may be systematically disadvantaged by AI based on cost-of-care data rather than on severity of illness, as the disparity in access to health care skews the algorithm.⁷

The use of big data in research has resulted in a shift of terminology describing the participants as ‘data subjects’ or ‘data sources’, rather than the traditional ‘human subjects’ or ‘research participants’. National and international guidelines, legislation, and regulation govern the use and sharing of data to various degrees in different countries in sub-Saharan Africa. In South Africa, legislation such as the *Protection of Personal Information Act* (POPIA)⁸ and the *Promotion of Access to Information Act* (PAIA)⁹ have also become relevant in research ethics. The concern related to this type of research is that researchers can access potentially sensitive data without any engagement with data subjects. Consequently, the risks relate to ‘informational harm’ rather than physical harm. Informational harm includes breaches of privacy and ‘algorithmic discrimination’.^{4,6} The resultant harms are to groups of people as well as individuals, with ensuing emotional distress and discrimination. The complexity of this research means that RECs may not have the expertise to review data science research, or that the research occurs without any involvement of a REC for the review process.⁴ An example of the latter was an experiment conducted via Facebook, in which the news feeds of 689 003 users were manipulated to expose them to greater or lesser amounts of emotional content to show that emotional contagion can occur without the awareness of the participants.¹⁰ The study was severely criticised for the lack of ethics oversight and for potentially exposing vulnerable participants to significant harm.^{11,12}

Regulatory and governance processes have emerged to establish oversight of new research contexts. As an example, Facebook established an Oversight Board in 2020; however, the value of such oversight has been debatable. Ferretti et al. question whether the growth in oversight mechanisms is simply a tick-box process, “motivated by the urge to fill the existing regulatory gaps, or whether it is just ‘ethics washing’”⁴. Oversight mechanisms involved in data sharing, such as data safety monitoring boards and data access committees, are not new and have existed for some time.⁴ However, data science review probably requires closer collaboration between RECs and data safety monitoring boards, with RECs becoming more involved in routine monitoring and oversight of data and data access.

© 2023. The Author(s). Published under a Creative Commons Attribution Licence.



Data access committees

Many countries have established guidelines and policies that govern data sharing within and across their borders. Data access committees (DACs) are tasked with protecting the rights and interests of the parties involved in genomic data sharing by reviewing requests for access. There are three types of DACs: (1) DACs in single research groups, where the study's principal investigator (PI) or co-investigator assists with managing requests; (2) DACs in consortia, where the PIs are assisted by legal and/or ethics experts; and (3) DACs attached to institutes, which function centrally and whose members include the necessary experts. The second type of DAC may have a data access officer to manage regular requests, and an advisory committee functioning at a higher level that includes the PIs and legal and ethics experts to advise on more difficult cases or to establish policy. The major reason for data access control is to protect the privacy of data subjects and foster public trust, together with protecting the professional interests of the data creators.¹³ A DAC requires a framework for good governance to guide data access decisions.¹⁴

Solutions to improve ethics review of data science research

Ferretti et al.⁴ propose several reforms to improve the ethics review of data science and big data research. These include regulatory reforms such as new guidance for RECs in the form of flow charts on the ethics of such research, procedural reforms with new working and assessment tools, upskilling of REC members in big data knowledge, and inclusion of subject experts as members of the REC or consulting external experts for specific issues.⁴ They also suggest the inclusion of other ethics committees, complementary to RECs in the review process, "to assess big data research and provide sectorial accreditation to researchers". The advantages of this would be to lessen the load on RECs and to obtain expert opinions for big data studies. The disadvantages are potential inefficiency of review procedures, erosion of responsibility of the REC, and questions about the role of the REC in big data ethics review.⁴

Establishing an ethics advisory committee for REDSSA

The research strategy for the REDSSA project included early integration of ELSI into data science research conducted at the research hub. Consequently, REDSSA bioethicists attend weekly meetings of the research hub and are immersed in the scientific and ethics challenges arising on the ground. Advice is provided in real time as issues emerge during the conduct of research. In addition, a Data Science Ethics Advisory Committee was established as a more formal structure to inform broader ethics questions in the research hub. This Committee is interdisciplinary, with representation from ethics, law, data science, social science and the community. All the members of this Committee are part of the REDSSA project and are funded through the grant. No additional funding exists for the EAC. It does not function as a REC or an institutional review board. Instead, substantive ethics issues within the hub or its projects can be referred for deliberation. Such referrals can occur before or after submission to a REC or institutional review board to allow deliberation on specific ELSI concerns.² As the REDSSA project is linked to the INFORM-Africa research hub in the DS-I Africa Consortium, referrals from the research hub to the EAC do not require additional funding or attract extra charges.

The terms of reference of the EAC were initially written by a small group of REDSSA team members affiliated with the Centre for Medical Ethics and Law at Stellenbosch University, and then refined by discussion at the first meeting of the EAC, held virtually, and subsequently via two rounds of email communication. The functions of the EAC are advisory, consultative and educational, and include development of recommendations and policy review.

The purpose of the EAC is threefold:

1. To promote and uphold respect for the dignity and rights of research participants/data donors/data subjects/data sources of the INFORM-Africa project.

2. To act as a consultative and resource base on data science ethical issues primarily for the research hub and various stakeholders.
3. To help develop ELSI policies and guidelines with relevant consultants as required.

The responsibilities of the EAC are:

1. To fulfil an advisory and consultative role with respect to data science ethics dilemmas in the research environment.
2. To advise on the development of protocols relating to research ethics dilemmas in conjunction with relevant researchers.
3. To clarify concepts around ethics pertaining to surveillance and research using surveillance data.

The interdisciplinary membership of the EAC includes the following: chairs; vice-chairs; data scientists; bioethicists from sub-Saharan Africa; external bioethicists; research ethics committee members; researchers knowledgeable about COVID-19 and HIV research; legal experts; INFORM-Africa Community Advisory Board members; independent members; and the Secretariat. Each position has two appointees (a primary and an alternative). The Chairs are appointed by the PI from the REDSSA management team for an initial period of 1 year, renewable annually thereafter up to a maximum of 3 years. The Chairs are assisted by the Vice-Chairs.

The PI and co-PI of the REDSSA project, together with the NIH Scientific Officer overseeing the REDSSA project, are ultimately responsible for the oversight of the EAC. The members of the EAC are required to declare conflicts of interest and recuse themselves from the discussion as appropriate.

When necessary, ad hoc members – such as relevant bioethics, legal or data science representatives, and experts in specific fields – are consulted on a case-by-case basis.

The REDSSA EAC is an advisory/consultative body and not a decision-making committee. Research requiring ethics approval must be reviewed by an institutional or national REC. The REC could consult experts in the ethics of data science if necessary, including the EAC which would be facilitated by the INFORM-Africa PI.

All committee deliberations remain confidential. The EAC recommendations are formulated via consensus. The recommendations are minuted and forwarded to the INFORM-Africa PI, who may share them with the REC. If the INFORM-Africa PI disagrees with the recommendation of the EAC, the decision and the reason(s) for that decision should be communicated to the EAC in writing. The EAC members discuss how to manage this on a case-by-case basis, but a report is forwarded to the NIH Scientific Officer. The final regulatory approval and oversight of research projects lies with the REC. If the REC disagrees with the advice provided by the EAC, a meeting should be convened between the Chair of the EAC, the Chair of the REC and the PI to discuss the project. Likewise, if a PI seeks advice from the EAC about an approved project, such advice should be communicated to the REC.

A two-part framework for assessing the ethical implications of big data health-related research projects is suggested by investigators from the United Kingdom Research Study into Ethnicity and COVID-19 Outcomes in Healthcare Workers (UK-REACH) study group. Firstly, "the specific legal and ethical issues raised by the project's aims and methods" must be identified; and, secondly, those issues must be addressed to promote positive aspects of the project (such as justice and respect for persons) while simultaneously modifying or eradicating negative aspects (such as stigmatisation, legal violations, and aggravation of social inequality and injustice).¹⁵ A framework for assessing the ethics of big data research is described by Xafis and co-authors¹⁶ in which they detail the following steps: (1) Identify and explicate the moral and ethical issue(s); (2) Identify the relevant values (both substantive and procedural) pertaining to the issue or problem; (3) Identify potential solutions and actions; (4) Evaluate the relative ethical importance of the different options; (5) Select the option that carries the greatest ethical weight while considering the roles and influence of the decision-makers in the group; and (6) Communicate



the decision clearly to all stakeholders. The REDSSA EAC deliberates on potential ethics frameworks that best suit its decision-making needs.

Conclusion

Data science has the potential to enhance health-related knowledge, particularly in the field of public health. However, big data projects are subject to ethical and legal concerns and RECs may experience challenges with the review process as data ethics is an emergent discipline in sub-Saharan Africa. EACs may play a supportive role in big data research for both researchers and RECs. The REDSSA EAC provides one viable way of fulfilling an advisory role that can better support researchers and RECs involved in big data research.

Competing interests

There are no competing interests to declare. The authors are all members of the REDSSA EAC.

References

1. Dhar V. Data science and prediction. *Communications of the ACM*. 2013;56(12):64–73. <https://doi.org/10.1145/2500499>
2. Moodley K, Rennie S. Research strategy of REDSSA. In: Unpublished NIH grant application.
3. Munyaradzi M. New funding set to transform data science research and innovation. *Nat Afr*. 15 November 2021 [cited 2022 Sep 02]. <https://doi.org/10.1038/d44148-021-00111-3>
4. Ferretti A, Ienca M, Sheehan M, Blasimme A, Dove ES, Farsides B, et al. Ethics review of big data research: What should stay and what should be reformed? *BMC Med Ethics*. 2021;22, Art. #51. <https://doi.org/10.1186/s12910-021-00616-4>
5. Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of big data challenges and analytical methods. *J Bus Res*. 2017;70:263–286. <https://doi.org/10.1016/j.jbusres.2016.08.001>
6. Doerr M, Meeder S. Big health data research and group harm: The scope of IRB review. *Ethics Hum Res*. 2022;44(4):34–38. <https://doi.org/10.1002/eahr.500130>
7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447–453. <https://doi.org/10.1126/science.aax2342>
8. Protection of Personal Information Act (POPIA) [homepage on the Internet]. c2019 [cited 2022 Sep 02]. Available from: <https://popia.co.za/>
9. Promotion of Access to Information Act 2 of 2000 (PAIA) [webpage on the Internet]. c2000 [cited 2022 Sep 02]. Available from: <https://www.gov.za/documents/promotion-access-information-act>
10. Adam D, Kramer I, Guillory JE, Hancock JT. Experimental evidence of massive scale emotional contagion through social networks. *Proc Natl Acad Sci USA*. 2014;111(24):8788–8790. <https://doi.org/10.1073/pnas.1320040111>
11. Vitak J, Shilton K, Ashktorab Z. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*; 2016 February 27 – March 02; San Francisco, CA, USA. New York: Association for Computing Machinery; 2016. p. 941–953. <http://dx.doi.org/10.1145/2818048.2820078>
12. Facebook faces UK probe over emotion study. *BBC World*. 2014 July 02 [cited 2022 Sep 02]. Available from: <https://www.bbc.com/news/technology-28102550>
13. Shabani M, Borry P. “You want the right amount of oversight”: interviews with data access committee members and experts on genomic data access. *Genet Med*. 2016;18(9):892–897. <https://doi.org/10.1038/gim.2015.189>
14. De Vries J, Abayomi A, Littler K, Madden E, McCurdy S, Ouwe Missi Oukem-Boyer O, et al. Addressing ethical issues in H3Africa research – the views of research ethics committee members. *The HUGO J*. 2015;9, Art. #1. <https://doi.org/10.1186/s11568-015-0006-6>
15. Reed-Berendt R, Dove ES, Pareek M. The ethical implications of big data research in public health: “Big Data Ethics by Design” in the UK-REACH Study. *Ethics Hum Res*. 2022;44(1):2–17. <https://doi.org/10.1002/eahr.500111>
16. Xafis V, Schaefer GO, Labude MK, Brassington I, Ballantyne A, Lim HY, et al. An ethics framework for big data in health and research. *Asian Bioeth Rev*. 2019;11:227–254. <https://doi.org/10.1007/s41649-019-00099-x>

**AUTHOR:**

Gonasagrie Nair¹ 
 Theresa L. Burgess^{1,2} 
 Adetayo E. Obasa¹ 
 Sharon Kling^{1,3} 
 Shenuka Singh^{1,4} 

AFFILIATIONS:

¹Centre for Medical Ethics and Law, WHO Bioethics Collaborating Centre, Department of Medicine, Stellenbosch University, Cape Town, South Africa

²Division of Physiotherapy, Department of Health and Rehabilitation Sciences, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

³Department of Paediatrics and Child Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

⁴Discipline of Dentistry, School of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:

Gonasagrie Nair

EMAIL:

lulu.nair13@gmail.com

HOW TO CITE:

Nair G, Burgess TL, Obasa AE, Kling S, Singh S. Revisiting community engagement methods in the context of data science research and big data use in South Africa. *S Afr J Sci.* 2023;119(5/6), Art. #14723. <https://doi.org/10.17159/sajs.2023/14723>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

KEYWORDS:

data science, big data, community engagement, data ownership, citizen science

PUBLISHED:

30 May 2023

FUNDING:

US National Institutes of Health (U01MH127704)



Revisiting community engagement methods in the context of data science research and big data use in South Africa

Significance:

Effective community engagement for the use of large data sets in health research is faced with challenges similar to those in investigator-driven research. The scope of community engagement has evolved in high-income countries to embrace citizen science by communities and regulators to build trust in data science research. In South Africa and other low- and middle-income countries, with varying levels of literacy and the influence of pre-existing beliefs and past negative experiences with research, advisory committees of diverse stakeholder composition still have a role to play in protecting the rights of researched communities.

Introduction

Access to existing large diverse data sets plays an important role in drug development research, precision medicine, diagnostic imaging, artificial intelligence (AI) platforms, medical decision support systems, and managing public health emergencies.¹

Large volumes of genomic and phenotypic health-related data are collected from various sources including computers, smartphones, tablets, and wearable devices. Electronic health data are also collected by medical insurance companies in the private health sector and from public health data bases.² These data are categorised as “big data”, given that the information originates from a variety of sources, is of large volume, and is processed at high speeds.³ Data science, which makes use of big data, is defined as the “study of the extraction of knowledge from data” and differs from statistics because the data sources and formats vary.⁴ Data may be presented as numbers, text, images, or video. A multidisciplinary approach, involving computer scientists, sociologists, clinicians and epidemiologists, is required for the analysis and interpretation of health data.⁴

Importantly, access to large pre-existing data sets may increase the efficiency of research by avoiding potential duplication and overburdening research participants and increasing statistical power and the generalisability of study findings.

While the value of data science cannot be underscored, there may be a lack of awareness among the public⁵, especially in low- and middle-income countries (LMICs), that these data are being collected and shared with in-country researchers or with researchers in other countries. At the same time, there are ethical and legal challenges associated with health data science research that need to be considered, particularly to maintain individual patient and community trust in research. This emphasises the need for sustained community and stakeholder engagement by researchers. In this Perspective, we therefore highlight the ethical implications of big data research, the use of community and stakeholder engagement to build data science literacy and public trust, the limitations of traditional community engagement, especially in LMICs and South Africa, and how these identified challenges could be addressed.

Ethical considerations

The use of routine clinical data for research purposes results in a blurring of boundaries between clinical care and research, and raises questions around data ownership, patient privacy, and autonomy.⁶

Some of the possible harms to research participants could include violation of privacy, and stigma based on health-seeking behaviour and health patterns of communities. Additionally there could be secondary discrimination from data sets used to generate algorithms, which could lack diversity and thereby introduce bias in the interpretation of the study findings.⁷ Apart from issues of privacy and confidentiality, questions of data ownership arise if health-related data have already been collected as part of routine clinical care and have subsequently been shared for research purposes.

From a consent perspective, while clinical research allows for broad consent or tiered consent for the storage of samples and use of data for future related research, consent for clinical care is typically only for specific clinical management. Additionally, data could be accessed purely for clinical purposes and later re-purposed for research, yet consent was only obtained for the clinical services provided.

Legislation related to consent for data use in South Africa adds to this unclear picture for big data research. As per the *South African National Health Act*⁸ and Health Professionals Council of South Africa guidelines⁹, a patient has the right to expect that their health-related data will be confidential and that sharing of this information will only occur after their consent has been obtained. In contrast, the *Protection of Personal Information Act of 2013*¹⁰ allows for the sharing of special information, that is, health data, if these are de-identified.

The need for community engagement

The highlighted ethico-legal considerations reiterate the need for better engagement with affected individuals and communities. Community engagement is seen as a vital process to optimise public trust in the research process.^{11,12} The levels of community engagement include stakeholder input, consultation, collaboration, and shared leadership, with shared leadership being the most collaborative and stakeholder input being least so.¹³



However, community engagement may also result in unintended consequences. Although researchers may have good intentions to initiate meaningful conversations with research communities around the research and protocol development, such efforts may be misinterpreted and intentions may be misconstrued, thereby leading to mistrust between the researched communities and the research teams involved.¹⁴

There are several reasons related to non-participation in research, including a lack of understanding of the research, and considering the research irrelevant either because of a feeling that it does not address the needs of communities or by misinterpreting it as being elitist. Past negative experiences and/or cultural barriers may also play a role.⁵ Community engagement has assisted in identifying these reasons and addressing them in a culturally sensitive manner to allow for research participation, and has thus been beneficial both to the researcher in allowing successful trial implementation and to the community in addressing health priorities. The success of research is dependent not only on the occurrence of community engagement, but as communities become more familiar with the advantages and pitfalls of participating in research, by the extent of community engagement.⁵

Limitations in the current frameworks to guide community engagement in data science

Frameworks for participatory research have been developed to promote authentic community engagement through a sense of ownership and to meet funder/s' requirements.¹¹ These frameworks, although useful, are not formally recognised by policymakers or research ethics committees (RECs). In settings in which community engagement is not mandatory or required, original participatory engagement intentions fall away and, due to mistrust and disillusionment, communities with capacity shift from collaborative engagement to one of capacitation, where the community identifies research priorities, implements research, analyses data and disseminates results.⁵ This shift to the 'capacitation' model, which is being recommended and adopted in high-income countries^{5,6,13}, may not be feasible in LMICs due to the scarcity of human and financial resources for sustainability.⁵

Additionally, strategic plans for health research highlight the value of community engagement⁵, but there are no identified processes to enforce its implementation. The Emanuel, Wendler, and Grady framework has been adopted by some RECs globally for evaluation of ethical, social, and scientific robustness of proposed research and a 2008 revision included collaborative partnership for the first time.¹⁶ In spite of this recommendation that collaborative partnership is one of the eight factors considered in ethics review, a South African study indicated collaborative partnership was less likely to raise queries during the review process in comparison to the other factors, if considered at all.¹⁷

In comparison to health research focused on a specific disease or condition, the community in health data science research is not as clearly defined. If we consider the diverse sources of big data, questions around who constitutes the community and stakeholders arise. All users of social media, owners of a cell phone or wearable device and those who seek health care in either the private or public sectors may be considered the research community. However, narrowing the health-seeking behaviour to a particular health condition such as Human Immunodeficiency Virus (HIV), tuberculosis (TB), or to a rare disease will facilitate the identification of the community even in big data research.⁵

Traditional mechanisms for community engagement within the context of health research and clinical trials have involved community advisory boards (CABs). CABs generally constitute influential community members who serve as a bridge between community members and researchers, thus ensuring optimal study implementation and protection of the rights of communities.

The need for a paradigm shift

Funders and governance structures of clinical research in high-income countries with stringent data protection and protection of personal information laws are requiring comprehensive involvement of patients, as owners of their health data, in research.^{6,12} This has seen the advent

of greater degrees of citizen science, with patients deciding for which projects their data will be used and uploading data directly to databases.⁶

Our proposal

Ferretti et al.¹⁸ state that the use of big data excludes engagement with study participants, but we are of the opinion that the CAB model for community engagement would still be appropriate in addition to models that encompass more participatory methods of community engagement. One such participatory approach would be 'crowdsourcing' – characterised by large groups of experts and non-experts from diverse backgrounds providing solutions to a problem. This is an approach that can be used in clinical health research.¹⁹

Where there are well-defined accessible communities, the ethical principles that govern research can be adhered to through CABs and REC review. CAB review of consent forms to ensure social and cultural appropriateness and advice on the consent process ensures respect for the autonomy of study participants. However, data science research involves the re-use of pre-existing data sets so consent is not sought from individuals or communities, but new mechanisms, such as dynamic or portable consent made possible through online platforms, may be a solution.²⁰ Protocol review, prior to study implementation, ensures that principles of beneficence/non-maleficence and justice are adhered to. This ensures that ineligible participants are linked to care, that benefits outweigh risks, that study participants are not required to waive any of their rights, and that post-trial access and benefit sharing mechanisms are in place to ensure access to successful interventions to those who endured the risks of study participation. While this approach may be regarded as paternalistic and non-empowering, it still has a role in research-naïve communities and many indigent communities in LMICs in which individuals may be coerced into study participation. Ferretti et al.⁷ note that RECs, which often include a community representative, may struggle to apply existing governance frameworks or regulatory tools for ethics review for data science research because data are anonymised and the research does not involve interaction with research participants. We agree with these authors. There is a need to build further capacity in RECs with regard to the review of protocols related to big data science. Such capacity building should include ongoing educational training as well as ensuring that RECs include members with appropriate skills and experience in this evolving area of research.

Conclusion

Community engagement for health research utilising large data sets should include public engagement or 'data science citizenship'. However, there is a role for traditional engagement to foster trust and transparency through CABs where stakeholders are existing research participants. RECs should be empowered to critically evaluate community engagement in data science health research. In-country regulations for data ownership and sharing should align with each other for easy interpretation by communities and researchers, both local and international.

Competing interests

We have no competing interests to declare.



References

1. Mehta N, Shukla S. Pandemic analytics: How countries are leveraging big data analytics and artificial intelligence to fight COVID-19? *SN Comput Sci.* 2022;3, Art. #54. <https://doi.org/10.1007/s42979-021-00923-y>
2. Vayena E, Blasimme A. Biomedical big data: New models of control over access, use and governance. *J Bioeth Inq.* 2017;14:501–513. <https://doi.org/10.1007/s11673-017-9809-6>
3. Benedek A, Molnar G, Szuts Z. Practices of crowdsourcing in relation to big data analysis and education methods. In: 2015 IEEE 13th International Symposium on Intelligent Systems and Informatics (SISY); 2015 September 17–19; Subotica, Serbia. IEEE; 2015. p. 167–172. <https://doi.org/10.1109/SISY.2015.7325373>
4. Dhar V. Data science and prediction. *Commun ACM.* 2013;56:64–73. <https://doi.org/10.1145/2500499>



5. Grayson S, Doerr M, Yu JH. Developing pathways for community-led research with big data: A content analysis of stakeholder interviews. *Heal Res Policy Syst.* 2020;18, Art. #76. <https://doi.org/10.1186/s12961-020-00589-7>
6. Winickoff DE, Jamal L, Anderson NR. New modes of engagement for big data research. *J Responsible Innov.* 2016;3:169–177. <https://doi.org/10.1080/23299460.2016.1190443>
7. Ferretti A, Ienca M, Velarde MR, Hurst S, Vayena E. The challenges of big data for research ethics committees: A qualitative Swiss study. *J Empir Res Hum Res Ethics.* 2022;17:129–143. <https://doi.org/10.1177/15562646211053538>
8. National Health Act 61 of 2003 (NHA) [webpage on the Internet]. c2003 [cited 2022 Sep 05]. Available from: <https://www.gov.za/documents/national-health-act>
9. Health Professions Council of South Africa (HPCSA). General ethical guidelines for good practice in telehealth: Booklet 10 [document on the Internet]. c2021 [cited 2022 Sep 05]. Available from: https://www.hpcsablogs.co.za/wp-content/uploads/2022/08/Booklet-10_Telehealth_Dec_2021_.pdf
10. Protection of Personal Information Act (POPI Act). Available from: <https://popia.co.za/>
11. Moodley K, Beyer C. Tygerberg Research Ubuntu-Inspired Community Engagement Model: Integrating community engagement into genomic biobanking. *Biopreserv Biobank.* 2019;17:613–624. <https://doi.org/10.1089/bio.2018.0136>
12. US National Institute for Health Research. Going the extra mile: Improving the nation's health and wellbeing through public involvement in research [document on the Internet]. c2015 [cited 2022 Sep 05]. Available from: https://www.researchgate.net/publication/279180522_%27Going_the_extra_mile_improving_the_nation%27s_health_and_wellbeing_through_public_involvement_in_research%27
13. Patient-Centered Outcomes Research Institute (PCORI). The value of engagement [webpage on the Internet]. c2018 [cited 2022 Sep 06]. Available from: <https://www.pcori.org/engagement/value-engagement>
14. Wilkinson A, Slack C, Thabette S, Salzwedel J. "It's almost as if stakeholder engagement is the annoying 'have-to-do'...": Can ethics review help address the "3 Ts" of tokenism, toxicity, and tailoring in stakeholder engagement? *J Empir Res Hum Res Ethics.* 2022;17:292–303. <https://doi.org/10.1177/15562646221078415>
15. South African Department of Health. National Health Research Strategy: Research priorities for SA – 2021–2024. No date [cited 2022 Aug 26]. Available from: <https://www.health.gov.za/wp-content/uploads/2022/05/National-Health-Research-Priorities-2021-2024.pdf>
16. Emanuel EJ, Wendler DD, Grady C. In: Emanuel EJ, Grady C, Crouch RA, Lie RK, Miller FG, Wendler DD, editors. *The Oxford textbook of clinical research ethics.* Oxford/New York: Oxford University Press; 2008. p. 123–135.
17. Tsoka-Gwegweni JM, Wassenaar DR. Using the Emanuel et al. framework to assess ethical issues raised by a biomedical research ethics committee in South Africa. *J Empir Res Hum Res Ethics.* 2014;9:36–45. <https://doi.org/10.1177/1556264614553172>
18. Ferretti A, Ienca M, Sheehan M, Blasimme A, Dove ES, Farsides B, et al. Ethics review of big data research: What should stay and what should be reformed? *BMC Med Ethics.* 2021;22, Art. #51. <https://doi.org/10.1186/s12910-021-00616-4>
19. Tucker JD, Day S, Tang W, Bayus B. Crowdsourcing in medical research: Concepts and applications. *PeerJ.* 2019;7, e6762. <https://doi.org/10.7717/peerj.6762>
20. Cato KD, Bockting W, Larson E. Did I tell you that? Ethical issues related to using computational methods to discover non-disclosed patient characteristics. *J Empir Res Hum Res Ethics.* 2016;11:214–219. <https://doi.org/10.1177/1556264616661611>



AUTHORS:
Suzanne Day¹ 
Stuart Rennie^{2,3} 

AFFILIATIONS:
¹Department of Medicine, Division of Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
²Department of Social Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
³UNC Center for Bioethics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

CORRESPONDENCE TO:
Suzanne Day

EMAIL:
suzanne.day@med.unc.edu

HOW TO CITE:
Day S, Rennie S. Data science research in sub-Saharan Africa: Ethical considerations in crowdsourcing for community engagement. *S Afr J Sci.* 2023;119(5/6), Art. #14911. <https://doi.org/10.17159/sajs.2023/14911>

ARTICLE INCLUDES:
 Peer review
 Supplementary material

KEYWORDS:
data science, ethics, crowdsourcing, community engagement, sub-Saharan Africa

FUNDING:
US National Institute of Mental Health of the National Institutes of Health (U01MH127704)

PUBLISHED:
30 May 2023

Data science research in sub-Saharan Africa: Ethical considerations in crowdsourcing for community engagement

Significance:

There is an increasing movement to 'digitise' health-related data on the African continent, and to improve local health and health systems using cutting-edge data analytics. While these big data initiatives may be beneficial, and engagement is needed to help maintain public trust in data science, the introduction of new digital technologies raises ethical concerns and challenges for engagement. In this Commentary, we focus on the ethics of using crowdsourcing as an approach to increasing community engagement in data science. We map out key areas of ethical concern related to data science and argue that crowdsourcing serves as a promising strategy for identifying ways in which communities can become more engaged in data science initiatives.

The growth in data science research in sub-Saharan Africa raises important ethical questions for the collection and use of 'big data' in this context, with particularly disparate implications for the most vulnerable and marginalised populations. While enhanced public involvement may be able to mitigate some of these risks, data science presents some unique barriers to community engagement efforts, including limited data literacy, lack of transparency in data collection and use, and little opportunity to 'opt out' from participation. The participatory approach of crowdsourcing offers a promising solution to address the critical need for community engagement. Crowdsourcing involves inviting a group to contribute solutions to a problem, and then publicly sharing the results for implementation. By crowdsourcing stakeholder ideas for innovative ways to enhance public involvement in data science research, the Research for Ethical Data Science in Southern Africa (REDSSA) project is leading the efforts to close the community engagement gap. Promising strategies that emerge from these efforts will ultimately help to shape more ethical and equitable data science research in Africa as this field continues to grow.

Ethical issues in data science in sub-Saharan Africa

Data science interventions developed through the collection and analysis of 'big data' have been touted for the possibility to address some of the most pressing social issues facing low- and middle-income countries (LMICs). Big data may enable decision-makers in LMICs to better understand patterns of human migration, track deforestation, estimate poverty among a population, and predict epidemic outbreaks.¹ In the context of sub-Saharan Africa (SSA), data science research and the collection of big data is an emerging field with the potential for rapid expansion, aided through increased use of digital social networks, availability of Internet access, and mobile smartphone usage.² The rise of data science research could have a number of beneficial applications across SSA nations, such as serving to enhance public health through reporting and containment of disease, establishing early outbreak warning systems, priming healthcare providers for timely response, prompting strategic healthcare planning, and mobilising domestic and international stakeholder support.³ While these applications have the potential for positive impact on public health and development, guidance to inform the ethical collection and use of big data has not kept pace with the growth in data science approaches in LMICs.⁴

A data justice perspective provides a potential framework for viewing the ethical concerns of data science in SSA. Data justice is an approach that borrows social justice concepts and applies them to pose ethical questions of rights, fairness and protections in the context of big data collection and use.⁵ From a data justice perspective⁶, there are three conditions to consider in order for data-driven approaches to be ethically sound: non-discrimination (i.e. the ability to challenge biased data and avoid discrimination), engagement in the technology (i.e. the ability to make autonomous decisions about how one's data are collected, shared and used), and visibility (i.e. the ability to be represented in the data while maintaining privacy protections). When applied to data science in the SSA context, these facets of data justice raise multiple ethical red flags.

First, pertaining to the condition of non-discrimination: it is unclear whether and/or to what extent algorithms in growing use in the SSA context based on the collection of big data are being checked for bias, and what potential harms may result from interventions developed based on biased models. For example, while machine learning predictive models of HIV risk in SSA have the potential to inform testing and other prevention services, predictive models may be biased in terms of which populations are identified as being at elevated HIV risk, which can in turn result in further unintended harms via discrimination and heightened monitoring.⁷ Second, pertaining to the condition of engagement in the technology: there are few opportunities to make autonomous decisions about how one's data are collected and used, and there are many ways that big data can be used by others for less-than-good intentions, including surveillance for population control and exclusion.⁸ In the SSA context, the growing ability to map human mobility using mobile phone geodata may be misused by governments to predict and prevent population migration in times of crisis.⁹ Third, regarding the condition of visibility: while ideally this condition would see the balance of equal representation with adequate protections, the potential risks associated with the growth of data science are unequally distributed; vulnerable and marginalised populations are at greater risk of being insufficiently represented¹⁰, as well as at risk of disproportionate government surveillance for criminalisation and control¹¹. In the SSA context, marginalised populations may be at particular risk given widespread legislation undermining encryption across SSA countries.¹²



The need for public involvement and community engagement in data science research has been increasingly recognised as essential for mitigating the above-noted risks and improving adherence to data justice principles.^{13,14} However, the topic of data science is not one that lends itself easily to established community engagement approaches that have developed for use in other fields. A lack of data literacy (in terms of understanding what kinds of data are collected, how they are collected, and with what purpose) has resulted in a growing rift between the elite (researchers) who are further 'in the know' and a largely unaware (or uninformed) public.¹⁵ In addition, the terms of participation in data science research do not follow typical research participation processes: one cannot exactly 'opt out' of the collection of their data via mobile phone technology, for example, without essentially opting out of dominant forms of social connectivity and economic systems. Good participatory practices for community engagement identify those who are 'participants' or potential participants in the research as a key group for engagement¹⁶; yet in the case of data science research, what choice does one truly have? The data that are collected from communities (some of it personal, even if de-identified) are collected without consent, or via consent processes that do not follow the typical informed consent processes used in other fields of research. While the community-led call-to-action for research to produce 'nothing about us, without us' has been essential for shaping engagement processes in other fields¹⁷, this approach has not similarly been a part of traditional public health surveillance.

Despite these challenges, greater community engagement is urgently needed in research involving big data for the sake of better data science and more equitably beneficial research outcomes. In addition to helping to bridge the information gap between data scientists and the public, community engagement can help data science research to better incorporate the values and interests of the public that are not readily captured in the data.¹³ Narrowing the information gap may also help preserve community trust in research institutions and mitigate misinformation about data science as its activities come to be more widely known. In addition, community engagement may help to address some of the unanticipated negative consequences of data science research and potential vulnerable points that are missed in algorithms by providing greater insights into community members' perceptions of risks and potential solutions for mitigating them.^{18,19} There is furthermore a need for community engagement efforts that are appropriate and feasible for use within the unique social, cultural, economic, and political contexts of data science research in the SSA context.²⁰ While there is limited work being done on community engagement for big data research globally, approaches developed in high-income country contexts may not be easily transferrable into SSA settings, for just as there are unique data justice concerns in SSA, so too are there potentially unique engagement needs.

Herein lies a complex dilemma for data science researchers in SSA seeking to enhance community engagement processes: what would be promising approaches for engaging the community on data science research when it is a topic that is not widely understood, when its processes are largely opaque, when the 'community' of affected stakeholders may be millions of people, and when people who are technically participating in the research via the collection of their data have little real choice to 'opt out', shape or impact the collection and use of their data? Furthermore, how can we tailor engagement approaches to the unique contexts of data science research in SSA? Finally, how can we ensure that engagement approaches for data science research are developed in ways that would be acceptable and of interest to the communities we seek to involve?

Stakeholder-driven solutions for community engagement

One promising approach for addressing the above-noted dilemmas may lie in crowdsourcing. Crowdsourcing involves inviting a group of experts and non-experts to contribute creative solutions to a problem, and then sharing the results with the public.²¹ Drawing on the concept of crowd wisdom, crowdsourcing is premised upon the idea that one need not be an 'expert' to contribute great ideas; thus, as a methodology for intervention development, crowdsourcing is well positioned to disrupt

the elitism that communities may experience as a barrier to engagement in data science research.²²

Crowdsourcing also serves a dual-purpose approach to problem-solving. It is both a way to gain promising stakeholder-driven ideas for potential implementation, and participating in crowdsourcing also serves as a form of community engagement in spreading awareness about a particular issue, involving communities/relevant stakeholders as key contributors to intervention development, and disseminating potential solutions at the community level.²³⁻²⁶ It is an inherently participatory process for intervention development, with solutions emerging through a 'bottom-up' community-driven process rather than 'top-down' researcher-led designs. Additionally, interventions developed through crowdsourced community ideas have been shown to be effective in addressing community concerns and priorities. Crowdsourcing has been successfully used to develop messaging to encourage community engagement in HIV cure research²⁷, to promote HIV testing among at-risk populations²⁸⁻³⁰, and to obtain feedback from community members on clinical trial designs²⁴. With demonstrated effectiveness in clinical trials³¹, crowdsourcing approaches have been used extensively by health and scientific research organisations as an innovative approach to problem solving, including the US National Academies of Sciences, Engineering and Medicine³², the US National Institutes of Health Research Office of Behavioral and Social Science Research³³, and *The Lancet* Healthy Cities Commission³⁴.

It may not be possible to crowdsource ideas that could solve all the dilemmas involved in data science research. For example, while members of the public could participate in crowdsourcing ideas to change how health surveillance data are collected and used, it is unlikely that such solutions would be implementable without being accompanied by substantial changes in the regulatory sphere. Furthermore, with community engagement for data science still in its infancy in SSA, it is unlikely that community members have sufficient understanding of how health surveillance data are currently collected and used to be able to consider how these processes may be intervened upon in ways more aligned with a data justice approach. Improving the baseline understanding of the wider public on the topic of data science could potentially help to improve the ability of lay communities to engage in crowdsourcing initiatives on this topic. For example, one strategy being examined by the REDSSA team is providing patients with infographics and pamphlets explaining how health data are collected and used for data science purposes. Efforts to make the topic of data science more broadly understood are essential for boosting participation in crowdsourcing efforts, and subsequently the quality of crowdsourced solutions; low participation in crowdsourcing runs the risk of producing designs based on only a small fraction of the potential pool of stakeholders, calling into question the extent to which the crowdsourced product reflects community concerns.³⁵

In contrast, crowdsourcing ideas for how to improve community engagement in data science research is a more promising possibility – one which avoids the need for in-depth understanding of data science. By instead asking stakeholders to contribute creative ideas for community engagement about data science, drawing on their own experiences, values and priorities regarding the collection and use of big data, we can develop engagement strategies that are reflective of and responsive to community concerns.²¹ In addition, community-driven ideas for engagement approaches in data science research may be potentially more effective than top-down designs, and would be grounded in the actual concerns/gaps identified by the people we need to hear from in said engagement processes, i.e. those who can identify vulnerabilities/unintended negative consequences, if offered the opportunity to participate in a meaningful way. In this way, crowdsourced solutions for overcoming the challenges identified with community engagement for data science (e.g. ideas for how to increase data literacy, and strategies to enhance transparency in data collection and use) would be developed by and for those communities most impacted by said challenges.

There are, however, some important caveats and limitations to consider. Crowdsourcing is not invulnerable to similar biases, exclusions and disproportionate negative impact as noted above regarding data science itself. Who we engage with to contribute ideas, and how we engage them, will substantially impact the kinds of ideas that are contributed to



a crowdsourcing approach.³⁶ In crowdsourcing ideas for how to enhance community engagement in data science research, there is much to consider regarding how ‘even’ the playing field is for participation in crowdsourcing: while not requiring expert insights into how data are collected and used, communities may still find it a challenging topic to consider given that data science is a topic that may feel highly irrelevant to or removed from people’s daily lives given their heretofore lack of inclusion in decision-making processes.³⁷ Crowdsourcing community engagement strategies therefore will require careful consideration to ensure that potential participants are sufficiently informed to feel like they can contribute an idea, as well as to feel like their contributions will be meaningful. In addition, crowdsourcing in SSA presents several unique considerations, including language diversity, a highly heterogeneous population spread over vast geographic areas, and the limits of implementing digital strategies in resource-constrained settings. However, successful crowdsourcing projects in diverse LMIC settings provide methodological blueprints for mitigating some of these challenges.^{29,30,38,39}

Engagement for ethical data science research

Crowdsourcing ideas for engagement strategies in data science research would be one small step towards addressing a heretofore overlooked aspect of the field: the lack of meaningful mechanisms for obtaining community input on ethical issues in the collection and use of big data. While crowdsourcing is not the only way to develop engagement strategies and has its own ethical challenges³⁶, it nonetheless offers a participatory starting point for developing meaningful engagement processes. Furthermore, while ethical challenges of crowdsourcing are fairly well known and there are emergent best practices to help mitigate them, the ethical issues related to data science as they play out in SSA is an as-yet little explored landscape. Increased social science research, both qualitative and quantitative, is needed to measure current community awareness of ‘big data’ research in SSA, and explore concerns that communities have in relation to its many forms. Engagement strategies are urgently needed now to elucidate these challenges more clearly if we are to have a hope of shaping the growing data science field in ways more aligned with the pillars of data justice. To this end, the REDSSA project is leading the way in crowdsourcing stakeholder-driven solutions to the problem of a lack of community engagement in research using big data.⁴⁰ The results of this study will have immediate practical use as new data science initiatives are being increasingly implemented across SSA.⁴¹ It is imperative for the ethical conduct of data science in Africa that innovations in community engagement keep pace with ‘big data’ research and its novel applications.

Acknowledgements

Research reported in this publication was supported by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

We have no competing interests to declare.

References

1. Chandy R, Hassan M, Mukherji P. Big data for good: Insights from emerging markets. *J Prod Innov Manage*. 2017;34(5):703–713. <https://doi.org/10.1111/jpim.12406>
2. Akinagbe A, Peiris KDA, Akinloye O. Prospects of big data analytics in Africa healthcare system. *Glob J Health Sci*. 2018;10(6):114. <https://doi.org/10.5539/gjhs.v10n6p114>
3. Amankwah-Amoah J. Emerging economies, emerging challenges: Mobilising and capturing value from big data. *Technol Forecast Soc Chang*. 2016;110:167–174. <https://doi.org/10.1016/j.techfore.2015.10.022>
4. Wyber R, Vaillancourt S, Perry W, Mannava P, Folaranmi T, Celi LA. Big data in global health: Improving health in low- and middle-income countries. *Bull World Health Organ*. 2015;93(3):203–208. <https://doi.org/10.2471%2FBLT.14.139022>

5. Dencik L, Hintz A, Redden J, Treré E. Exploring data justice: Conceptions, applications and directions. *Inform Commun Soc*. 2019;22(7):873–881. <https://doi.org/10.1080/1369118X.2019.1606268>
6. Taylor L. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data Soc*. 2017;4(2). <https://doi.org/10.1177/2053951717736335>
7. Nichol AA, Bendavid E, Mutenherwa F, Patel C, Cho MK. Diverse experts’ perspectives on ethical issues of using machine learning to predict HIV/AIDS risk in sub-Saharan Africa: A modified Delphi study. *BMJ Open*. 2021;11(7), e052287. <https://doi.org/10.1136/bmjopen-2021-052287>
8. Bircan T, Korkmaz EE. Big data for whose sake? Governing migration through artificial intelligence. *Humanit Soc Sci*. 2021;8(1):241. <https://doi.org/10.1057/s41599-021-00910-x>
9. Taylor L. No place to hide? The ethics and analytics of tracking mobility using mobile phone data. *Environ Plan D Soc Space*. 2015;34(2):319–336. <https://doi.org/10.1177/0263775815608851>
10. Gilman M, Green R. The surveillance gap: The harms of extreme privacy and data marginalization. *NYU Rev L Soc Change*. 2018;42:253.
11. Valentine S. Impoverished algorithms: Misguided governments, flawed technologies, and social control. *Fordham Urb LJ*. 2019;46:364.
12. CIPESA. How African governments undermine the use of encryption [document on the Internet]. c2021 [cited 2022 Sep 29]. Available from: https://cipesa.org/?wfb_dl=477
13. Aitken M, Tully MP, Porteous C, Denegri S, Cunningham-Burley S, Banner N, et al. Consensus statement on public involvement and engagement with data intensive health research. *Int J Popul Data Sci*. 2019;4(1):586. <https://doi.org/10.23889/ijpds.v4i1.586>
14. Johnson H, Davies JM, Leniz J, Chukwusa E, Markham S, Sleeman KE. Opportunities for public involvement in big data research in palliative and end-of-life care. *J Palliat Med*. 2021;35(9):1724–1726. <https://doi.org/10.1177/02692163211002101>
15. Bhargava R, Deahl E, Letouzé E, Noonan A, Sangokoya D, Shoup N. Beyond data literacy: Reinventing community engagement and empowerment in the age of data. New York: Data Pop Alliance; 2015 [cited 2022 Sep 29]. <https://hdl.handle.net/1721.1/123471>
16. UNAIDS/AIDS Vaccine Advocacy Coalition. Good participatory practice: Guidelines for biomedical HIV prevention trials. Geneva: UNAIDS; 2011.
17. Charlton JI. Nothing about us without us. Berkeley, CA: University of California Press; 1998.
18. Aitken M, Porteous C, Creamer E, Cunningham-Burley S. Who benefits and how? Public expectations of public benefits from data-intensive health research. *Big Data Soc*. 2018;5(2). <https://doi.org/10.1177/2053951718816724>
19. McWhirter RE, Critchley CR, Nicol D, Chalmers D, Whitton T, Otlowski M, et al. Community engagement for big epidemiology: Deliberative democracy as a tool. *J Pers Med*. 2014;4(4):459–474. <https://doi.org/10.3390%2Fjpm4040459>
20. Akondeng C, Njamnshi WY, Mandi HE, Agbor VN, Bain LE, Njamnshi AK. Community engagement in research in sub-Saharan Africa: Approaches, barriers, facilitators, ethical considerations and the role of gender – a systematic review protocol. *BMJ Open*. 2022;12(5), e057922. <https://doi.org/10.1136/bmjopen-2021-057922>
21. Tucker JD, Day S, Tang W, Bayus B. Crowdsourcing in medical research: Concepts and applications. *PeerJ*. 2019;7, e6762. <https://doi.org/10.7717%2Fpeerj.6762>
22. Grayson S, Doerr M, Yu J-H. Developing pathways for community-led research with big data: A content analysis of stakeholder interviews. *Health Res Policy Syst*. 2020;18(1):76. <https://doi.org/10.1186/s12961-020-00589-7>
23. Day S, Mathews A, Blumberg M, Vu T, Rennie S, Tucker JD. Broadening community engagement in clinical research: Designing and assessing a pilot crowdsourcing project to obtain community feedback on an HIV clinical trial. *Clin Trials*. 2020;17(3):306–313. <https://doi.org/10.1177/1740774520902741>
24. Day S, Mathews A, Blumberg M, Vu T, Mason H, Rennie S, et al. Expanding community engagement in HIV clinical trials: A pilot study using crowdsourcing. *AIDS*. 2020;34:1195–1204. <https://doi.org/10.1097/qad.0000000000002534>



25. Day S, Li C, Hlatshwako TG, Abu-Hijleh F, Han L, Deitelzweig C, et al. Assessment of a crowdsourcing open call for approaches to university community engagement and strategic planning during COVID-19. *JAMA Netw Open*. 2021;4(5), e2110090. <https://doi.org/10.1001/jamanetworkopen.2021.10090>
26. Mathews A, Farley S, Hightow-Weidman L, Muessig K, Rennie S, Tucker JD. Crowdsourcing and community engagement: A qualitative analysis of the 2BeatHIV contest. *J Virus Erad*. 2018;4(1):30–36. [https://doi.org/10.1016/S2055-6640\(20\)30239-9](https://doi.org/10.1016/S2055-6640(20)30239-9)
27. Mathews A, Farley S, Blumberg M, Knight K, Hightow-Weidman L, Muessig K, et al. HIV cure research community engagement in North Carolina: A mixed-methods evaluation of a crowdsourcing contest. *J Virus Erad*. 2017;3(4):223–228. [https://doi.org/10.1016/S2055-6640\(20\)30318-6](https://doi.org/10.1016/S2055-6640(20)30318-6)
28. Tang W, Mao J, Liu C, Mollan K, Zhang Y, Tang S, et al. Reimagining health communication: A noninferiority randomized controlled trial of crowdsourced intervention in China. *Sex Transm Dis*. 2019;46(3):172–178. <https://doi.org/10.1097/olq.0000000000000930>
29. Rosenberg NE, Obiezu-Umeh CS, Gbaja-Biamila T, Tahlil KM, Nwaozuru U, Oladele D, et al. Strategies for enhancing uptake of HIV self-testing among Nigerian youths: A descriptive analysis of the 4YouthByYouth crowdsourcing contest. *BMJ Innov*. 2021;7(3):590–596. <http://dx.doi.org/10.1136/bmjinnov-2020-000556>
30. Hlatshwako T, Conserve D, Day S, Reynolds Z, Weir S, Tucker JD. Increasing men's engagement in HIV testing and treatment programs through crowdsourcing: a mixed-methods analysis in Eswatini. *Sex Transm Dis*. 2021;48(10):789–797. <https://doi.org/10.1097%2FOLQ.0000000000001408>
31. Wang C, Han L, Stein G, Day S, Bien-Gund C, Mathews A, et al. Crowdsourcing in health and medical research: A systematic review. *Infect Dis Poverty*. 2020;9(1), Art. #8. <https://doi.org/10.1186/s40249-020-0622-9>
32. The National Academies of Sciences Engineering and Medicine. The Impact of Social Networking and Crowdsourcing on Research, the Enterprise, and the Workforce: A Workshop. 2011 [cited 2022 September 29]. Available from: <https://www.nationalacademies.org/our-work/the-impact-of-social-networking-and-crowdsourcing-on-research-the-enterprise-and-the-workforce-a-workshop>
33. Office of Behavioral and Social Sciences Research. Scientific Priorities for Behavioral and Social Sciences Research at NIH [webpage on the Internet]. c2020 [cited 2022 Sep 29]. Available from: <https://obsr.ideascale.com/>
34. Wu D, Best LL, Stein G, Tang W, Tucker JD. Community participation in a Lancet Healthy Cities in China Commission. *Lancet Planet Health*. 2018;2(6):e241–e242. [https://doi.org/10.1016/S2542-5196\(18\)30083-4](https://doi.org/10.1016/S2542-5196(18)30083-4)
35. World Health Organization, TDR, Social Innovation in Health Initiative. Crowdsourcing in health and health research: A practical guide. Geneva: World Health Organization; 2018. Available from: <https://apps.who.int/iris/handle/10665/273039>
36. Tucker JD, Pan SW, Mathews A, Stein G, Bayus B, Rennie S. Ethical concerns of and risk mitigation strategies for crowdsourcing contests and innovation challenges: scoping review. *J Med Internet Res*. 2018;20(3), e75. <https://doi.org/10.2196/jmir.8226>
37. Saliternik M. Big data and the right to political participation. *U Pa J Const L*. 2018;21:713.
38. Tang W, Wei C, Cao B, Wu D, Li KT, Lu H, et al. Crowdsourcing to expand HIV testing among men who have sex with men in China: A closed cohort stepped wedge cluster randomized controlled trial. *PLoS Med*. 2018;15(8), e1002645. <https://doi.org/10.1371/journal.pmed.1002645>
39. Iwelunmor J, Ezechi O, Obiezu-Umeh C, Gbaja-Biamila T, Nwaozuru U, Oladele D, et al. The 4 Youth by Youth HIV self-testing crowdsourcing contest: A qualitative evaluation. *PLoS ONE*. 2020;15(5), e0233698. <https://doi.org/10.1371%2Fjournal.pone.0233698>
40. Research for Ethical Data Science in Southern Africa (REDSSA) [webpage on the Internet]. No date [cited 2022 Sep 29]. Available from: <http://www.sun.ac.za/english/faculty/healthsciences/cmcl/redssa>
41. NIH Office of Strategic Coordination – The Common Fund. Program snapshot: Harnessing data science for health discovery and innovation in Africa (DS-I Africa) [webpage on the Internet]. No date [updated 2022 Jul 15; cited 2022 Sep 29]. Available from: <https://commonfund.nih.gov/africadata>



Setting up data science research in Africa and engagement of stakeholders

AUTHORS:

Fati Murtala-Ibrahim¹

Jibreel Jumare²

Manhattan Charurat³

Chenfeng Xiong⁴

Vivek Naranbhai⁵

Patrick Dakum¹

Shirley Collie⁶

Waasila Jassat⁷

Gambo Aliyu⁸

Adetifa Ifedayo⁹

Alash'le Abimiku²

AFFILIATIONS:

¹Institute of Human Virology Nigeria (IHVN), Abuja, Nigeria

²International Research Center of Excellence (IRCE), Institute of Human Virology, Abuja, Nigeria

³Institute of Human Virology, School of Medicine, University of Maryland, Baltimore, Maryland, USA

⁴Department of Civil and Environmental Engineering, College of Engineering, Villanova University, Philadelphia, Pennsylvania, USA

⁵Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

⁶Discovery Health, Johannesburg, South Africa

⁷National Institute for Communicable Diseases (NICD), Johannesburg, South Africa

⁸National Agency for the Control of AIDS (NACA), Abuja, Nigeria

⁹Nigeria Centre for Disease Control (NCDC), Abuja, Nigeria

CORRESPONDENCE TO:

Fati Murtala-Ibrahim

EMAIL:

fmurtalaibrahim@ihvnigeria.org

HOW TO CITE:

Murtala-Ibrahim F, Jumare J, Charurat M, Xiong C, Naranbhai V, Dakum P, et al. Setting up data science research in Africa and engagement of stakeholders. *S Afr J Sci.* 2023;119(5/6), Art. #14726. <https://doi.org/10.17159/sajs.2023/14726>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

KEYWORDS:

data science, research, stakeholders, Africa

FUNDING:

US National Institutes of Health (5U54TW012041-02)

PUBLISHED:

30 May 2023



© 2023. The Author(s). Published under a Creative Commons Attribution Licence.

Significance:

Data science explores the use of big data to gain deeper insights and generate new knowledge and innovations which can lead to economic growth and sustainable development. However, setting up data science research comes with challenges. How we engage stakeholders is a major factor that determines success. This Commentary highlights important considerations for stakeholder engagement based on the experiences of investigators in a data science for health discovery project underway in Nigeria and South Africa. The perspectives presented will guide implementation in this relatively new but rapidly growing research domain.

Background

Health sciences research has been defined to include basic, clinical, and applied science on human health and well-being. It explores the determinants, prevention, detection, treatment, and management of diseases, and can be extended to data science research.¹ Setting up health sciences research in Africa will promote a strong health science industry as part of broader efforts to establish a robust research and development (R&D) environment, accelerating the emergence of knowledge-based economies that engender sustainable growth and development.¹ In a seminal 1990 report, the Commission on Health Research and Development stated that strengthening research capacity in low- and middle-income countries (LMICs) is “one of the most powerful, cost-effective and sustainable means of advancing health and development”¹. Applying data science to health sciences research provides an opportunity to use large data sets generated in public health settings to gain deeper insight and generate new knowledge and innovations. It also provides better ways of implementing research to achieve greater health benefits, improving the economy of countries.²⁻⁴ The Data Science for Health Discovery and Innovation in Africa (DS-I Africa) initiative aims to leverage data science technologies to transform biomedical and behavioural research. This initiative ultimately intends to develop solutions that would improve health for individuals and populations in Africa.⁵ Likewise, the INFORM-Africa (Role of Data Streams in Informing Infection Dynamics in Africa) Research Hub partners with the governments, health facilities, industry, and communities in Nigeria and South Africa to advance data science in Africa by closing the gap in utilisation of big data and analytical capacity. The core objective is to strengthen the use of existing population-scale epidemiologic data sources as a cornerstone of future pandemic preparedness, using HIV and COVID-19 pandemics as examples.

Evolution of data science

Data science is an emerging and evolving discipline, especially in LMICs, and needs to be explored in sub-Saharan Africa to maximise the gains. Data science has been described by Beyene et al.² as an integrated interdisciplinary approach used to develop tools, templates, and processes to conduct complex analyses of big data sets. The authors attribute the slow evolution of data science in Africa and other resource-limited settings to a lack of well-trained data scientists.² ‘Data science’ as a term was created in the early 1960s and used to describe a discipline that supports the synthesis and interpretation of the large amount of data that had been generated over time, but it has evolved from statistics and data analysis to include computer science concepts like artificial intelligence, machine learning, and the Internet of Things.²⁻⁴

Stakeholder engagement

Data science research provides the opportunity for global collaboration amongst a wide range of expertise for maximum impact to be achieved. Tembo et al.⁶ describe this collaboration in different ways depending on the region; for instance, in high-resource settings, it is known as ‘patient and public involvement’, ‘engagement’, or ‘participation’. In LMICs, these approaches are termed ‘community engagement’, ‘participation’, and ‘community engagement and involvement’.⁶ In the same vein, the INFORM-Africa Research Hub has assembled experienced researchers with complementary expertise in big data analytics, quantum information processing, spatial statistics and analysis, genetics, computational biology, agent-based and data-driven modelling, clinical infectious diseases, infectious disease epidemiology, molecular virology, and geospatial analytics to address its research goal as outlined above.

Importance of stakeholders

The importance of stakeholders cannot be overemphasised. In most health-related research, stakeholder engagement and involvement can add value to the implementation of the research, in addition to building new knowledge and innovation.⁷ It is important to make the stakeholders as broad and varied as reasonably possible, to engage them at the planning stage of the research when the priorities are being set, and to involve them in the design phase of the research project.⁸ This helps to incorporate culturally acceptable norms into the research study proposed, ensures alignment with the priorities of the communities, supports the recruitment and retention of research participants, and facilitates overall implementation and dissemination of the research findings.⁸⁻¹⁰ In summary, stakeholder involvement provides an opportunity for inclusion rather than exclusion.^{11,12} Governments in Africa and most LMICs are important stakeholders for health-related data research as they own most health facilities at all levels of care. They also own most of the laboratories that generate results of clinical investigations. This makes the government an important stakeholder when big data generated by health facilities and laboratories is required for research. The INFORM-Africa Research Hub has benefitted from significant input

from all its stakeholders during grant proposal development, planning, and implementation. Its expanded and multidisciplinary stakeholders, including policymakers and communities of both countries, will play a significant role in disseminating its findings and products.

Challenges to data science research

In Africa, several challenges exist regarding access to big data and other aspects of data science research, including:

- lack of trained data scientists and inability to retain well-trained scientists (brain drain);
- limited infrastructure (facilities for curating research data, integrated Electronic Medical Record Systems, establishing national databases, electronic surveillance systems, national vital statistics repository, etc.);
- limited awareness of the value of data science research among researchers and research institutions;
- limited resources/funding for data science research;
- limited data sharing culture and opportunities;
- limited engagement of communities in research through community participatory research initiatives;
- limited training and availability of adequate provisions on regulatory/ethical guidelines for data science research;
- limited engagement of private health facilities and health insurance data sources; and
- poor clinical documentation, record keeping, and data management practices.

Standards in engaging stakeholders for maximum benefit

When standards that use the internationally recognised four foundational principles for scientific data management and stewardship – Findability, Accessibility, Interoperability, and Reusability (FAIR)⁵ – are developed within the appropriate cultural context, they guide research teams to engage stakeholders and minimise the challenges often experienced. Some of the standards and guidance also provided by Tembo et al. include adopting the principles of power-sharing, building relationships, acknowledging diverse perspectives, reciprocity, and respecting different knowledge bases.^{6,8} By partnering with government agencies, health data custodians, community gatekeepers, notable leaders in the scientific community, and research ethics boards in Nigeria and South Africa, INFORM-Africa has started these critical steps of engagement. We have incorporated representatives of our stakeholders in our standing committees, in addition to sharing and reviewing documents such as protocols, standard operating procedures and data sharing agreements, to ensure that elements of the FAIR principles are captured within the cultural context of these countries. Globally, engaging stakeholders and involving them in research efforts from conception to disseminating results will play a huge role in data science and changing policies.

Acknowledgements

We acknowledge the role of the DSI-Africa Consortium in making this research possible. The US National Institutes of Health provided the core funding for the DSI-Africa Consortium (more information is available at <https://dsi-africa.org/>). This research was specifically funded by the US National Institutes of Health (grant number 5U54TW012041-02). We also acknowledge the contributions of the INFORM-Africa project team, staff

and management of the Institute of Human Virology Nigeria (IHVN), University of Maryland, Baltimore (UMB), Villanova University, Centre for the AIDS Programme of Research in South Africa (CAPRISA), Nigeria Centre for Disease Control (NCDC), Discovery Health in South Africa, National Institute for Communicable Diseases (NICD) in South Africa, National Agency for the Control of AIDS (NACA), Nigeria National AIDS and STD Control Programme (NASCP), and all participants, patients, investigators, clinicians and personnel involved in generating the primary data used in the INFORM-Africa research project.

Author information

All the authors are part of the INFORM-Africa Research Study Group.

Competing interests

We have no competing interests to declare.

References

1. Wenham C, Wouters O, Jones C, Juma PA, Mijumbi-Deve RM, Sobngwi-Tambekou JL, et al. Measuring health science research and development in Africa: mapping the available data. *Health Res Policy Sys.* 2021;19(1), Art. #142. <https://doi.org/10.1186/s12961-021-00778-y>
2. 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>
3. Cao L. Data science: A comprehensive overview. *ACM Comput Surv.* 2017;50(3), Art. #43. <https://doi.org/10.1145/3076253>
4. Navarro FCP, Mohsen H, Yan C, Li S, Gu M, Meyerson W, et al. Genomics and data science: An application within an umbrella. *Genome Biol.* 2019;20(1), Art. #109. <https://doi.org/10.1186/s13059-019-1724-1>
5. US National Institutes of Health (NIH). National Institutes of Health strategic plan for data science 2018 [document on the Internet]. c2018 [cited 2022 Sep 05]. Available from: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf
6. Tembo D, Hickey G, Montenegro C, Chandler D, Nelson E, Porter K, et al. Effective engagement and involvement with community stakeholders in the co-production of global health research. *BMJ.* 2021;372, Art. #178. <https://doi.org/10.1136/bmj.n178>
7. Staunton C, Tindana P, Hendricks M, Moodley K. Rules of engagement: Perspectives on stakeholder engagement for genomic biobanking research in South Africa. *BMC Med Ethics.* 2018;19(1), Art. #13. <https://doi.org/10.1186/s12910-018-0252-y>
8. Goodman MS, Sanders Thompson VL. The science of stakeholder engagement in research: Classification, implementation, and evaluation. *Transl Behav Med.* 2017;7(3):486–491. <https://doi.org/10.1007/s13142-017-0495-z>
9. Musesengwa R, Chimbari MJ, Mukaratirwa S. Initiating community engagement in an ecohealth research project in southern Africa. *Infect Dis Poverty.* 2017;6(1), Art. #22. <https://doi.org/10.1186/s40249-016-0231-9>
10. Hinchcliff R, Greenfield D, Braithwaite J. Is it worth engaging in multi-stakeholder health services research collaborations? Reflections on key benefits, challenges and enabling mechanisms. *Int J Qual Health Care.* 2014;26(2):124–128. <https://doi.org/10.1093/intqhc/mzu009>
11. Boaz A, Hanney S, Borst R, O'Shea A, Kok M. How to engage stakeholders in research: Design principles to support improvement. *Health Res Policy Sys.* 2018;16(1), Art. #60. <https://doi.org/10.1186/s12961-018-0337-6>
12. Laird Y, Manner J, Baldwin L, Hunter R, McAteer J, Rodgers S, et al. Stakeholders' experiences of the public health research process: Time to change the system? *Health Res Policy Sys.* 2020;18(1), Art. #83. <https://doi.org/10.1186/s12961-020-00599-5>



Revealing human mobility trends during the SARS-CoV-2 pandemic in Nigeria via a data-driven approach

AUTHORS:

Weiyu Luo¹
Chenfang Xiong¹
Jiajun Wan²
Ziteng Feng¹
Olawole Ayorinde³
Natalia Blanco^{4,5}
Man Charurat^{4,5,6}
Vivek Naranbhai^{7,8}
Christina Riley⁹
Anna Winters⁹
Fati Murtala-Ibrahim³
Alash'le Abimiku^{3,4,5}

AFFILIATIONS:

¹Department of Civil and Environmental Engineering, College of Engineering, Villanova University, Philadelphia, Pennsylvania, USA
²Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
³Institute of Human Virology Nigeria (IHVN), Abuja, Nigeria
⁴University of Maryland School of Medicine Institute of Human Virology, Baltimore, Maryland, USA
⁵Center for International Health Education and Biosecurity, University of Maryland School of Medicine, Baltimore, Maryland, USA
⁶Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA
⁷Centre for the AIDS Programme of Research in South Africa (CAPRISA), Nelson R. Mandela School of Medicine, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa
⁸HIV Pathogenesis Programme, Nelson R. Mandela School of Medicine, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa
⁹Akros, Lusaka, Zambia

CORRESPONDENCE TO:

Chenfang Xiong

EMAIL:

chenfang.xiong@villanova.edu

DATES :

Received: 09 Sep. 2022
Revised: 06 May 2023
Accepted: 17 May 2023
Published: 30 May 2023

HOW TO CITE:

Luo W, Xiong C, Wan J, Feng Z, Ayorinde O, Blanco N, et al. Revealing human mobility trends during the SARS-CoV-2 pandemic in Nigeria via a data-driven approach. *S Afr J Sci.* 2023;119(5/6), Art. #14727. <https://doi.org/10.17159/sajs.2023/14727>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

DATA AVAILABILITY:

- [Open data set](#)
- All data included
- On request from author(s)
- Not available
- Not applicable

EDITORS:

Jennifer Fitchett
Adriaan van der Walt

KEYWORDS:

SARS-CoV-2, human mobility, trips, policy, Nigeria

FUNDING:

US National Institutes of Health (U54TW012041-02)

© 2023. The Author(s). Published under a Creative Commons Attribution Licence.

Research Article

<https://doi.org/10.17159/sajs.2023/14727>

We employed emerging smartphone-based location data and produced daily human mobility measurements using Nigeria as an application site. A data-driven analytical framework was developed for rigorously producing such measures using proven location intelligence and data-mining algorithms. Our study demonstrates the framework at the beginning of the SARS-CoV-2 pandemic and successfully quantifies human mobility patterns and trends in response to the unprecedented public health event. Another highlight of the paper is the assessment of the effectiveness of mobility-restricting policies as key lessons learned from the pandemic. We found that travel bans and federal lockdown policies failed to restrict trip-making behaviour, but had a significant impact on distance travelled. This paper contributes a first attempt to quantify daily human travel behaviour, such as trip-making behaviour and travelling distances, and how mobility-restricting policies took effect in sub-Saharan Africa during the pandemic. This study has the potential to enable a wide spectrum of quantitative studies on human mobility and health in sub-Saharan Africa using well-controlled, publicly available large data sets.

Significance:

- The mobility measurements in this study are new and have filled a major data gap in understanding the change in travel behaviour during the SARS-CoV-2 pandemic in Nigeria. These measurements are derived from high-quality data samples by state-of-the-art data-driven methodologies and could be further adopted by other quantitative research related to human mobility.
- Additionally, this study evaluates the impact of mobility-restricting policies and the heterogeneous effects of socio-economic and socio-demographic factors by a time-dependent random effect model on human mobility. The quantitative model provides a decision-making basis for the Nigerian government to provide travel-related guidance and make decisions in future public health events.

Background

The spread of the coronavirus (SARS-CoV-2) in 2020 became an enduring war of global public health. Up until 23 May 2023, over 760 million confirmed cases had accumulated worldwide, claiming almost 6.9 million lives.^{1,2} Despite the many effective vaccines rolled out to fight such an unprecedented war, more transmissible variants still emerged. A spike of deaths and new infections was still observed in January 2023 in the Western Pacific region.² For us to better learn from past experiences and recommend the path forward, it is imperative to look back and learn from the ways in which human beings fought this battle, so that we can rethink our strategy for future pandemics.

One key lesson worth evaluating is the effectiveness of non-pharmaceutical interventions (NPIs) such as stay-at-home orders and travel regulations. At the early stage of the pandemic, governments and their citizens adopted such NPIs at different levels globally to contain the spread of the SARS-CoV-2 virus.³⁻⁷

Perra⁸ conducted a thorough review of NPIs during the pandemic and summarised data sets, modelling methods and findings. Researchers have looked into NPIs implemented by several nations, such as the USA, South Korea, China and countries of the European Union, and found them to be quite effective in delaying and containing the spread of disease.⁹⁻¹¹

With respect to the travel restriction policy as an NPI, researchers have found it to be particularly useful in the early stages of an outbreak, and specifically when the policy is confined to an area that is considered to be the major source of the virus. Previous studies have proven a positive relationship between human mobility and SARS-CoV-2 cases.¹²⁻¹⁴ Mobility restrictions may become less effective once the outbreak is more widespread at a later stage.^{15,16} In terms of data sources, dedicated surveys and passively collected smartphone location data were the most used data sets in NPI-related studies. Different surveys have been conducted to quantify the impacts of NPIs on human activities^{17,18}, social distancing and close contacts^{19,20}, as well as well-being indicators on, for instance, mental health^{21,22} and health behaviour^{23,24}. These dedicated surveys covered a variety of sample sizes, ranging from 500 to greater than 100 000. Compared with survey data, an emerging data source for estimating the human behavioural response to mobility restrictions and how that was associated with the onset of SARS-CoV-2 was smartphone location data collected passively via mobile devices by location intelligence and measurement platforms.^{13,14,25-27} Technology advances have led to an increasingly higher penetration of smartphones and the vital roles they play in people's daily lives, making such data a unique, high-resolution and cost-effective source of information on human movement and possible changes in movement without compromising the confidentiality of these data.

Research gaps and motivation

From reports in the literature, it is important to note that most studies have focused on data-rich societies such as the USA, China and European countries.⁸ Very little research attention has been paid to the rest of the world, and there are



limited data sets on this topic. Only two studies on NPIs were found in sub-Saharan Africa.^{28,29} Evans et al.²⁸ developed a prediction model of SARS-CoV-2 cases in Madagascar. Zandvoort et al.²⁹ studied Nigeria's NPIs with a modified SEIR (Susceptible, Exposed, Infective, and Recovered) model and concluded that physical distancing alone might not have been enough to contain the virus without lockdown. Their paper highlighted the need for reliable data sources on mobility and social distancing. Because of the lack of such data, they could only adopt synthetic contact matrices to model the effect of lockdown and behaviour.

Considering that little scientific evidence is available, sub-Saharan Africa is in need of proper data and research on human activities, mobility and the association of mobility with disease transmission due to the high disease burden that exists in sub-Saharan Africa. Tuberculosis is among the leading causes of death of African citizens.³⁰ Several SARS-CoV-2 variants and the new outbreak of monkeypox were first reported in sub-Saharan Africa.¹⁶ An appropriate human mobility measurement process can best supplement studies of the transmission of these diseases and other emerging and re-emerging infections. Crucial research questions need to be answered, such as the speed of the outbreaks, how human movements and gatherings contribute to them, and how effective different lockdown policies were. While conducting dedicated travel surveys on individuals and/or households remains a high-cost approach to understanding people's travel and mobility changes, smartphone penetration rate, which has increased steadily in sub-Saharan countries, makes mobile phone-based data collection an attractive alternative. In 2021, the rate reached 20% in Nigeria and 25% in South Africa.³¹ These rates make using passively collected location information from smartphone platforms a feasible and promising option to study human mobility in sub-Saharan countries.

Motivated by the need to develop human mobility measurements and models for Africa, we studied mobile device location data and measured individual-level travels based on a data-driven approach. The methodology was adapted from a parallel research effort in the USA in developing US national travel patterns and origin-destination trip matrices.^{32,33} Two human mobility measurements, i.e. daily average number of trips per person and daily average distance travelled per person, were taken from a filtered high-quality sub-sample using Nigeria as the study area. The study covered the period of 01 January 2020 to 25 April 2020, with the aim of depicting daily human mobility changes at the beginning of the SARS-CoV-2 pandemic. Then, a time-series model of human mobility was developed and estimated

to quantify changes in people's travel patterns, and how the pandemic and its associated travel restriction policies affected these patterns. At the time of writing this paper, the team had also started working on mobility data production for South Africa. The study of a two-country statistical comparison is the immediate next step. To our best knowledge, this is the first attempt to use emerging passively collected mobile device location data to measure travel behaviour, such as trip-making behaviour and travelling distances in the sub-Saharan region. This study will fill a critical and long-lasting data gap in transportation and mobility studies. Based on such measurements, we also empirically tested the effectiveness of mobility restriction policies to reach a number of policy implications supported by emerging data evidence.

SARS-CoV-2 in Nigeria and associated human mobility policies

The development of SARS-CoV-2 in Nigeria and its associated public policy decisions in relation to human mobility are summarised in Figure 1.

This study focuses on the beginning of the SARS-CoV-2 pandemic in Nigeria. This initial period can be divided into three stages based on transitions of public policy on human mobility:

- **Pre-lockdown Stage³⁴** (indicated in blue in Figure 1): The first confirmed SARS-CoV-2 case was discovered in Ogun State, Nigeria, on 27 February 2020. Travel bans on countries with ongoing high transmissions such as China, Italy and Germany were issued on 18 March, and on the same day, Lagos and Ogun banned mass gatherings and religious activities for more than 50 people. Schools were closed on the subsequent day. On 23 March, in order to prevent disease importation, all international flights were banned and land borders were closed. Mandatory quarantine and testing were required for international returnees.
- **Lockdown Stage^{34,35}** (indicated in red in Figure 1): Phase 1 of federal lockdown was issued for Lagos, Ogun and the Federal Capital Territory (FCT) on 30 March 2020. The lockdown was in effect for 2 weeks and included several measures to slow virus transmission (such as workplace closure, the banning of social gatherings and public events, and curfews). These states were selected based on several risk factors, including large numbers of confirmed cases



Figure 1: A timeline of Nigeria's SARS-CoV-2 situation and mobility-restricting policies.^{34,35}

and their high population densities. Starting on 02 April, other states entered lockdown as well. The first was Bauchi, followed by many others. After the 14-day Phase 1 lockdown, another 21-day federal lockdown (Phase 2) was issued for Lagos, Ogun, the FCT and Kano (due to a rapid increase in cases). On 23 April, inter-state travel restrictions were imposed in all states, as well as the FCT, and domestic flights were grounded.

- **Lockdown Easing Stage³⁵** (indicated in green in Figure 1): After the two phases of federal lockdown, Nigeria started to loosen its restriction measures through gradual lockdown easing. Phase 1 of lockdown easing commenced on 04 May, which was initially planned for two weeks (until 17 May), but was extended for another two weeks (until 01 June). Lagos and the FCT were included in this first phase of lockdown easing. A national curfew of 20:00 to 06:00 was declared in accordance with the lockdown easing measures. Phase 2 of lockdown easing started on 02 June. It lasted for four weeks and ended on 29 June. Restriction measures loosened, including a shortened national curfew from 22:00 to 04:00, the reopening of banks, and the exemption from inter-state travel restrictions of providers of essential services and manufacturers of produce. Phase 3 of lockdown easing lasted for 4 weeks, commencing on 30 June and ending on 27 July. Restriction measures continued to loosen. These included the re-opening of local flights 'based on close monitoring' and the resumption of schools for certain grades. However, there was no change in the national curfew, and the failure to use face masks in public was still punishable by law until 23 June 2022.³⁶

Data sources and methodology of mobility data analytics and modelling

Data sources

In this study, the primary data source used to measure human mobility was smartphone locations licensed from third-party data providers who supplied opted-in and anonymised mobile device location pins via Global Positioning System (GPS), wireless fidelity (Wi-Fi), Internet Protocol (IP), and Internet of Things (IoT) signalling. The raw data panel contained about 62 500 opted-in samples on a daily basis for Nigeria, generating some 2 570 000 sightings (i.e. one location point with a time stamp) daily from 01 January to 25 April 2020. Devices of smartphone operating systems collected the data of anonymised samples of people who had opted in to share their locations. The data collection processes do not collect any personal-identifiable information and employ privacy protection techniques to substantially reduce the risk of reidentification, e.g. aggregate the home and work location to a coarser geographical level.

To facilitate understanding of mobility and changes in mobility before and during the pandemic, supplementary data was collected and digitised in parallel with the smartphone locations. First of all, events and government policies described in the previous section were digitised

into dates and dummy variables that were later incorporated into the mobility model. These variables included the announcement of the first SARS-CoV-2 case in Nigeria, the ban on travel and mass gatherings, and the Phase 1 and Phase 2 federal lockdown levels. Also related to the SARS-CoV-2 situation, daily new number of confirmed cases of each state from the Humanitarian Emergency Response in Africa (HERA) was integrated into the model.³⁷

The inherent discrepancy in mobility across different states within Nigeria should also be linked to time-invariant covariates such as population structure, economy and number of facilities. Population-specific information such as age and gender were extracted from United Nation Population Fund data (<https://pdp.unfpa.org/>). In particular, the percentage of members of the population below 14 years and above 65 years of age were entered into the mobility model. Then, to capture the impact of household and individual income levels on travel, an indicator of general economy was adopted as a proxy. This indicator, the Relative Wealth Index (RWI⁹), was recently developed to micro-estimate the relative wealth and poverty levels in low- and middle-income countries at a 2.4 km resolution. In this study, the RWI points were spatially joined at county level and the values were averaged. The county-level to state-level values, weighted by county population, were then aggregated. The percentile of the RWI values of all states were then calculated and ranked. A value of 1 represents the highest-income state and a value of 0 represents the lowest-income state. Lastly, the availability of points of interest would play a crucial role in mobility. Due to a lack of sufficient point-of-interest data records, we were only able to incorporate the number of health facilities as covariates.³⁸ Three levels of healthcare delivery in Nigeria were included in the data set. We aggregated the count of healthcare delivery to state level, including hospitals, pharmacies, clinics, health centres, medical centres, maternity homes, laboratories and other entities that provide medical and/or healthcare services. It was believed that the availability of healthcare facilities would play a special role in impacting people's travel decisions during the pandemic^{12,39}, as providing quality health and medical services at the travel destination could reduce the fear of travel³⁹.

Data analytical method of measuring human mobility

Figure 2 gives an overview of the analytical framework that was developed to measure human mobility. The framework is based on our existing research.³³ Beginning from raw location data, a series of quality metrics was first developed to confirm data frequency, stability and data consistency. Partially for preserving data privacy, smartphone location data were not collected frequently. In an extreme case, one sample in the raw data was only observed in one sighting (i.e. a location point with a time stamp). To address this limitation, we filtered only the regular active users (RAU) as a sub-sample for subsequent mobility identification. An RAU must be observed at least eight times at different locations in a single day, and then has to be observed at eight different unique hours on that day. This is defined based on a trade-off between sample size and statistical biases of mobility measurements.³³ With the implementation of this RAU quality

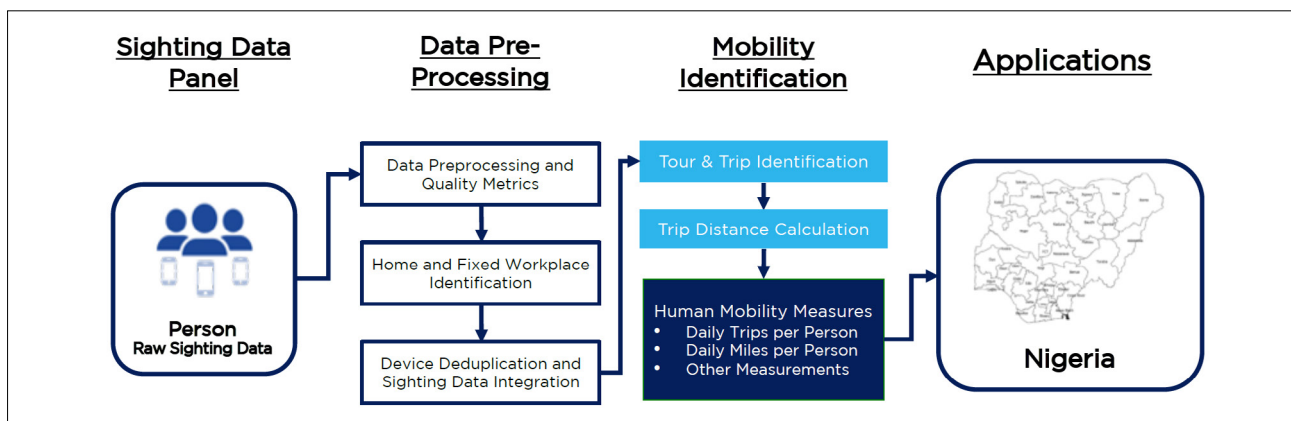


Figure 2: The analytical framework of measuring human mobility via passively collected smartphone location data.

filter, we assembled a sub-sample of 281 458 RAUs (from January 2020 to April 2020) in Nigeria for the subsequent data analytics steps.

A behaviour-based method was used to impute home and fixed workplaces based on the most frequently visited places at night and during the day. This step offered each sample an anchor for its daily life centres and for the subsequent identification of travels. Then, a deduplication step removed repetitive data from the observations as one individual could possess multiple mobile devices and share data with multiple data vendors. This process avoided the over-representation of individuals using multiple devices and sharing data with multiple data vendors, and consolidated unique device sightings. In the Nigeria study presented in this paper, about 0.04% of the devices were removed as duplicates.

The most critical pillar of the framework is the mobility identification steps. We employed a tour-based approach to properly identify all tours and trips from the raw location data, including trip origin, destination, start time and end time. A tour means a sequence of linked trips that fulfil similar mobility needs for a person. The tour-based method enables one to consider trip chaining and differentiate between linked and unlinked trips. Many traditional methods can only identify unlinked trips, for instance, a single transit commute trip with a long time of waiting at the origin, and/or transfer transit stations would be identified as multiple unlinked trips. Figure 3 illustrates how the tour-based algorithm produced more accurate trip identification results than traditional methods. Figure 3a and 3b show how the tour-based method differentiates true activity clusters (e.g. a home cluster and a work cluster) from mid-trip transfer points (e.g. waiting at a transit station). More details about this algorithm can be found in Zhang et al.³³

The tour and trip identification approach is then applied to all RAU sub-samples in the study area and for the designated study period to yield a roster of trips. The location points of each trip are then used to estimate the great-circle distance. Because of the lack of appropriate multimodal transportation network data, especially for transit and rail modes, it is difficult to reconstruct accurate turn-by-turn movements of sub-samples and thereby estimate trip distance. As this research focused more on analysing the mobility and travel behaviour change along the timeline of COVID-19, an unbiased travel distance approximation⁴⁰ was found to be acceptable and sufficient for this study, rather than elaborating efforts and computing powers to obtain network-based distance. Therefore, great-circle distance is employed to approximate the actual trip distance. This limitation will be addressed once a routable transportation network is developed for analysis. Finally, the trip roster, together with the approximated trip distance information, was employed to generate two aggregated human mobility measurements:

- Daily trips per person: The number of trips made by each person per day

- Daily distance per person: The total approximated distance travelled by each person per day

To date, there are limited data to validate the mobility findings generated from the passively collected smartphone location data. Because of the sparsity of such location data, it is possible that the proposed methodology under-estimates the number of trips and/or distance travelled per person per day. Without appropriate validation data and a calibration process, the possible measurement bias may not be properly identified and mitigated. This will remain a critical research topic yet to be completed. On the other hand, the consistency of the data in terms of number of devices and number of sightings per device on a daily basis has been thoroughly evaluated. The study also filtered high-quality RAUs as the sub-sample used in the analytical framework. We are confident that the development mobility measurements reasonably reflect the actual behaviour shifts in Nigeria.

With this overarching framework and its capability to analyse individual-level tours and trips, additional mobility measures can be derived with additional future research and development effort. This paper is focused on analysing and modelling these two measurements in Nigeria, which is one of two country-level study and application sites.

Modelling the time-dependent human mobility measures

Individual-level trip information was aggregated to state level for Nigeria. We employed a random-effect model of panel data to capture the relationship between human mobilities and government policies, SARS-CoV-2 cases and several time-invariant covariates. A one-day lag variable was embedded in the model to capture the first-order autocorrelation of the dependent variable. The formulation of the model is described in Equation 1:

$$Y_{kit} = c_{it} + \sum_{M,i} \beta_{mi} X_{mi} + \sum_{N,i,t} \beta'_{nit} X_{nit} + \sum_{K,i,t-1} Y_{k,i,t-1} Y_{k,i,t-1} + u_i + e_{it} \quad \text{Equation 1}$$

where Y_{kit} represents the k^{th} dependent variable of state i at time t ; c_{it} is the constant term serving as the intercept of the model for each state at each time; X_{mi} is the m^{th} time-invariant variable of state i , and β_{mi} is the corresponding coefficient; X_{nit} is the n^{th} time-variant variable of state i , and β'_{nit} is the corresponding coefficient; $Y_{k,i,t-1}$ stands for the k^{th} time-series variable of state i at time $t-1$, which is lagged by one day; $Y_{k,i,t-1}$ is the corresponding coefficient of $Y_{k,i,t-1}$; u_i is the random effect term, which is independent of all X_{mi} and $Y_{k,i,t}$, but common to all states i ; e_{it} stands for the error term. In our experiment, there were $M=5$ time-invariant variables, $N=4$ time-variant no-lag variables and $K=2$ time-variant variables lagged by one day. The description and the type of variables are described in Table 1.

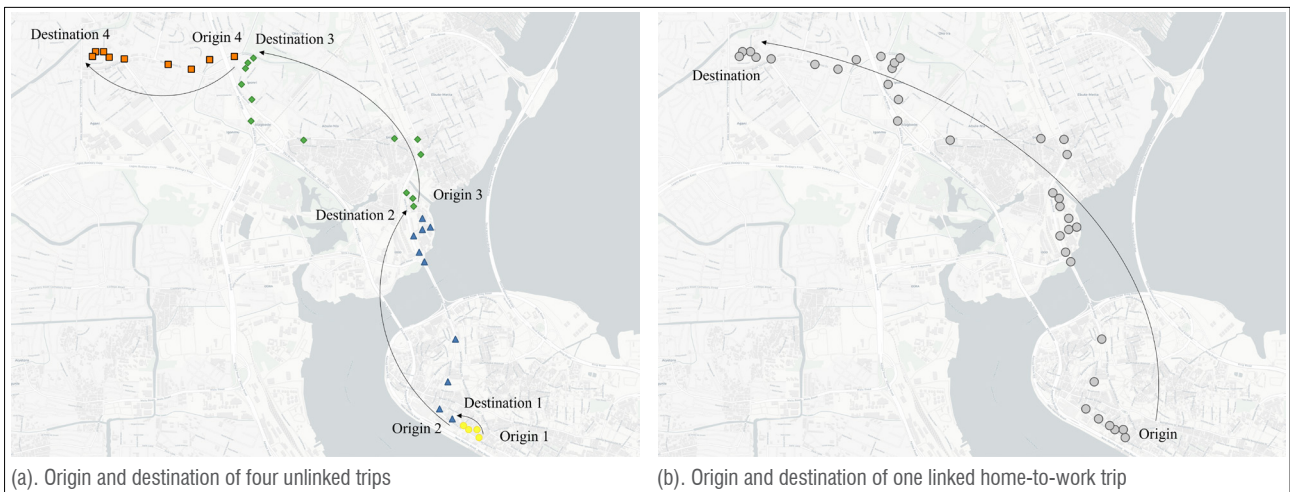


Figure 3: Tour identification and trip chaining demonstration.

Data measurements and modelling results

Human mobility measurements

Based on the data sources and methodology elaborated on in the previous section, we derived daily trips and daily distances travelled per person as human mobility measurements for Nigeria. The results are illustrated in Figure 4. The measurements are averaged at the national level and cover the period 01 January to 25 April 2020, demonstrating the overall mobility changes in Nigeria at the beginning of the pandemic. The dates when different mobility-restricting policies were implemented are annotated in Figure 4.

Using the ban on travel and mass gatherings (19 March 2020) as the pandemic breakpoint, before that date, people made 2.0 trips and travelled 13.2 km per day. After the breakpoint, the daily distance

travelled per person dropped steadily to an average of 9.66 km, which represented a decrease of over 25%. The number of trips per person displayed a more notable fluctuation and averaged 2.1 trips per person per day after 19 March 2020, i.e. a 5% increase compared with the pre-pandemic level. Overall, the results show that human mobility patterns displayed an unprecedented modification during the study period. The pandemic, along with the series of mobility restrictions, seemed to have had a significant effect in limiting the distance travelled, while people were still making a good number of daily travels, measured by trips. This indicated that, when adapting to the pandemic, people reduced the number of longer trips and replaced them with shorter trips.

Another robust way of measuring the mobility changes during the SARS-CoV-2 pandemic was to analyse the relative percentage change in these mobility measures.

Table 1: Descriptive statistics of variables used in the random-effect model of human mobility panel data in Nigeria

Variable name	Description	Mean	S.D.	Max	Min	Unit
Trips/person	Daily trips travelled per person	2.05	0.40	4.26	1.24	–
Distance/person	Daily distance travelled per person	12.41	7.87	166.55	2.32	km
RWI percentile	Percentile of Relative Wealth Index (RWI) among all states in Nigeria	0.51	0.29	1.00	0.03	–
Age_0_14	Percentage of population below age 14	41.88	5.11	48.66	33.02	%
Age_65+	Percentage of population above age 65	3.24	0.73	4.72	1.48	%
Health facilities	Count of health facilities in each state, in the unit of 1000	1.25	0.48	2.33	0.39	1000
New cases	Daily new confirmed cases of SARS-CoV-2 in each state	0.28	2.87	80.00	0.00	–
First case	A dummy variable = 1 if the first confirmed case in Nigeria was announced	0.51	0.50	1.00	0.00	–
Travel ban	A dummy variable = 1 if the ban of foreign travel and gathering in Nigeria was issued	0.33	0.47	1.00	0.00	–
Fed-lockdown-p1	A dummy variable = 1 if Phase 1 of Nigeria federal lockdown was issued	0.23	0.42	1.00	0.00	–
Fed-lockdown-p2	A dummy variable = 1 if Phase 2 of Nigeria federal lockdown was issued	0.11	0.32	1.00	0.00	–

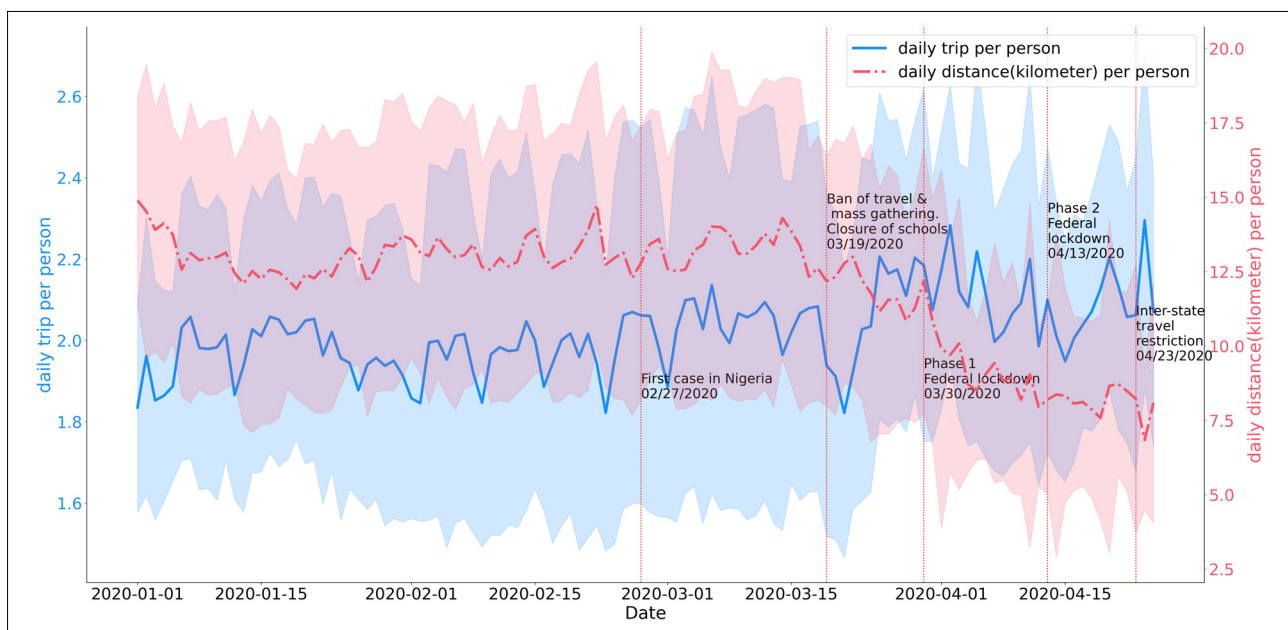


Figure 4: Daily trips per person and daily distance travelled per person measured using smartphone location data collected in Nigeria (01 January 2020 to 25 April 2020).

The average numbers measured for January 2020 were used as the benchmark to calculate the percentage change of the mobility measures for the following months of the study period. Figure 5a and 5b visualise the human mobility trends in percentage changes when compared with the January average. The statistics are reported for three different groups of states in Nigeria. The grouping was based on the groups' relative wealth according to RWI percentile rankings. Two thresholds for RWI (i.e. 0.33 and 0.66) were selected to divide the states into three groups of equal sample sizes. As shown in Figure 5a and 5b, the green curve with an RWI percentile ranking greater than 0.66 represents the situation for the high-income group of states, while the blue curve with an RWI percentile ranking lower than 0.33 represents the low-income states. The orange curve with an RWI percentile ranking between 0.33 and 0.66 represents the situation of the mid-income states. Both trips and distances showed some level of increase shortly after the announcement of the first SARS-CoV-2 case in Nigeria on 27 February 2020. This could be due to the panic facing the uncertainties of a new pandemic. People were travelling to get together or to get a part, stock up on goods or even relocate to another place with decent medical resources. On 19 March 2020, Nigeria banned entry for travellers from 13 countries with higher public health concerns, followed by the closure of schools. Around the same time, the World Health Organization (WHO) defined SARS-CoV-2 as a global pandemic. The trips and distances travelled per person in all state groups dropped in response to the announcement. Daily distances travelled in all groups continued to tumble, dropping to only 40–50% compared with the January average. However, the trends in the daily number of trips per person performed differently and showed discrepancies across state groups, especially after

the announcement of the federal lockdown. The daily trips per person of richer states rebounded drastically and stayed at the level of 110% compared with January 2020. For the relatively lower-income states, the daily number of trips decreased to around 83% of the January average.

Model estimation of time-dependent human mobility measures

The results of the random-effect model for daily trips per person and daily trip distance per person are reported in Table 2 and Table 3, respectively. The variable with the prefix "Lag1_" represents one day lag of the data. Goodness-of-fit indices (R-squared) are 0.6622 for the model of daily trips per person, and 0.2142 for the model of daily distance travelled per person, respectively. The codes, spreadsheet data and results have been deposited in GitHub (<https://github.com/villanova-transportation/Nigeria-mobility-COVID19-SAJS>) and are publicly accessible.

Regarding the lagged time-variant variable, as expected, one-step lagged daily trips per person and daily distance travelled per person both played a positive and statistically significant role in the two models. It also implies the existence of autocorrelation in the time-series data. For the trips per person, the magnitude of the effect of trip distance per person was limited. A trip distance of 1 km more per person on the previous date only increased to an additional trips per person at the current date. However, for the trip distance per person, the effect of trips per person was significant. One more trip per person on the previous date increased to an additional trip distance of per person at the current date.

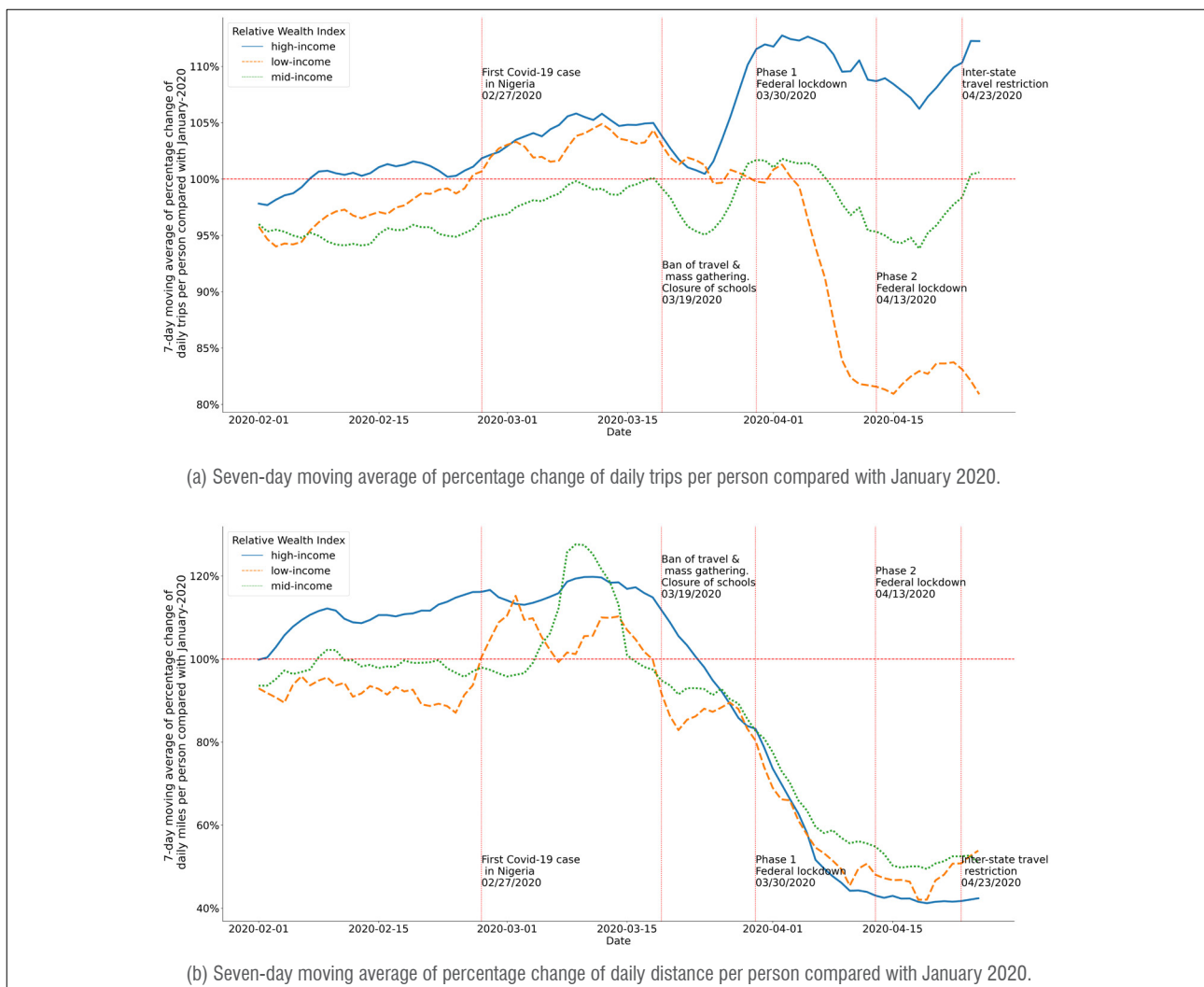


Figure 5: The trend of the seven-day moving average of (a) daily trips per person and (b) daily distance travelled per person for the three state groups differentiated by level of wealth.

Table 2: Model estimation results for the random effect model of daily trips per person

	Estimated coefficient	S.E.	p-value	
Constant	1.318	0.105	0.000	***
RWI percentile	-0.155	0.027	0.000	***
Lag1_trip-distance/person	0.001	0.001	0.011	*
Lag1_trips/person	0.762	0.010	0.000	***
Age_0_14	-0.014	0.002	0.000	***
Age_65+	-0.039	0.007	0.000	***
Health facilities	-0.030	0.008	0.000	***
New cases	0.000	0.001	0.765	
First case	0.021	0.010	0.034	*
Travel ban	-0.023	0.014	0.118	
Fed-lockdown-p1	-0.006	0.016	0.710	
Fed-lockdown-p2	-0.007	0.015	0.668	

Significant codes: 0***0.001***0.01**0.05'.0.1**1

Table 3: Model estimation results for the random-effect model of daily distance per person

	Parameter	S.E.	p-value	
Constant	-0.6172	3.1236	0.8434	
RWI percentile	2.6289	0.7991	0.001	***
Lag1_trip-distance/person	0.3795	0.0143	0	***
Lag1_trips/person	2.3935	0.2979	0	***
Age_0_14	0.1132	0.0519	0.0294	*
Age_65+	-0.9585	0.21	0	***
Health facilities	0.6659	0.2358	0.0048	**
New cases	-0.1019	0.0394	0.0097	**
First case	0.0264	0.2952	0.9289	
Travel ban	-0.3543	0.4275	0.4074	
Fed-lockdown-p1	-1.1668	0.4632	0.0118	*
Fed-lockdown-p2	-0.0743	0.4462	0.8677	

Significant codes: 0***0.001***0.01**0.05'.0.1**1

As for the impact of new SARS-CoV-2 cases, it was estimated to be statistically insignificant in the trips-per-person model. On the other hand, the daily new COVID-19 case number³⁷ was found to influence the daily distance travelled per person negatively. One more confirmed case led to a mild reduction in the daily trip distance per person. Again, the model corroborates what has been observed in the data. People may give up longer-distance travel, such as vacations and visiting family, amid public health concerns. This was as expected and verified by several previous research studies.¹²⁻¹⁴ The announcement of the first SARS-CoV-2 case in Nigeria had a slightly significantly positive effect on trips per person, while it did not have a significant effect on trip distance per person. The trips per person on the date on which the first case was announced were only more than they had been on the dates before the announcement was made, on average.

The model also tested the influence of mobility-restriction policies. The travel ban policy showed a negative, but limited effect on trips per person, as expected, and did not have a significant effect on trip distance per person. The impact of the Phase 1 and Phase 2 federal lockdown policy had an insignificant effect on trips per person, while the Phase 1 federal lockdown policy had a significantly negative effect on trip distance per person. The Phase 1 federal lockdown policy reduced to a noteworthy daily trip distance per person, which is about 9.47% of its mean value. This indicates that the Phase 1 lockdown policy was effective in restricting human mobilities and subsequently slowed down the propagation of the virus. The effectiveness of the Phase 2 federal lockdown, however, had a limited effect in further restricting human mobility. These findings are roughly in line with international studies on the effect of different Phase 1 policies in countries such as the USA^{14,41} and Japan⁴².

From the view of the socio-economic and sociodemographic factors, RWI percentile, Age_0_14 and Age_65+ had a significantly negative influence on trips per person. In other words, relatively wealthier states with a higher proportion of 0–14 or 65+ age groups were more likely to make fewer daily trips per person. On the other hand, Age_65+ had a significantly negative impact on trip distance per person, while, on the contrary, RWI percentile and Age_0_14 had a significantly positive impact on trip distance per person. That means the states with a higher proportion of older people were inclined to travel a shorter distance, while the relatively wealthier states with a higher proportion of younger people tended to travel longer distances. The number of health facilities had significant influences in both models, but the impact was in different directions. States with more health facilities tended to make fewer daily trips per person, but travelled longer daily distance per person. While this was significant, the impact of health facilities was limited due to the small coefficients compared with their scale.

Conclusion

This paper is one of the first attempts to quantify travel behaviour (i.e. trip-making behaviour and daily travelling distances) and its changes at the beginning of the SARS-CoV-2 pandemic in Nigeria. The study is part of a research consortium entitled 'Role of Data Streams in Informing Infection Dynamics in Africa' (INFORM-Africa, <https://dsi-africa.org/project/7>). Fully recognising the data gap in sub-Saharan Africa in understanding human mobility, the study employs a data-driven analytical framework that utilises passively collected smartphone location data and algorithms that have been previously developed and tested. The proposed approach enables the quantification of daily human mobility in terms of the number of trips and distances travelled by each person. The paper demonstrated this novel data-driven approach and how it can fill the critical data gap using Nigeria as the application. The measurements were produced for the period 01 January to 25 April 2020. A steady decrease in daily distances travelled per person during the pandemic was evident, while the daily number of trips travelled fluctuated and increased slightly. The mobility measurements were clustered into three groups based on level of wealth. Our study was able to highlight differences in mobility trends at the state level, revealing spatial and temporal differences in mobility patterns during a pandemic.

Another highlight of the paper is the assessment of the effectiveness of mobility-restricting policies as key lessons learned from the SARS-CoV-2 pandemic. We found that travel bans and federal lockdown policies failed to restrict trip-making behaviour, but had a significant impact on distance travelled. This led to a corollary in that people changed their mobility patterns by reducing their number of long-distance trips and replaced them with shorter trips. While this suggested some moderate policy effects of the government's orders, how it eventually benefitted (or deteriorated) the public health situation under the pandemic needs further evaluation. The fewer longer trips could mean fewer gatherings at long-distance bus stations and airports. But would the increased number of shorter local trips (and activities) lead to new public health hotspots and community transmissions? The proposed approach can be further developed to look at location-specific population density and assess how that influenced local outbreaks.

This paper contributes a first attempt to quantify human travel behaviour and how mobility-restricting policies took effect. Its innovations are three-fold:

- The mobility measurements are new to the field and have filled a major data gap in understanding how people travel and how travel behaviour changed during the SARS-CoV-2 pandemic. The mobility data could facilitate a variety of quantitative research studies related to transportation and health that could not have been done before.
- The study is driven entirely by high-quality data samples and a data-driven analytical framework. The framework can be directly applied to study other localities and periods of interest. The framework can also be adapted to develop additional measurements and quantitative models with regard to human mobility patterns and relevant policies and regulations.
- The study assesses the mobility-restricting policies via a time-dependent random effect modelling practice. It quantifies the effectiveness of those policies and derives policy implications that are critical to develop travel-related guidelines in response to future unprecedented epidemics and other infectious diseases.

Limitations of this study are acknowledged and will be the next research focus. Admittedly, the study did not assess human movement between areas, but focused primarily on understanding the magnitude of daily human travel behavioural patterns and changes under the influence of pandemic and mobility-restricting policies. As an immediate future research direction, information on trip origins and destinations will be incorporated to develop spatio-temporal models of human mobility. The data representativeness was also not studied in terms of how the quantified mobility resembles the actual mobility of an average Nigerian. A large-scale household travel survey is not yet available in Nigeria to be used as a benchmark for comparison. When developing human movement patterns between areas and spatio-temporal models, we will gather transportation network observations such as traffic data on highways and tollways, and air and rail ticket sales as possible ways to verify the data representativeness.

Secondly, the raw location data employed in this study was licensed from third-party smartphone location data providers. As it is strictly prohibited to make such highly sensitive data publicly available, part of the study involving raw location data processing cannot be replicated. We acknowledge this limitation. Such limitation will last long and hold true for any research employing location-based service data. Possible solutions may be developed using artificial intelligence generated contents and synthetic data-generation methods. This will be a promising research direction to enable a general understanding and a wider acceptance of using such location-based data.

The research team will also work on expanding the study to South Africa. As another important sub-Saharan African country with a significant population, South Africa's human mobility behaviour under different waves of SARS-CoV-2 variants would offer valuable empirical evidence for policymakers and health practitioners. The team plans to conduct a bi-country statistical comparison of Nigeria and South Africa to analyse the differences and similarities between the two countries when facing a pandemic.

Acknowledgements

The research approach adopted in the study, including data sources, data processing and modelling methodologies, was reviewed and approved by the Institutional Review Boards at the Villanova University and the University of Maryland Baltimore and the National Health Research Ethics Committee of Nigeria (NHREC). This work was financially supported by a US National Institutes of Health (NIH) award (grant number U54TW012041-02) entitled "Role of Data Streams in Informing Infection Dynamics in Africa – INFORM-Africa". The views and opinions stated in this paper are those of the authors and do not necessarily reflect the views or positions of the project sponsor.

Competing interests

We have no competing interests to declare.

Authors' contributions

W.L.: Methodology, data collection, sample analysis, data analysis, validation, data curation, writing – the initial draft, writing – revisions, project management. C.X.: Conceptualisation, methodology, data collection, sample analysis, data analysis, validation, data curation, writing – the initial draft, writing – revisions, project leadership, funding acquisition. J.W.: Methodology, data collection, data analysis, validation, data curation, writing – revisions. Z.F.: Writing – the initial draft, data analysis, writing – revisions. O.A.: Writing – the initial draft, writing – revisions, data analysis, project management. N.B.: Conceptualisation, writing – the initial draft, writing – revisions. M.C.: Conceptualisation, project leadership, writing – the initial draft, writing – revisions. V.N.: Conceptualisation, project leadership, writing – the initial draft, writing – revisions. C.R.: Conceptualisation, writing – the initial draft, writing – revisions. A.W.: Conceptualisation, writing – the initial draft, writing – revisions, project leadership. F.M.-I.: Conceptualisation, writing – revisions, project management. A.A.: Conceptualisation, writing – revisions, project leadership, funding acquisition.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis.* 2020;20(5):533–534. [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1)
2. World Health Organization (WHO). WHO Coronavirus (COVID-19) Dashboard [webpage on the Internet]. No date [cited 2023 May 26]. Available from: <https://covid19.who.int>
3. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science.* 2020;368(6489):395–400. <https://doi.org/10.1126/science.aba9757>
4. Courtemanche C, Garuccio J, Le A, Pinkston J, Yelowitz A. Strong social distancing measures in the United States reduced the COVID-19 growth rate. *Health Affairs.* 2020;39(7):1237–1246. <https://doi.org/10.1377/hlthaff.2020.00608>
5. Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: An observational study. *Lancet Public Health.* 2020;5(5):e279–88. [https://doi.org/10.1016/S2468-2667\(20\)30090-6](https://doi.org/10.1016/S2468-2667(20)30090-6)
6. Lee M, Zhao J, Sun Q, Pan Y, Zhou W, Xiong C, et al. Human mobility trends during the early stage of the COVID-19 pandemic in the United States. *PLoS ONE.* 2020;15(11), e0241468. <https://doi.org/10.1371/journal.pone.0241468>
7. White ER, Hébert-Dufresne L. State-level variation of initial COVID-19 dynamics in the United States. *PLoS ONE.* 2020;15(10), e0240648. <https://doi.org/10.1371/journal.pone.0240648>
8. Perra N. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Phys Rep.* 2021;913:1–52. <https://doi.org/10.1016/j.physrep.2021.02.001>
9. Chi G, Fang H, Chatterjee S, Blumenstock JE. Microestimates of wealth for all low- and middle-income countries. *Proc Natl Acad Sci USA.* 2022;119(3), e2113658119. <https://doi.org/10.1073/pnas.2113658119>
10. Luo W, Guo W, Hu S, Yang M, Hu X, Xiong C. Flatten the curve: Empirical evidence on how non-pharmaceutical interventions substituted pharmaceutical treatments during COVID-19 pandemic. *PLoS ONE.* 2021;16(10), e0258379. <https://doi.org/10.1371/journal.pone.0258379>
11. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science.* 2020;368(6489):395–400. <https://doi.org/10.1126/science.aba9757>
12. Hu S, Xiong C, Yang M, Younes H, Luo W, Zhang L. A big-data driven approach to analyzing and modelling human mobility trend under non-pharmaceutical interventions during COVID-19 pandemic. *Transp Res Part C Emerg Technol.* 2021;124:102955. <https://doi.org/10.1016/j.trc.2020.102955>

13. Xiong C, Hu S, Yang M, Luo W, Zhang L. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proc Natl Acad Sci USA*. 2020;117(44):27087–27089. <https://doi.org/10.1073/pnas.2010836117>
14. Xiong C, Hu S, Yang M, Younes H, Luo W, Ghader S, et al. Mobile device location data reveal human mobility response to state-level stay-at-home orders during the COVID-19 pandemic in the USA. *J R Soc Interface*. 2020;17(173), Art. #20200344. <https://doi.org/10.1098/rsif.2020.0344>
15. Kraemer MUG, Tegally H, Pigott DM, Dasgupta A, Sheldon J, Wilkinson E, et al. Tracking the 2022 monkeypox outbreak with epidemiological data in real-time. *Lancet Infect Dis*. 2022;22(7):941–942. [https://doi.org/10.1016/S1473-3099\(22\)00359-0](https://doi.org/10.1016/S1473-3099(22)00359-0)
16. Kraemer MUG, Yang C-H, Gutierrez B, Wu C-H, Klein B, Pigott DM, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*. 2020;368(6490):493–497. <https://doi.org/10.1126/science.abb4218>
17. Håkansson A. Changes in gambling behaviour during the COVID-19 pandemic—a web survey study in Sweden. *Int J Environ Res Public Health*. 2020;17(11):4013. <https://doi.org/10.3390/ijerph17114013>
18. Ugolini F, Massetti L, Calaza-Martínez P, Carriñanos P, Dobbs C, Stoicik SK, et al. Effects of the COVID-19 pandemic on the use and perceptions of urban green space: An international exploratory study. *Urban For Urban Green*. 2020;56:126888. <https://doi.org/10.1016/j.ufug.2020.126888>
19. Jarvis CI, Van Zandvoort K, Gimma A, Prem K, CMMID COVID-19 working group, Klepac P, et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med*. 2020;18(1):124. <https://doi.org/10.1186/s12916-020-01597-8>
20. Zhang J, Litvinova M, Liang Y, Zheng W, Shi H, Vespignani A, et al. The impact of relaxing interventions on human contact patterns and SARS-CoV-2 transmission in China. *Sci Adv*. 2021;7(19), eabe2584. <https://doi.org/10.1126/sciadv.abe2584>
21. Tintori A, Cerbara L, Ciancimino G, Crescimbeni M, La Longa F, Versari A. Adaptive behavioural coping strategies as reaction to COVID-19 social distancing in Italy. *Eur Rev Med Pharmacol Sci*. 2020;24(20):10860–10866. https://doi.org/10.26355/eurrev_202010_23449
22. Alzueta E, Perrin P, Baker FC, Caffarra S, Ramos-Usuga D, Yuksel D, et al. How the COVID-19 pandemic has changed our lives: A study of psychological correlates across 59 countries. *J Clin Psychol*. 2021;77(3):556–570. <https://doi.org/10.1002/jclp.23082>
23. Balanzá-Martínez V, Kapczynski F, De Azevedo Cardoso T, Atienza-Carbonell B, Rosa AR, Mota JC, et al. The assessment of lifestyle changes during the COVID-19 pandemic using a multidimensional scale. *Rev Psiquiatr Salud Ment*. 2021;14(1):16–26. <https://doi.org/10.1016/j.rpsm.2020.07.003>
24. Rogers NT, Waterlow NR, Brindle H, Enria L, Eggo RM, Lees S, et al. Behavioural change towards reduced intensity physical activity is disproportionately prevalent among adults with serious health issues or self-perception of high risk during the UK COVID-19 lockdown. *Front Public Health*. 2020;8, Art. #575091. <https://doi.org/10.3389/fpubh.2020.575091>
25. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *Lancet Infect Dis*. 2020;20(11):1247–1254. [https://doi.org/10.1016/S1473-3099\(20\)30553-3](https://doi.org/10.1016/S1473-3099(20)30553-3)
26. Hsiehchen D, Espinoza M, Slovic P. Political partisanship and mobility restriction during the COVID-19 pandemic. *Public Health*. 2020;187:111–114. <https://doi.org/10.1016/j.puhe.2020.08.009>
27. Weill JA, Stigler M, Deschenes O, Springborn MR. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc Natl Acad Sci USA*. 2020;117(33):19658–19660. <https://doi.org/10.1073/pnas.2009412117>
28. Evans MV, Garchitorena A, Rakotonanahary RJL, Drake JM, Andriamihaja B, Rajanarifara E, et al. Reconciling model predictions with low reported cases of COVID-19 in sub-Saharan Africa: Insights from Madagascar. *Glob Health Action*. 2020;13(1), Art. #1816044. <https://doi.org/10.1080/16549716.20.1816044>
29. Van Zandvoort K, Jarvis CI, Pearson CAB, Davies NG, Ratnayake R, Russell TW, et al. Response strategies for COVID-19 epidemics in African settings: A mathematical modelling study. *BMC Med*. 2020;18, Art. #324. <https://doi.org/10.1186/s12916-020-01789-2>
30. WHO Regional Office for Africa. Tuberculosis (TB) [webpage on the Internet]. No date [cited 2022 Sep 01]. Available from: <https://www.afro.who.int/health-topics/tuberculosis-tb>
31. Statista. Penetration rate of smartphones in selected countries [webpage on the Internet]. c2021 [cited 2022 Sep 01]. Available from: <https://www.statista.com/statistics/539395/smartphone-penetration-worldwide-by-country/>
32. Federal Highway Administration (FHWA). NHTS NextGen OD Data [webpage on the Internet]. No date [cited 2022 Sep 01]. Available from: <https://nhts.ornl.gov/od/>
33. Zhang L, Darzi A, Pan Y, Yang M, Sun Q, Kabiri A, et al. Next generation National Household Travel Survey national origin destination data passenger origin-destination data methodology documentation. Washington DC: Federal Highway Administration, US Department of Transportation; 2021. Available from: https://nhts.ornl.gov/od/assets/doc/2020_NextGen_NHTS_Passenger_OD_Data_Methodology_v2.pdf
34. The Center for Policy Impact in Global Health. Nigeria's policy response to COVID-19 [webpage on the Internet]. No date [cited 2023 Apr 04]. Available from: <https://centerforpolicyimpact.org/our-work/the-4ds/nigeria-policy-response-to-covid-19/>
35. Jacobs ED, Malachy IO. A critical evaluation of Nigeria's response to the first wave of COVID-19. *Bulletin of the National Research Centre*. 2022;46(1):44. <https://doi.org/10.1186/s42269-022-00729-9>
36. South African Government. Minister Joe Phaahla: Repeal of regulations regarding Covid-19 pandemic and monkey-pox [media release on the Internet]. 23 June 2022 [cited 2023 May 02]. Available from: <https://www.gov.za/speeches/statement-minister-health-dr-joe-phaahla-repeal-regulations-notifiable-medical-conditions>
37. HERA. HERA – the Covid-19 data project [webpage on the Internet]. No date [cited 2023 Apr 05]. Available from: <https://hera-ngo.org/projects/the-covid-19-data-project>
38. openAFRICA. Nigerian health care facilities (primary, secondary and tertiary) [data set]. c2021 [updated 2021 Jun 29; cited 2022 Sep 01]. Available from: <https://africaopendata.org/dataset/nigerian-health-care-facilities-primary-secondary-and-tertiary1>
39. Cezar M, Tiba A, Basarin B, Vujičić M, Valjarević A, Niemets L, et al. Predictors of changes in travel behavior during the COVID-19 pandemic: The role of tourists' personalities. *Int J Environ Res Public Health*. 2021;18(21): Art. #111169. <https://doi.org/10.3390/ijerph182111169>
40. Qureshi MA, Ho-Ling H, Shih-Miao C. Comparison of distance estimates for Commodity Flow Survey: Great circle distances versus network-based distances. *Transp Res Rec*. 2002;1804(1):212–216. <https://doi.org/10.3141/1804-28>
41. Li Y, Li M, Rice M, Zhang H, Sha D, Li M, et al. The impact of policy measures on human mobility, COVID-19 cases, and mortality in the US: A spatiotemporal perspective. *Int J Environ Res Public Health*. 2021;18(3), Art. #996. <https://doi.org/10.3390/ijerph18030996>
42. Yabe T, Tsubouchi K, Fujiwara N, Wada T, Sekimoto Y, Ukkusuri SV. Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic. *Sci Rep*. 2020;10(1), Art. #18053. <https://doi.org/10.1038/s41598-020-75033-5>



Public health research using cell phone derived mobility data in sub-Saharan Africa: Ethical issues

AUTHORS:

Stuart Rennie^{1,2}
Caesar Atuire³
Tiwonge Mtande⁴
Walter Jaoko⁵
Sergio Litewka⁶
Eric Juengst²
Keymanthri Moodley⁴

AFFILIATIONS:

¹Department of Social Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

²UNC Center for Bioethics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

³Department of Philosophy and Classics, University of Ghana, Accra, Ghana

⁴Centre for Medical Ethics and Law, Department of Medicine, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

⁵KAVI-Institute of Clinical Research, University of Nairobi, Nairobi, Kenya

⁶Institute for Bioethics and Health Policy, Miller School of Medicine, University of Miami, Miami, Florida, USA

CORRESPONDENCE TO:

Stuart Rennie

EMAIL:

stuart_rennie@med.unc.edu

DATES:

Received: 14 Sep. 2022

Revised: 09 May 2023

Accepted: 12 May 2023

Published: 30 May 2023

HOW TO CITE:

Rennie S, Atuire C, Mtande T, Jaoko W, Litewka S, Juengst E, et al. Public health research using cell phone derived mobility data in sub-Saharan Africa: Ethical issues. *S Afr J Sci.* 2023;119(5/6), Art. #14777. <https://doi.org/10.17159/sajs.2023/14777>

ARTICLE INCLUDES:

- Peer review
- Supplementary material

DATA AVAILABILITY:

- Open data set
- All data included
- On request from author(s)
- Not available
- Not applicable

EDITOR:

Floretta Boonzaier

KEYWORDS:

ethics, mobility data, public health, community engagement, surveillance

FUNDING:

US National Institutes of Health (U01MH127704)

The movements of humans have a significant impact on population health. While studies of such movements are as old as public health itself, the COVID-19 pandemic has raised the profile of mobility research using digital technologies to track transmission routes and calculate the effects of health policies, such as lockdowns. In sub-Saharan Africa, the high prevalence of cell phone and smartphone use is a source of potentially valuable mobility data for public health purposes. Researchers can access call data records, passively collected in real time from millions of clients by cell phone companies, and associate these records with other data sets to generate insights, make predictions or draw possible policy implications. The use of mobility data from this source could have a range of significant benefits for society, from better control of infectious diseases, improved city planning, more efficient transportation systems and the optimisation of health resources. We discuss key ethical issues raised by public health studies using mobility data from cell phones in sub-Saharan Africa and identify six key ethical challenge areas: autonomy, including consent and individual or group privacy; bias and representativeness; community awareness, engagement and trust; function creep and accountability; stakeholder relationships and power dynamics; and the translation of mobility analyses into health policy. We emphasise the ethical importance of narrowing knowledge gaps between researchers, policymakers and the general public. Given that individuals do not really provide valid consent for the research use of phone data tracking their movements, community understanding and input will be crucial to the maintenance of public trust.

Significance:

- Mobility data derived from cell phones are being increasingly used for health research and public health purposes in sub-Saharan Africa, with minimal individual consent and largely without public awareness.
- While such data can have significant potential public health benefits, risks and concerns related to their collection and use in sub-Saharan African contexts have not been widely discussed.
- Innovative community engagement initiatives, which are appropriate and responsive to sub-Saharan African contexts, need to be developed to address ethical challenge areas and help warrant public trust in mobility research.

Introduction

The use of big data for public health promotion, clinical care improvement and health system strengthening is increasingly globalised. Until the recent past, the practice of collecting, merging, storing and using large data sets for these purposes was mostly limited to high-income countries. This is no longer the case. Persistent health challenges and the rising integration of digital technologies in the daily lives of people in sub-Saharan Africa (SSA) have engendered increased interest in big data initiatives throughout the region. Traditional health statistics, such as those gathered by government agencies through demographical and health surveys, are unlikely to be as comprehensive as vast volumes of health-related data gathered in real time from a diversity of born digital sources by both public agencies and private companies.

The development and adoption of Internet telephony has been of particular interest to those who have worked for many years to improve health in SSA. In the colonial and post-colonial periods, the use of landlines in Africa was limited to government offices, businesses in major cities and socio-economic elites. Over the past two decades, the rarity of landlines has been eclipsed by the ubiquitous use of cell phones.¹ According to the Global System for Mobile Communications report, *The Mobile Economy Sub-Saharan Africa 2022*, smartphones are being rapidly adopted: on average accounting for 49% of connections in 2022, smartphones are predicted to account for 61% of connections by 2025. However, this means that nearly 40% of people in SSA use basic cell phones or have no mobile phone access at all.² The widespread use of cell phones, in turn, has sparked the rise of mobile health – or mHealth – initiatives and research studies that seek to improve health and healthcare services by enhancing communication between patients and healthcare providers, study participants and researchers, and citizens and public health professionals. Many mHealth applications (apps) for public health purposes have been developed, implemented and evaluated in low- and middle-income countries (LMICs).³⁻⁵ Tomlinson et al.⁶ reported on the use of cell phones by community health workers in South Africa to conduct electronic household surveys and questionnaires. Cell phone surveys could be a cost-effective approach to gather data about non-communicable diseases in LMICs.⁷ Similarly, Brinkel et al.⁸ describe the use of cell phones in public health surveillance, i.e. where health workers gather patient (and sometimes Global Positioning System [GPS]) data on cell phones and send the information to the server of a local cell phone service provider, upon which the data are forwarded to district health offices and research institutes.

This paper addresses ethical issues that arise with one particularly powerful form of such big data research: the use of cell phone data to track human mobility patterns in efforts to improve public health in SSA. How people move affects population health, most obviously in the case of infectious disease epidemics, such as the West

African Ebola outbreak⁹ and the COVID-19 pandemic¹⁰⁻¹². This research utilises call data records (CDRs) that cell phone companies passively collect in real time from millions of clients. CDRs list the cellular base station or tower and the code of the subscriber identification module (SIM) card involved in each call or text. If these data are available, in conjunction with a map of the relevant towers, the location of each call or text can be identified and, from this, an individual's movement between calls can be derived.¹³ The resulting data, which often involves hundreds of thousands of trajectories, can be associated with other data sets to generate insights, make predictions or draw possible policy implications. For example, Gibbs et al.¹⁴ used data from CDRs in Ghana to identify the relationship between reductions in human mobility and decreases in the real-time reproduction number (R_t) of COVID-19 during the early stages of the pandemic. In addition to CDRs, other types of location data are collected through cell phone apps or the Bluetooth functionality on smartphones. Although this paper focuses on data from CDRs, the ethics discussion is also applicable to other types of cell phone location data.

Cell phone companies can share data from CDRs with third parties (including researchers) in different ways. De Montjoye et al.¹⁵ offer four 'privacy-conscious' models of CDR data sharing, where anonymisation and the spatial or temporal aggregation of shared data are central. According to a limited release model, the telecommunications company technically transforms a cell phone data set so as to make the reidentification of individuals very difficult. According to a precomputed indicators and synthetic data model, third parties are not given transformed data, but receive information derived from a cell phone data set (such as number of calls) or synthetic data that convey the predefined statistical properties of the original data set. According to a remote access model, the data are not released, but stay under the control of the telecommunications company (or authorised authority) and can be conditionally accessed by third parties remotely. Finally, in a question-and-answer model, the data set remains under the control of the telecommunications company and is accessed by third parties through a question-and-answer system: third parties ask specific and standardised questions about a data set, and the telecommunications company provides answers that have been vetted by a security and auditing system. The limited release model is closest to traditional data sharing. Given that it requires fewer technical and human resources than the alternatives, this model is likely the predominant CDR data-sharing model in SSA.

The research use of cell phone derived mobility data is likely to lead to a variety of public health benefits in SSA, even if it is difficult to identify or quantify them at this early stage. Drawing lessons from the COVID-19 pandemic response, Oliver et al.¹⁶ describe four areas in which mobility data can be beneficial for epidemic control: understanding the dynamic environment of an epidemic; tackling cause-and-effect questions by identifying the key mechanisms and consequences of epidemic containment; predicting the likelihood of future outcomes; and developing impact assessments to determine how various interventions impact epidemic spread. To optimise these and other social benefits, it will be important for researchers and data managers to anticipate the ethical and social challenges that may arise along the way. Our focus here is on key ethical challenges raised by the use of these data in the SSA context, particularly considering that there are no formal ethics guidelines specific to mobility science, and such research is likely to be unfamiliar to most research ethics committees in the region. This review makes use of diverse literature: mobility research related to development, migration and humanitarian crises; anthropological research on cell phone use in Africa; current debates about mobility justice; and mobility data related to health promotion in LMICs, particularly in Africa. The core themes are also derived from the involvement of the authors in the Data Science for Health Discovery and Innovation in Africa (DS-I Africa) initiative and from relevant discussions within bioethics circles in SSA. They are not additionally embedded within dominant bioethics frameworks, which originate from high-income countries and whose universality has been placed in question in current discourses surrounding the decolonisation of bioethics.¹⁷ Based on this review, we address six challenge areas: autonomy, including consent and individual or group privacy; bias and representativeness; community awareness, engagement and trust;

function creep and accountability; stakeholder relationships and power dynamics; and the translation of mobility analyses into health policy.

Autonomy

Respect for the autonomy of research participants is a core value in research, usually represented in terms of obtaining their free and informed consent, protecting participants' privacy, and ensuring their ability to withdraw at any time. However, each of these standard methods of respecting the rights and interests of human data sources faces particular challenges in the research use of cell phone derived mobility data in the SSA context.

In their privacy policies and the terms of their service agreements, cell phone companies often disclose that client data may be shared with third parties. Mobility researchers are among these parties, and it could be argued that a client who agreed to the cell phone companies' privacy policies and terms of service therefore consented to the collection, sharing and use of their call records. Although this might be legally adequate, whether this is an ethically valid form of consent has long been debated.¹⁸ Extensive, densely written and sometimes buried policies, which require the client to agree in full or forgo the service, are often poorly read or understood, drawing doubt on whether consent is voluntary and informed.¹⁹ This is even more relevant in the African rural context where users of cell phones may have less information about the technological infrastructure underlying cell phone use and how data are collected. Cell phone operators generally do not offer consent forms in local languages, making it difficult for the less literate to comprehend to what they may be consenting. In addition, it is not clear that the routine collection and/or potential research use of CDRs is clearly disclosed in the policies of the major cell phone providers in SSA. In addition, awareness of research using CDRs among typical cell phone users in SSA is likely to be extremely low. For these reasons, it would not be uncharitable to mark this as an unconsented use of cell phone derived location data. The approach can be contrasted with other mobility research designs where participants are asked to explicitly agree to the collection and use of their data, such as the study by De Gruchy et al.²⁰, which piloted the use of WhatsApp for the administration of surveys and collection of location data.

Arguably, the unconsented use of CDRs can be ethically acceptable if the data are fully anonymised, i.e. if the privacy dimension of autonomy is absolutely protected by rendering the reidentification of data sources impossible by all parties. Full anonymity is to be contrasted with deidentification (the removal of a person's identifying characteristics, such as date of birth, from a data set) and pseudonymisation (where personal identifiers are replaced by pseudonyms or codes and cannot be attributed to a specific person without the aid of additional information).²¹ The latter imply that measures have been taken to render identification difficult, but not impossible. If the data are fully anonymised and reported in the aggregate, a balance can be struck between the potential social benefits of mobility research and the right to privacy. Such a balance is ethically ideal, but fragile, as the potential for reidentification needs to be continuously revisited as data sets are merged and algorithms become more sophisticated.²² Mobility data, as such, carries a risk of reidentification because movement patterns, even those of deidentified individuals, can reveal personal and possibly sensitive information.²³ Repeated and frequent trajectories between one location (likely a person's home or work), even of deidentified individuals, can reveal personal and possibly sensitive information, unless special care has been taken to 'coarsen' or aggregate spatial data. Deanonimisation via movement tracking can also pose threats to group privacy. Tracking the movements of displaced groups during humanitarian crises can be important for the provision of care and support, but the same technology can also be used, for example, by authoritarian regimes to track the activities of opposition groups. Even in countries like Ghana, which has a *Data Protection Act (Act 2012)*, there is a long list of exemptions – public order, public safety, public morality, national security or public interest – that enables a government to encroach on the rights of citizens.²⁴ It should be noted that recent mobility research and big data capacity-building efforts in some African countries during the COVID-19 pandemic have established or consolidated close public-private partnerships between research institutions,

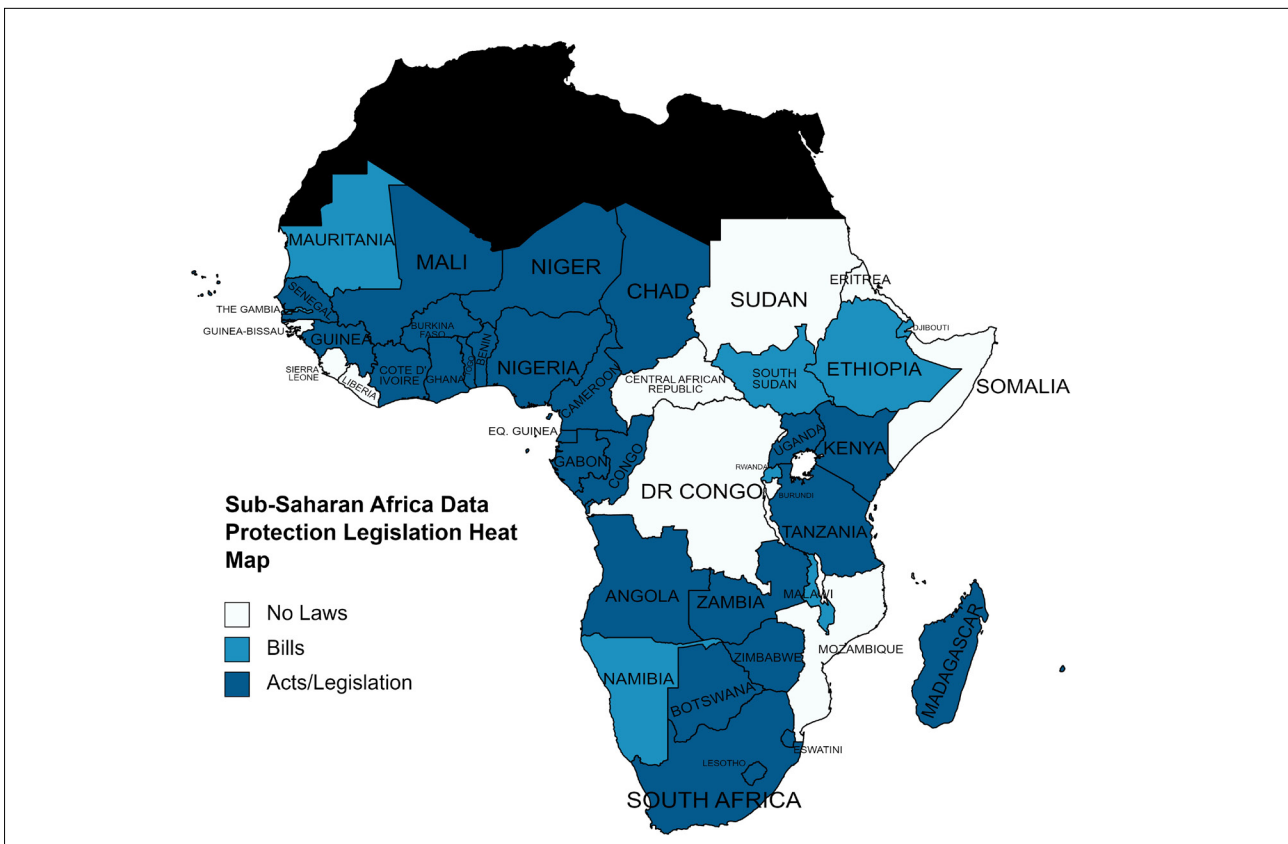
national governments and telecommunications companies. Initiatives in Malawi²⁵, the Democratic Republic of Congo²⁶ and The Gambia²⁷ aim to create a 'pipeline' of cell phone data for public health purposes, from telecommunications companies to ministries of health. How privacy is protected in these and other African pipelines will depend on the existence, nature and enforcement of national data legislation. Figure 1 provides a general overview of the status of data protection legislation in SSA as of April 2023. Enhancing the visibility of groups by governments through mobility data has ethical risks in the SSA context, particularly in relation to politicised ethnic and other divisions.²⁸

Bias and representativeness

The use of reliable scientific methods is a basic ethical requirement of health research.²⁹ Unreliable methods risk producing invalid data, wasting research resources, and potentially harming individuals and communities if the data inform health policy. Like other types of research, mobility research faces challenges in regard to bias, accuracy and representativeness. In the SSA context, concerns have been raised about how to interpret cell phone derived mobility patterns when, for instance, individuals own and use more than one phone, or use different SIM cards in one phone, or when phones are shared by family members or friends.³⁰ In their study in Uganda, Milusheva et al.³¹ noted the potential for bias in mobility data based on how phones are used: Ugandans (particularly those who are struggling financially) often prepay for phone services and use them intermittently. It is common for some phone users to switch off their GPS to conserve their cell phone's battery, or they may be unable to keep their phones on due to the frequent power cuts that are common in many SSA countries. The result is that mobility data may disproportionately represent the movements of those who are male, relatively wealthy and largely urban, and therefore invite misleading inferences.³¹ For example, smartphone-derived data may suggest decreased mobility due to COVID-19 lockdown policies, but the decrease will be less than estimated if the data set does not include or

account for the movements of those without smartphones and without the possibility to work from home.³² A study by Wesolowski et al.³³ linked CDR and socio-economic survey data in Kenya to correct for cell phone ownership bias and attempted to produce a more representative estimate of mobility patterns. Contextualising phone-derived mobility data sets by comparing them with other data sources, and the use of complex filtering techniques are approaches to detect and reduce bias in mobility data sets that will be important considering the behaviour and circumstances of cell phone users in SSA.³⁴

'Mobility justice' refers to how differences in class, gender, race, ethnicity, nationality, sexual identity and physical ability, as well as the built environment, social practices and public policies, influence human movement.³⁵ The capacity and means to move about are not equitably distributed, and in SSA, mobility patterns commonly are shaped by class, gender and racial colonial histories and their legacies in the present. Nyamai and Schramm³⁶ document how the colonial concentration of services and opportunities in Nairobi's Central Business District and the prioritisation of public infrastructure for use by private vehicles compel the majority of citizens to travel long distances by (often inadequate) public transport or on foot. One could speak here of 'thin' and 'thick' conceptions of mobility. According to a thin conception, mobility is simply the movement of bodies through space and time as tracked by mobile devices. Under a thick conception, patterns of movement (or the lack thereof) are the result of a complex network of social determinants. This raises the question whether mobility research can or should engage with thick conceptions of mobility when interpreting patterns of movement and the justice-related considerations attached to them. For example, Deng and Wang³⁵ found persistent disparities in the representativeness of movement data collected by social media and from cell phones in Texas and Louisiana during Hurricane Harvey, i.e. the data best reflected the movements of those in majority-white and non-poor neighbourhoods.



Created with mapchart.net (CC BY 4.0)

Figure 1: This map indicates data protection acts and legislation that govern the collection and use of personal data, which identifies living individuals directly or indirectly. Data protection bills are proposed laws or draft versions of laws or Acts under discussion and have not yet been passed by Parliament.

The precision of movement data also significantly differed by neighbourhood, which (like representativeness) is likely to influence how resources are allocated during natural disasters.³⁷ They concluded that those collecting and interpreting mobility data should be aware of how minorities and low-income communities may become less visible and therefore less likely to receive assistance. In short, to avoid entrenching existing inequalities, mobility research needs to incorporate social justice considerations.

Community awareness, engagement and trust

Although COVID-19 has accelerated public health research using cell phone data, studies of public awareness, knowledge or attitudes towards the use of such data for public health purposes are conspicuous by their absence. Jones et al.³⁸ reviewed published research on the challenges and opportunities of using CDRs in health research, and could not find any literature on public perceptions of using such records. More recently, a scoping review by Sekandi et al.³⁹ on ethical, legal and socio-cultural issues in the use of CDRs for public health in the East African region found no published research on public views on this topic. Relatedly, there appears to have been little or no community engagement in health-related mobility research using CDRs in Africa or elsewhere.³⁹ Revealingly, a recent set of guiding principles to maintain public trust in the use of mobile operator data proposed by a group of statisticians, data analytics specialists and academics does not appear to regard community input as important to the maintenance of trust.⁴⁰ They appear to assume that as long as key stakeholders (i.e. government agencies, data analysis organisations and mobile device operators) follow principles of necessity and proportionality, professional independence, privacy protection, quality control and international comparability, community engagement is unnecessary. On this view, public trust simply follows from stakeholder trustworthiness:

In all, explicitly addressing the five principles in the preparation of a project should give confidence to the statistical agency and its partners, that enough care has been exercised in the set up and implementation of the project, and should convey trust to public and government in the use of mobile operator data for policy purposes.^{40(p.e24-1-e24-21)}

The neglect of community engagement in health-related mobility research using cell phone data is problematic for a number of reasons. First, at least some degree of community engagement is increasingly expected (if not demanded) by regulators and ethics committees as a basic requirement of the responsible conduct of research.⁴¹ Raising awareness that mobility research using CDRs is taking place would be a bare minimum level of engagement. Second, using data from communities without their awareness or input, even if following professional standards, is a potential source of mistrust, suspicion and misinformation. Community acceptance of basically unconsented data collection for public health purposes cannot be assumed. A study by Garrett and Young⁴² on patient views on the use of digital data for public health surveillance suggests that the public may be significantly less comfortable about the collection and use of location data than with data from social media accounts or electronic health records. Third, not involving the public is a lost opportunity for improving the quality of mobility data by addressing possible gaps between how movements are represented and how (and perhaps why) people are actually moving. Lastly, community engagement could help identify risks to community members posed by mobility research that may be invisible to data scientists, particularly those who have little familiarity with the societies they are studying.⁴³

Function creep and accountability

'Function creep' refers to the phenomenon of a technology being used for something other than its originally intended purpose. Drivers' licences, originally meant to promote traffic safety, gradually took on the role of authorised personal identification. The use of CDRs for public health promotion is itself a form of function creep, although ethical concerns about function creep typically are about when the new function of the technology is less benign than its original purpose.

As some in technology studies have observed, the 'creep' in function creep often takes the form of a gradual expansion from a context of care to a context of control.⁴⁴ Here, there is a concern that techniques developed in cell phone data-driven mobility research to promote public health in LMICs will end up being used for questionable surveillance, commercial or political purposes. For example, some experts suggest that uses of cell phone data to help tackle humanitarian crises and infectious diseases in LMICs are very likely to be repurposed to predict, track and prevent unwanted migration.^{45,46} Foreign involvement in the African communication technology sector, particularly that of the Chinese government, has also raised questions about function creep and accountability.⁴⁷ Chinese companies like Huawei and the Transion Group have invested heavily in the cell phone infrastructure of SSA. While the provision of loans, equipment and training initiatives has stimulated needed growth in this sector, the Chinese government and its corporate track record of the digital surveillance of its own citizens raises concern that mobile data from Africans may be transmitted abroad and that African governments are being assisted to use mobile technology to increase social control over their citizens.⁴⁸ The Chinese government and Chinese companies are not the sole focus of these concerns: the actions of Western governments and companies have also increased the risk for 'digital authoritarianism' in Africa.⁴⁹

As more and more digital phone data are being collected, analysed and shared, potentially harmful manifestations of function creep are likely inevitable. A question then is: who is accountable for minimising the risks raised by function creep in regard to cell phone derived mobility data in SSA? COVID-19 and the rise of data digitisation have stimulated the further development of legal and regulatory frameworks for data protection and privacy, although currently the result is a patchwork, with some African countries having few or no relevant policies, while others have extensive governance frameworks.⁵⁰ However, the use of CDR data for public health is relatively new in SSA, and it is unclear how well this particular form of data is covered, even by the most developed policies. In the meantime, ad hoc agreements continue to be reached between cell phone companies, government agencies, researchers and data analytics organisations in their collaborative projects, with each stakeholder answering to their own internal regulatory regime. At present, there do not appear to be overarching governance structures within SSA countries to minimise the risk of cell phone derived mobility data being used for ethically questionable purposes. Developing such structures appears to be a matter of urgency, as African governments leverage cell phone technology to take actions that can significantly impact citizens, while bypassing public debate. In Ghana, for example, the government has made a digital identification card compulsory, while requiring SIM cards from all cell phones to be linked to the digital ID.⁵¹ At the same time, the government has introduced a 1.5% levy on all financial transactions conducted by cell phone, which will disproportionately impact Ghanaians of lower socio-economic status.⁵² This suggests that, while national governments are the 'natural' authorities responsible for minimising negative forms of function creep, governments themselves need to be held publicly accountable for how they make use of cell phone data and why.

Stakeholder relationships and power dynamics

Ethical issues regarding research data governance in Africa have been raised for many years, particularly in response to large-scale genomic research initiatives, such as H3Africa.⁵³ However, what sets mobility data apart is that researchers do not collect the primary data themselves: they depend on commercial entities, i.e. cell phone operators, from whom they gain CDR data through data use agreements. The ethical implications of this relationship in the SSA context are relatively unexplored, but, clearly, cell phone operators become key public health stakeholders in an arrangement where researchers and governments come to depend on them for public health related data. This is part of a larger global issue: digital technology companies are increasingly playing a central and profit-seeking role in public sectors traditionally governed by states (such as emergency response, national security, education and law enforcement), but without being subject to the accountability, transparency or legitimacy of state agencies.⁵⁴ Our main interest here, however, is where mobility researchers and their research institutions are situated within this new

landscape. Although not seeking to make a profit, they use CDR data obtained by commercial entities from individuals with minimally informed and questionably voluntary consent. Even if the mobility data are anonymised, the original commercial consent standards under which the data were obtained falls significantly below traditional research standards for consent, and, given the vast numbers involved, re-consent is out of the question. This arrangement also complicates the ethical review of research: even with 'broad consent' for the use of biological samples, research participants at least know their data are being collected and will be used in future studies. Cell phone users, whose data are collected passively, know far less about the potential use of their data in mobility (and other) research. Under what conditions should research ethics committees approve the use of such data for research purposes? Further, to the extent that mobility analysis informs public policy, mobility researchers and their institutions in this way entrench and normalise the influence of commercial technology firms on the public domain, despite their primary motivation being public health promotion.

In short, the stakeholder relationships in this relatively new field in LMICs risk being marked by dominance, dependence, a lack of transparency and disempowerment along a number of lines: the power held by cell phone companies with vast amounts of citizens' mobility data; the data dependency of governments and researchers on the companies; the lack of control by individuals and communities over the collection, sharing and use of the data collected from them; and the unwillingness or inability of governments to hold companies accountable in ways commensurate with their growing public influence. This risk is particularly significant in SSA countries, where national governments in resource-limited settings with weak health infrastructures may be highly vulnerable to coming under the sway of powerful transnational corporations. In addition, some African governments have a poor track record in regard to public accountability. The commercially mediated use of big data by political authorities could further widen the rift between government and the governed.

Translation of mobility analyses into health policy

Public health research is conducted on the assumption that its findings can be used to improve health by informing relevant policies and practices. Leaving to one side the issues of data representativeness and accuracy, a central question is what health policymakers should do with cell phone derived mobility analyses. As with other emerging and exciting information technologies, there is a risk of this approach being regarded as inherently superior to other ways of generating evidence or to other considerations that are important to health policy decision-making. This can lead to an overly technology-driven policy approach, such as can be seen in critiques of how big data have been utilised in the development of 'smart cities'. As Kitchin⁵⁵ argues, heavy reliance of urban policy on 'real time' big data analytics, combined with a neglect of ethical considerations and the lived experience of city dwellers, threatens to make cities *less* inhabitable.

Current debates about evidence-based policy indicate that, while having good evidence is crucially important, health policy is always, to some extent, underdetermined by empirical evidence.⁵⁶ For example, a number of studies have been conducted comparing the implementation of COVID-19 lockdown policies with mobility as derived from cell phone location data in different settings over specific time periods.⁵⁷ Many mobility studies suggest that stay-at-home policies slowed the spread of COVID-19 at the beginning of the pandemic by inhibiting movement and association, but were less effective as time went on. What does such mobility information imply for future pandemic policymaking? Different policy directions are possible. One could argue that stay-at-home orders have a limited effect over time, and such policies should be used sparingly in the future, particularly in the light of the negative social effects of the large-scale inhibition of movement, i.e. impacts on mental health and child development. One could also argue, with the same data, that stay-at-home orders had a very significant effect on viral spread, and such policies should be more strictly enforced and should be enforced for longer durations in the future. Whatever path is taken will be an evidence-informed result of political, social, legal and ethical deliberations. The same holds for mathematical models, which make use of CDR data to capture (for example) the extent of a disease outbreak or

predict the effects of different health policy options.⁵⁸ Mobility data alone cannot answer health policy questions that are essentially normative in nature, and which incorporate issues of fairness, as well as those of economics and viral control.

In LMICs, including those in SSA, health policymaking is often ad hoc and fragmented in many chronically under-resourced ministries of health.⁵⁹ Human and infrastructural resources will need to be significantly strengthened before many public health systems are in a position to meaningfully utilise the digital data that mobility researchers are gathering. To translate mobility data into valuable policy information, mobility data experts have underlined the importance of developing standardised procedures and mechanisms that are responsive to legal and ethical considerations.¹⁶ Even in high-income countries such as the USA, the massive amounts of digital data collected during the COVID-19 pandemic had little public health impact due to enormous gaps in the translational pipeline.⁶⁰ A situation marked by scarcity of local data science expertise, weak regulatory regimes, little to no community engagement and public health systems not yet prepared to absorb digital data collected from Africans is bound to raise questions about what will be done with the data that continue to be collected, what the real benefits are, and who stands most to gain.

Conclusion

Improving our understanding of human behaviour is vitally important for efforts to improve public health. Insights from cell phone derived mobility data could be beneficial in many contexts, including humanitarian disasters, infectious disease outbreaks and responses to climate change. Considering persistent population health challenges and the sharp growth in cell phone use in SSA, it is understandable that public health researchers, organisations and policymakers are excited about the potential beneficial applications of mobility data. In this paper, we identified key challenges to be taken into account in the collection, sharing, management and use of mobility data in this setting. Moving forward, greater attention will need to be paid to the governance environments in respective SSA countries in regard to this specific type of data, and, in particular, how mobility data are shared between private mobile operators, researchers, national governments and other third parties. It is important to have accurate local knowledge about circumstances where seemingly innocuous information about human movement can become ethically sensitive, such as regions with territorial disputes, jurisdictions that criminalise sexual minorities, or places where religious groups are persecuted. To date, very little social science research has been conducted in SSA about the potential risks of social harm related to mobility data. Social science research on community attitudes about mobility data use is also in its infancy; a recent qualitative study in South Africa suggests that only a minority of those interviewed were concerned about the use of their location data, but also noted that the majority did not really know how that data were being used.⁶¹ Relatedly, increasing the engagement of communities and civil society organisations will be important for the ethical use of mobility data in public health research and policy, especially in efforts to hold both private companies and governments accountable. Local research ethics committees can also contribute to accountability efforts, although their effectiveness will likely depend on increasing knowledge of big data research among committee members.^{62,63} Data ethics tools have been developed in a number of countries (such as The Data Ethics Canvas of the Open Data Institute, The Box by AI Ethics Lab, and the Data Ethics Decision Aid of Utrecht University) that could be of some use for research ethics committees in SSA. In short, there are some identifiable challenges, but much is unknown, and much is left to be done in regard to the ethical use of cell phone derived mobility data in the SSA context.

Acknowledgements

Research reported in this publication was supported by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Dr Kenneth Goodman at the University of Miami for his insights and feedback on earlier versions of this article, and Nezerith Cengiz for her research in support of Figure 1.



Competing interests

We have no competing interests to declare.

Authors' contributions

S.R.: Conceptualisation, literature review and data collection, writing – the initial draft, writing – revisions. C.A.: Conceptualisation, writing – the initial draft, writing – revisions. T.M.: Writing – the initial draft, writing – revisions. W.J.: Writing – the initial draft, writing – revisions. S.L.: Conceptualisation, writing – the initial draft, writing – revisions. E.J.: Conceptualisation, writing – the initial draft, writing – revisions. K.M.: Conceptualisation, literature review and data collection, writing – the initial draft, writing – revisions.




References

- De Bruijn M, Nyamnjoh F, Brinkman I. Mobile phones: The new talking drums of everyday Africa. Bamenda: Langaa Research and Publishing Common Initiative Group; 2009. <https://doi.org/10.2307/j.ctvk3gmgv>
- GSMA Intelligence. The Mobile Economy Sub-Saharan Africa 2022 [document on the Internet]. c2022 [cited 2023 May 26]. Available from: <https://www.gsma.com/mobileeconomy/wp-content/uploads/2022/10/The-Mobile-Economy-Sub-Saharan-Africa-2022.pdf>
- Stephan LS, Almeida ED, Guimaraes RB, Ley AG, Mathias RG, Assis MV, et al. Processes and recommendations for creating mHealth apps for low-income countries. *JMIR Mhealth Uhealth*. 2017;5(4), e41. <https://doi.org/10.2196/mhealth.6510>
- Beratarrechea A, Lee AG, Willner JM, Jahangir E, Ciapponi A, Rubenstein A. The impact of mobile health interventions on chronic disease outcomes in developing countries: a systematic review. *Telemed e-Health*. 2014;20(1):75–82. <https://doi.org/10.1089/tmj.2012.0328>
- Sondaal SFV, Browne JL, Amoakoh-Coleman M, Borgstein A, Miltenburg AS, Verwijs M, et al. Assessing the effect of mHealth interventions in improving maternal and neonatal care in low- and middle-income countries: A systematic review. *PLoS ONE*. 2016;11(5), e0154664. <https://doi.org/10.1371/journal.pone.0154664>
- Tomlinson M, Solomon W, Singh Y, Doherty T, Chopra M, Ijumba P, et al. The use of mobile phones as a data collection tool: A report from a household survey in South Africa. *BMC Med Inform Decis Mak*. 2009;9, Art. #51. <https://doi.org/10.1186/1472-6947-9-51>
- Hyder AA, Wosu AC, Gibson DG, Labrique AB, Ali J, Pariyo GW. Noncommunicable disease factors and mobile phones: A proposed research agenda. *J Med Internet Res*. 2017;19(5), e133. <https://doi.org/10.2196/jmir.7246>
- Brinkel J, Kramer A, Krumkamp R, May J, Fobil J. Mobile-phone based mHealth approaches for public health surveillance in sub-Saharan Africa: A systematic review. *Int J Environ Res Public Health*. 2014;11(11):11559–11582. <https://doi.org/10.3390/ijerph111111559>
- Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ. Commentary: Containing the Ebola outbreak – the potential and challenge of mobile phone data. *PLoS Curr*. 2014;Sept 29, Edition 1. <https://doi.org/10.1371%2Fcurrents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e>
- Tokey AI. Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data. *J Transp Health*. 2021;22, Art. #101135. <https://doi.org/10.1016/j.jth.2021.101135>
- Willberg E, Jarv O, Vaisanen T, Toivonen T. Escaping from cities during the COVID-19 crisis: Using mobile phone data to trace mobility in Finland. *ISPRS Int J Geo-Inf*. 2021;10(2):103. <https://doi.org/10.3390/ijgi10020103>
- Haddawy P, Lawpoolsri S, Sa-ngamuang C, Yin MS, Barkowsky T, Wiratsudakul A, et al. Effects of COVID-19 government travel restrictions on mobility in a rural border area of Northern Thailand: A mobile phone tracking study. *PLoS ONE*. 2021;16(2), e0248542. <https://doi.org/10.1371/journal.pone.0245842>
- Wesolowski A, Buckee CO, Engo-Monsen K, Metcalf CJE. Connecting mobility to infectious diseases: The promise and limits of mobile phone data. *J Infect Dis*. 2016;214(suppl 4):S414–S420. <https://doi.org/10.1093/infdis/jiw273>
- Gibbs H, Liu Y, Abbott S, Baffoe-Nyarko I, Laryea DO, Akyereko E, et al. Association between mobility, non-pharmaceutical interventions, and COVID-19 transmission in Ghana: A modelling study using mobile phone data. *PLoS Glob Public Health*. 2022;2(9), e0000502. <https://doi.org/10.1371/journal.pgph.0000502>
- De Montjoye YA, Gams S, Blondel V, Canright G, De Cordes N, Deletaille S, et al. On the privacy-conscious use of mobile phone data. *Sci Data*. 2018;5(1):1–6. <https://doi.org/10.1038/sdata.2018.286>
- Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci Adv*. 2020;6(23), Art. #eabc0764. <https://doi.org/10.1126/sciadv.abc0764>
- Fayemi AK, Macaulay-Adeyelu OC. Decolonizing bioethics in Africa. *BEOnline*. 2016;3(4):68–90.
- McDonald AM, Cranor LF. The cost of reading privacy policies. *J Law Policy Inform Soc*. 2008;4(3):543–568. <http://hdl.handle.net/1811/72839>
- Obar JA, Oeldorf-Hirsch A. The biggest lie on the Internet: Ignoring privacy policies and terms of service policies of social networking services. *Inf Commun Soc*. 2020;23(1):128–147. <https://doi.org/10.1080/1369118X.2018.1486870>
- De Gruchy T, Vearey J, Opiti C, Mlotshwa L, Manji K, Hanefeld J. Research on the move: Exploring WhatsApp as a tool for understanding the intersections between migration, mobility, health and gender in South Africa. *Global Health*. 2021;17, Art. #71. <https://doi.org/10.1186/s12992-021-00727-y>
- Vokinger K, Stekhoven, D, Krauthammer M. Lost in anonymization – a data anonymization reference classification merging legal and technical considerations. *J Law Med Ethics*. 2020;48(1):228–231. <https://doi.org/10.1177/1073110520917025>
- Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertain Fuzz*. 2002;10(5):571–588. <https://doi.org/10.1142/S021848850200165X>
- Gams S, Killijan M-O, De Prado MN. Show me how you move and I will tell you who you are. *Trans Data Priv*. 2010;4:103–126. <https://doi.org/10.1145/1868470.1868479>
- Government of Ghana. Data Protection Act 2012, art. 60–61 [document on the Internet]. c2012 [cited 2023 May 26]. Available from: <https://www.dataprotection.org.gh/media/attachments/2021/11/05/data-protection-act-2012-act-843.pdf>
- Green D, Moszczynski M, Asbah S, Morgan C, Klyn B, Foutry G, et al. Using mobile data for epidemic response in low resource settings – a case study of COVID-19 in Malawi. *Data Policy*. 2021;3, e19. <https://doi.org/10.1017/dap.2021.14>
- Gueguen C, Snel N, Mutonji E. Turning big data insights into public health responses in the times of pandemics: Lessons learnt from the Democratic Republic of Congo. *Data Policy*. 2022;4, e8. <https://doi.org/10.1017/dap.2021.30>
- Arai A, Knippenberg E, Meyer M, Witayangkurn A. The hidden potential of call detail records in The Gambia. *Data Policy* 2021;3, e9. <https://doi.org/10.1017/dap.2021.7>
- Taylor L. Safety in numbers? Group privacy and big data analytics in the developing world. In: Taylor L, Floridi L, Van der Sloot B, editors. *Group privacy*. Philosophical Studies Series vol. 126. Cham: Springer; 2017. p. 13–36. https://doi.org/10.1007/978-3-319-46608-8_2
- Emanuel EJ, Wendler D, Killen J, Grady C. What makes clinical research in developing countries ethical? The benchmarks of ethical research. *J Infect Dis*. 2004;189(5):930–937. <https://doi.org/10.1086/381709>
- Erikson SL. Cell phones as an anticipatory technology: Behind the hype of big data for Ebola detection and containment. In: Engel U, Rottenburg R, editors. *Adaptation and creativity in Africa: Technologies and significations in the making of order and disorder*. Working Papers of the Priority Programme 1448. Leipzig/Halle: German Research Foundation; 2018. p. 2–14. https://lost-research-group.org/wp-content/uploads/2018/01/WP24_Erikson_180115.pdf
- Milushesva S, Bjorkegren D, Viotti L. Assessing bias in smartphone mobility estimates in low income countries. *COMPASS '21: ACM SIGCAS Conference on Computing and Sustainable Societies*; 2021 June 28 – July 02. New York: Association for Computing Machinery; 2021;364–378. <https://doi.org/10.1145/3460112.3471968>



32. Schellhase J. Using Google mobility data to access COVID-19 mitigation strategies in East Africa [webpage on the Internet]. c2020 [cited 2023 May 17]. Available from: <https://milkeninstitute.org/article/covid-19-google-mobility-data-africa>
33. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO. The impact of biases in mobile phone ownership on estimates of human mobility. *J R Soc Interface*. 2013;10, Art. #20120986. <https://doi.org/10.1098/rsif.2012.0986>
34. Rodriguez-Carrion A, Garcia-Rubio C, Campo C. Detecting and reducing biases in cellular-based mobility data sets. *Entropy*. 2018;10, Art. #736. <https://doi.org/10.3390/e20100736>
35. Sheller M. Theorizing mobility justice. In: Cook N, Butz D, editors. *Mobilities, mobility justice and social justice*. London: Routledge; 2018. p. 22–36. <https://doi.org/10.4324/9780815377047-2>
36. Nyamai DN, Schramm S. Accessibility, mobility, and spatial justice in Nairobi, Kenya. *J Urban Aff*. 2022;45(1):367–389. <https://doi.org/10.1080/07352166.2022.2071284>
37. Deng H, Wang Q. Examining mobility data justice during 2017 Hurricane Harvey. *arXiv:2103.2021*, Art. #13879. <https://doi.org/10.48550/arXiv.2103.13879>
38. Jones KH, Daniels H, Heys S, Ford DV. Challenges and potential opportunities of mobile phone call detail records in health research. *JMIR Mhealth Uhealth*. 2018;6(7), e161. <https://doi.org/10.2196/mhealth.9974>
39. Sekandi JN, Murray K, Berryman C, Davis-Olwell P, Hurst C, Kakaire R, et al. Ethical, legal and sociocultural issues in the use of mobile technologies and call detail records for public health in the East Africa region: Scoping review. *Interact J Med Res*. 2022;11(1), e35062. <https://doi.org/10.2196/35062>
40. Jansen R, Kovacs K, Esko S, Saluveer E, Sostra K, Bengtsson L, et al. Guiding principles to maintain public trust in the use of mobile operator data for policy purposes. *Data Policy*. 2021;3, e24. <https://doi.org/10.1017/dap.2021.21>
41. Adhikari B, Pell C, Cheah PY. Community engagement and ethical global health research. *Glob Bioeth*. 2020;31(1):1–12. <https://doi.org/10.1080/1287462.2019.1703504>
42. Garrett R, Young SD. Ethical views on sharing digital data for public health surveillance: Analysis of survey data among patients. *Front Big Data*. 2022;5, Art. #871236. <https://doi.org/10.3389/ftdata.2022.871236>
43. Taylor L, Schroeder R. Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal*. 2015;80:503–518. <https://doi.org/10.1007/s10708-014-9603-5>
44. Lyon D. Surveillance society. Presented at: Festival del Diritto, Piacenza, Italia, 28 September 2008 [document on the Internet]. c2008 [[cited 2023 May 17]]. Available from: http://www.festivaldeldiritto.it/2008/pdf/interventi/david_lyon.pdf
45. Luca M, Barlacchi G, Oliver N, Lepri B. Levering mobile phone data for migration flows. *arXiv:2105.2021*, Art. #14956. <https://doi.org/10.48550/arXiv.2105.14956>
46. Vinck P, Pham PN, Salah AA. “Do no harm” in the age of big data: Data, ethics and the refugees. In: Salah AA, Pentland A, Lepri B, Letouze E, editors. *Guide to mobile data analytics in refugee scenarios*. Cham: Springer; 2019. p. 87–99. https://doi.org/10.1007/978-3-030-12554-7_5
47. Yusuf M. China’s research into Africa’s digital sector worries experts [webpage on the Internet]. c2021 [cited 2023 May 17]. Available from: <https://www.voanews.com/a/china-reach-into-africa-digital-sector-worries-experts/6281543.html>
48. Parkinson J, Bariyo N, Chin J. Huawei technicians helped African governments spy on political opponents. *Wall Street Journal*. 2019 August 15. Available from: <https://www.wsj.com/articles/huawei-technicians-helped-african-governments-spy-on-political-opponents-11565793017>
49. Woodhams S. China, Africa and the private surveillance industry. *Georget J Int Aff*. 2020;21:158–165. <https://doi.org/10.1353/gia.2020.0002>
50. Daigle B. Data protection laws in Africa: A pan-African survey and noted trends. *J Int Commer Econ*. 2021;Feb. Available from: https://www.usitc.gov/publications/332/journals/jice_africa_data_protection_laws.pdf
51. Oduro-Marfo S, Falconer TA. Digital identity in Ghana. Case study conducted as part of a ten-country exploration of socio-digital ID systems in parts of Africa [document on the Internet]. c2021 [cited 2023 May 26]. Available from: https://researchictafrica.net/wp/wp-content/uploads/2021/11/Ghana_31.10.21.pdf
52. Macdonald A. Ghana imposes fee for biometric SIM registration with self-service app [webpage on the Internet]. c2022 [cited 2023 May 26]. Available from: <https://www.biometricupdate.com/202208/ghana-imposes-fee-for-biometric-sim-registration-with-self-service-app>
53. Tindana P, Yakubu A, Staunton C, Matimba A, Littler K, Madden E, et al. Engaging research ethics committees to develop an ethics and governance framework for best practices in genomic research and biobanking in Africa: The H3Africa model. *BMC Med Ethics*. 2019;20(1), Art. #69. <https://doi.org/10.1186/s12910-019-0398-2>
54. Taylor L. Public actors without public values: Legitimacy, domination and the regulation of the technology sector. *Philos Technol*. 2021;34:897–922. <https://doi.org/10.1007/s13347-020-00441-4>
55. Kitchin R. The ethics of smart cities and urban science. *Phil Trans R Soc A*. 2016;374, Art. #20160115. <https://doi.org/10.1098/rsta.2016.0115>
56. Parkhurst J. *The politics of evidence: From evidence-based policy to the good governance of evidence*. Abingdon: Routledge; 2017. <https://doi.org/10.4324/9781315675008>
57. Lee M, Zhao J, Sun Q, Pan Y, Zhou W, Xiong C, et al. Human mobility trends during the early stage of the COVID-19 pandemic in the United States. *PLoS ONE*. 2020;15(11), e0241468. <https://doi.org/10.1371/journal.pone.0241468>
58. Perrotta D, Frias-Martinez E, Pastore y Piontti A, Zhang Q, Luengo-Oroz M, Paolotti D, et al. Comparing sources of mobility for modelling the epidemic spread of Zika virus in Colombia. *PLoS Negl Trop Dis*. 2022;16(7), e0010565. <https://doi.org/10.1371/journal.pntd.0010565>
59. Lane J, Andrews G, Orange E, Brezak A, Tanna G, Lebese L, et al. Strengthening health policy development and management systems in low- and middle-income countries: South Africa’s approach. *Health Policy OPEN*. 2020;1, Art. #1000010. <https://doi.org/10.1016/j.hpopen.2020.100010>
60. Buckee C, Balsari S, Schroeder A. Making data for good better. *PLoS Digit Health*. 2020;1(1), e0000010. <https://doi.org/10.1371/journal.pdig.0000010>
61. Usadolo SE, Mbinda BB, Maome IJ. ‘We just want to be heard!’ Dataveillance and location data – do South Africans care? *Afr J Inter-Multidiscip Sud*. 2022;4(1):64–75. <https://doi.org/10.51415/ajims.v4i1.981>
62. Ienca M, Ferretti A, Hurst S, Puhon M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. *PLoS ONE*. 2018;13(10), e0204937. <https://doi.org/10.1371/journal.pone.0204937>
63. Ferretti A, Ienca M, Velarde MR, Hurst S, Vayena E. The challenges of big data for research ethics committees: A qualitative Swiss study. *J Empir Res Hum Res Ethics*. 2022;17(1–2):129–143. <https://doi.org/10.1177/15562646211053538>

**AUTHORS:**

Dirk Brand¹ 
 Annelize G. Nienaber McKay^{2,3} 
 Nezerith Cengiz⁴ 

AFFILIATIONS:

¹School of Public Leadership, Stellenbosch University, Stellenbosch, South Africa
²Division of Law, Abertay University, Dundee, Scotland, United Kingdom
³Department of Public Law, University of Pretoria, Pretoria, South Africa
⁴Centre for Medical Ethics and Law, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

CORRESPONDENCE TO:

Nezerith Cengiz

EMAIL:

ncengiz@sun.ac.za

DATES:

Received: 16 Sep. 2022

Revised: 16 Mar. 2023

Accepted: 20 Mar. 2023

Published: 30 May 2023

HOW TO CITE:

Brand D, Nienaber McKay AG, Cengiz N. What constitutes adequate legal protection for the collection, use and sharing of mobility and location data in health care in South Africa? *S Afr J Sci.* 2023;119(5/6), Art. #14605. <https://doi.org/10.17159/sajs.2023/14605>


ARTICLE INCLUDES:

- Peer review
- Supplementary material

DATA AVAILABILITY:

- Open data set
- All data included
- On request from author(s)
- Not available
- Not applicable

EDITOR:

Floretta Boonzaier 

KEYWORDS:

data legislation, data sharing, mobility data, location data

FUNDING:

US National Institutes of Health (1U01MH127704-01)



What constitutes adequate legal protection for the collection, use and sharing of mobility and location data in health care in South Africa?

Mobile phone technology has been a catalyst that has added an innovative dimension in health care and created new opportunities for digital health services. These digital devices can be viewed as an extension of the person using them due to the deluge of personal information that can be collected and stored on them. Data collected on mobile phones are used extensively in health services and research. Personal, mobility and location data are constantly collected. The unique mobile phone architecture provides for an easy flow of data between various role players such as application developers and phone manufacturers. The collection, storage and sharing of personal information on mobile phones elicit various legal questions relating to the protection of privacy, consent, liability and the accountability of stakeholders such as health insurance providers, hospital groups and national departments of health.

Significance:

We analyse the major legal concerns of mobility and location data collection and processing through mobile phones in the context of health care and provide recommendations to develop data protection guidelines that are built on the principles of lawfulness, fairness and transparency. The issues explored are of relevance in an African context and to a broader international audience.

Introduction

Mobile phones have become an integral part of daily life and can be viewed as an extension of their owners given the extent of personal information collected and stored.¹ Although initially intended for communication, mobile phones have transcended their original use and purpose to perform more versatile functions such as electronic payments, Global Positioning System (GPS) navigation, entertainment and social media applications (apps), and health monitoring.² These extended functions escalate concerns about privacy and data protection as the information collected often is used by or sold to third parties.³

Data protection legislation largely is designed to safeguard against the exploitation of personal information through governing data collection, processing, and sharing. This protection includes data collected and processed through mobile phone use.⁴

Often data are generated and processed as an essential part of providing healthcare.² The increased use and advancement of technology allow for data that would traditionally have been collected directly from patients to now be collected through mobile phones.² Examples include cases of urgent medical care where real-time location is shared with healthcare professionals (HCPs) through smartphones or smartwatches and cases of remote health monitoring via digital applications that transmit data to HCPs to better bridge the barrier of access.²

Yet the way in which data protection legislation translates into practice, raises concerns. Are data subjects aware and adequately informed about the digital collection and processing of their personal information? How should privacy rights be managed to better protect them and legally allow for such data to be used in healthcare services?

In this article, we aim to offer guidance on the protection of privacy in the use of mobile phone data in healthcare services by addressing the above and other related questions. We include a comparative perspective about recent developments in this area in the United Kingdom (UK).

Data collection via mobile phones

The replacement of conventional paper-based methods with digital devices has significantly improved the efficiency of data collection, storage, and sharing.⁵ The rapid pace and phenomenal scope of technological development provided by smartphones have facilitated the advanced ability to relay information on speed and direction of movement, together with visual and auditory media. This ability is fostered through the various built-in sensors and multimedia functions such as a gyroscope, digital compass, and accelerometer.⁵

Cloud service providers, developers, manufacturers and proprietors of apps, operating systems, and devices are industriously involved in the complex mobile phone landscape that includes various software layers and they serve as role players in the mobility and location data ecosystem.⁵ These role players, also referred to as responsible parties in terms of legislation, are accountable for the lawful processing of personal information that complies with the applicable data protection legislation.⁶

Section 1 of South Africa's *Protection of Personal Information Act 4 of 2013* (POPIA)⁷ includes a broad definition of personal information which encompasses any information that can be used to identify a natural person. In the context of mobile phone users, their personal information includes location data, contact numbers, unique device and customer identifiers, credit card and payment data, telephone call logs, Internet browsing history, emails, pictures and videos, and biometric data.^{8,9} According to the European Union Agency for Cybersecurity, personal data further includes information related to the device itself, such as metadata, device identifiers and location data.⁸ Figure 1, developed by the World Intellectual Property Organization, illustrates various types of personal data that potentially could be collected by mobile devices.⁶

Although users actively collect and store such data on their mobile phones, data collection also occurs in large volumes in the background unbeknownst to the user; for example, activated device location services allow for the detection of geographical location.⁶ This capability raises questions about whether such personal information can be protected.⁶

Hence, responsible parties must ensure that users are aware of and unequivocally consent to the processing of their personal information.⁹ Consent equates to the ‘voluntary, specific, and informed expression of will’, which is a critical requirement for the lawful processing of data as indicated in section 1 of POPIA.⁷ Responsible parties must accede to appropriate data-sharing agreements.

App developers have access to the personal and non-personal data of their users and often are responsible for granting access to or selling their users’ data to third parties – data which can be used in behavioural advertising by retailer and marketing agencies.¹⁰ A mobile phone’s operating system is linked to various apps that provide a comprehensive set of functions to the user. Operating systems and device manufacturers have access to personal information needed to ensure smooth device and system functionality.¹⁰ Also, they are responsible for the application programming interface (API), which is software that enables the processing of personal information by apps on mobile devices⁹, which increases the risk of a data breach or unauthorised third-party use of personal data¹⁰.

The key responsibility of operating systems and mobile device manufacturers is to ensure the protection of the personal information of their users.¹¹ This responsibility necessitates legally that they inform users about the processing of personal information on devices and apps and provide the users with the opportunity to opt out of any conditions or agreements relating to such processing of information.¹¹ However, the manner in which the various role players or responsible parties present their privacy policies and request consent for the use and processing of personal data from users may be problematic. Problems arise often because privacy policies are lengthy and composed in technical terms, making them incomprehensible to average users of mobile phones.¹¹ Complexity in the presentation of language is a violation of section 22 of the *Consumer Protection Act 68 of 2008*.¹²

Although transparency is an underlying principle of lawful data processing⁶, it is beyond the control of the individual. Often mobile phone users ignorantly or uncritically grant apps access and permission to collect and process their data where their sole purpose is to utilise the functionalities of the app in question.^{13,14}

The context in which personal data are collected and the nature of the data collected are important in determining and assessing the potential risks, as sensitive information could be inappropriately integrated or contained.¹³ This possibility is because different types of data often are combined, cross referenced and used for different purposes by different role players.

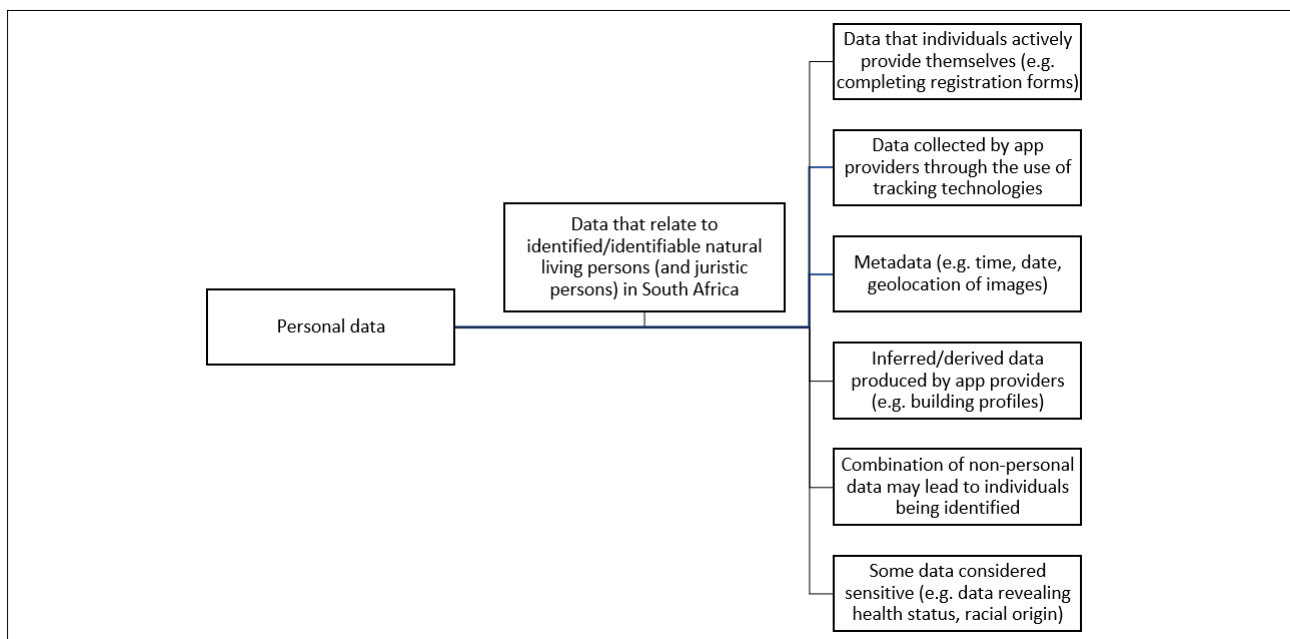
Moreover, artificial intelligence (AI) is an important component at such interfaces due to its ability to use algorithms for data analysis to further link data from different apps.^{14,15} An example could be a fitness app that collects data on a user’s physical activity and connects other data from the user’s food diary app to provide an overall model of the user’s health. Thus, the integration of AI creates another layer of personal data use and risk.¹⁴ Analysing the use and impact of AI in health services is beyond the scope of this article, but it is important to reflect briefly on this issue as mobile phone data are fed into algorithms used to develop AI-driven products and services used in health contexts.¹⁵

Use of mobile phone data and AI in health care

Health data are regarded as more sensitive than other forms of personal data, which place them higher in the level of interest for cyber criminals. Thus, this type of data receives special attention in data protection legislation such as the European Union (EU)’s General Data Protection Regulation (GDPR) and POPIA.^{7,16} Health information qualifies as “special personal information” in terms of section 26(1)(a) of POPIA, and therefore it qualifies for special protection. However, if the personal data are anonymised and cannot be re-identified, they fall outside the scope of POPIA and/or the GDPR.

In South Africa, the *National Health Act 61 of 2003* stipulates that all patient (user) information is confidential and HCPs may share or disclose that information only upon consent obtained from the patient.¹⁷ This requirement serves as a level of protection over patients’ personal information although the *National Health Act* is not focused on data protection as such.

As personal information collected through health- or fitness-related apps can be used by HCPs to provide healthcare services to individuals, so can digitally collected health data and even medical insurance data be used in medical research.^{14,15} According to Ventola¹⁸, five key categories exist for medical apps, namely “administration, health-record maintenance and access, communications and consulting, reference and information gathering, and medical education”.



Source: Modified from WIPO⁶ under a CC BY 4.0 licence

Figure 1: Types of data that could be considered personal data.

In the provision of healthcare services, mobile phone data can provide up-to-date information about an individual's state of health, which allows for remote health monitoring to better foster a HCPs clinical assessments and decision-making regarding a patient's treatment.¹⁵ In remote settings or during an emergency, the use of mobile phone apps may facilitate the provision of healthcare services through obtaining immediate access to data to remotely monitor the patient's health.¹⁸ By optimising the use of smartphones and health-related apps, the efficiency and value of healthcare provision may be improved through maximising time and resources. A variety of medical and health apps are available and are used in South Africa (and the UK), some of which are primarily for patients and others are aimed at HCPs. All these apps use personal data which often are combined with other data, as well as provide the services for which the app is designed. Examples of the most popular health and fitness and medical apps are provided in Table 1.¹⁹⁻²¹

Table 1: Most popular medical and/or health and fitness apps by sub-Saharan African country¹⁹⁻²¹

App	First ranking status (in country) on Apple App Store and/or Google Play Store	Overall downloads	Star rating
Amma: Pregnancy & Baby Tracker	Cabo Verde Guinea-Bissau Mozambique	10 M+	4.7
BetterMe: Health Coaching	South Africa	10 M+	4.1
Blood Pressure: Heart Health	Ghana Kenya Nigeria Tanzania	10 M+	4.4
Faso Santé	Burkina Faso	50 K+	4.0
Flo Period Tracker & Calendar*	Namibia Niger Mauritius Mozambique Uganda Zimbabwe	100 M+	4.6
Glow Baby Tracker & Growth App	Uganda	1 M+	4.5
HiMommy - daily pregnancy app	Nigeria	500 K+	4.7
Medscape	Zimbabwe	5 M+	4.6
Menstrual Cycle Tracker by Anastasai Kovba	Ghana	500 K+	4.7
Motivation - Daily quotes	Ghana	1 M+	4.8
Pregnancy + Tracker App*	Niger	10 M+	4.7
Pulse - Heart Rate Monitor app	Namibia	5 M+	4.5
SICOM Health	Mauritius	500 K+	–
Smart Access!	Kenya	50 K+	
Useful healthcare apps for patients			
App	Function	Overall downloads	Star rating
Better Help	Online therapy	1 M+	3.9
MDacne	Custom acne treatment	500 K+	4.5
MySugr	Diabetes tracker log	1 M+	4.4
Teladoc Health	Telehealth and telemedicine provider (virtual care)	1 M+	4.1

*Ranked first in the medical or health and fitness apps categories.

If the personal data on a fitness or health app are sent to medical insurers or HCPs, the recipients are allowed to process that health data in terms of the exception under section 32(1) of POPIA.⁷ In the EU, the GDPR allows for such health data to be lawfully processed by HCPs and to be used in medical diagnosis and healthcare provision or treatment (Art 9(2)(h) GDPR).^{16,22}

Similarly, in low- and middle-income countries where patients experience challenges in accessing health care, the use of mobile phone data enables HCPs instant access to patients' up-to-date information.²³ On the other hand, in high-income settings where advanced healthcare services are available, data collection through portable technological devices is essential. Smart hospitals, which are characterised by high-tech infrastructure and high-speed communication networks that "create new value and insights on patient safety, quality of care, cost-effectiveness, and patient-centeredness", are further fostered by AI and mobile phone data.²⁴

AI systems, consisting of one or more algorithms, can be used to complement the decision-making of HCPs in the diagnosis and treatment of patients.²⁵ Health apps on mobile phones often operate with AI and can be utilised as a source of personal information in assessing the health of a patient. However, the training, testing and use of AI models in health care require large amounts of health data, which raises questions around the privacy and protection of patients' personal data and, again, whether informed consent was obtained.²⁵ Mittelstadt²⁵ argues that these questions should be addressed on a case-by-case basis to reflect the extent to which the AI model is used to provide health care.

In addition to pertinent questions on how personal data are protected in the development and use of AI models, other important questions around the interpretability, transparency and traceability should not be ignored.^{15,25} Such questions include how AI models produce their specific output, how they are governed and what other data are required for auditing purposes? The use of AI models in the diagnosis and the treatment of patients brings into question if informed consent was obtained, or could be obtained, and, thus, impacts the doctor-patient relationship.^{25,26}

Protection of personal data concerns

When consent is requested for the processing of personal information in an app, care should be taken to ensure clarity about the purpose and scope of such processing. It is common that apps are interlinked, e.g. a fitness app that provides the possibility of sharing data on various social media apps, which increases the risk of a data breach or the unauthorised use of the personal data. In the sharing of personal data between apps, how can privacy and protection still be ensured to prevent the risk of misuse or theft by unauthorised third parties?

Mulder²² argues that vague language is used frequently by app providers in their statements and requests to collect and share data and, thereby, transgresses the fundamentals of informed consent and hinders the ability of individuals to provide true informed consent. This matter is cause for concern and has led to various court cases in the European Union relating to contraventions of the GDPR.¹⁶ In 2021, the Irish Data Protection Commission found Meta guilty of non-adherence to the GDPR's transparency requirement to inform the users of WhatsApp of how their personal data are treated.¹⁶ Consequently, a fine of EUR225 million was issued.²⁷

Added to the complex challenge of obtaining consent for mobile phone app use in South Africa is the low literacy levels in certain populations in the country. A study by the Department of Higher Education and Training indicates that 3.7 million adults in South Africa are illiterate.²⁸ Consequently, a significant portion of the population might struggle to understand the terms and conditions of app use, let alone the implications of sharing personal health information with third parties. To address this challenge, app developers must take a user-centred approach in designing and developing apps that are easy to use and understand. Achieving this goal involves using simple language, visual aids and audio cues to convey important information to users. Also, app developers should prioritise user testing and feedback to ensure that their apps are accessible for and understandable by people with low literacy levels.

Other data protection risks in mobile phone use include the constant power-up and Internet connection which facilitate data access by unauthorised third parties. Smartphones have various sensors that collect a variety of personal data and identifiers such as the device ID, metadata, and geolocation which, together, increase the risk of tracking and user profiling without consent.^{5,8} Such collated data from different trackers installed on apps feed behavioural advertising, with users often having only limited or no control.^{5,8,29}

Processing of children's personal information receives special attention in data protection legislation such as the GDPR and POPIA, because children are regarded as a vulnerable group in society and they may be less aware of the risks involved (Recital 38, GDPR).¹⁶ Their personal data, for example, can be used to manipulate and influence their behaviour. A responsible party must thus take extra care when processing the personal information of children. Prior consent by a competent person, such as a parent or legal guardian, is a requirement for the lawful processing of children's personal information (sections 34 and 35 of POPIA).⁷ These requirements apply to responsible parties in the mobile phone environment. When a mobile phone is used or an app is accessed, personal information is collected and processed, which has application to children as well. If consent is requested, it is doubtful that a competent person will always be there to provide it. If proper consent is not provided, the child's personal information is processed unlawfully, unless another legal ground applies. Children have the same rights as adults regarding the protection of their personal information, including when they use a mobile phone.

Users of mobile phones often do not have a clear understanding of the permission required to use an app, and some apps may require more permission than is needed to function properly. This circumstance raises concerns about the legal compliance of the app providers. It is the responsibility of operating systems and app providers on mobile phones to ensure the lawful processing of personal information and, in accordance with the applicable data protection legislation, they should take extra care when the personal information of children is processed.

According to the World Intellectual Property Organization⁶, the following key principles, often found in data protection legislation, should apply to all processing of personal data in the mobile phone context: lawfulness, fairness and transparency. Application of these principles implies that:

- there must always be a legal basis for processing personal data on a mobile phone, which could be consent provided by a data subject or another legal basis specified in the relevant legislation⁶;
- processing may not lead to unfair discrimination and should avoid importing any bias⁶; and
- appropriate information about the processing must be provided in an understandable and clear way, and this could include publishing an appropriate privacy policy before installation of the app or before processing the data, and the provision of icons or privacy notifications during use of the app^{5,6}.

Currently, there is no set of guidelines on mobile phone applications in South Africa. However, given the similitude between POPIA and the GDPR, the 'Guidelines on the Protection of Personal Data Processed by Mobile Applications Provided by European Union Institutions' may serve as guidance in our jurisdiction.^{7,16} These guidelines state that apps should collect only data that are strictly necessary for its functioning and that users must be provided with clear and accurate information to make an informed decision, with the option to withdraw their consent at any time.³⁰ In Europe, the Oviedo Convention is a further legal instrument in the health context that is aimed at the protection of human rights, including the right to privacy.³¹ Article 5 of this Convention confirms the requirement of informed consent in the provision of health care to a patient.³¹

A comparative perspective from the UK

The UK's data protection framework predates that of South Africa, making it instructive to look at how the UK handles issues related to data privacy and the use of mobile devices to discover learning opportunities from the UK experience.

The oldest instrument in the UK's data protection framework is an international data protection treaty to which the UK is a party, namely, the Convention for the Protection of Individuals regarding Automatic Processing of Personal Data³² (CETS 1981). The Automatic Data Processing Convention entered into force in October 1985 and to date has 55 ratifications or accessions.³² The Convention is aimed at ensuring respect for individual rights and fundamental freedoms and the right to privacy regarding automatic processing of personal data (Preamble and Article 1).³²

The Automatic Data Processing Convention provides the data subject with rights of access to, and correction of data held by third parties (Article 8).³² Principles such as accuracy of data, the minimisation of data, fairness, lawfulness, and transparency in data processing are all included in the Convention (Articles 4–8).³² The Convention distinguishes between personal and more sensitive personal data and prohibits sensitive personal data from being processed unless appropriate safeguards are in place (Articles 5–8).³²

In 1998, the UK enacted the *Data Protection Act* (DPA 1998).³³ It enacted the provisions of the EU's Data Protection Directive (Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995) and was aimed at the protection, processing, and movement of personal data.³⁴

In 2016, the EU enacted the GDPR 61 of 2016 (Regulation 2016 679 EU).¹⁶ The GDPR replaced the Data Protection Directive mentioned above.³⁴ The GDPR is aimed at harmonising data processing practices and the level of data protection provided to data subjects in EU member states (Preamble, GDPR).¹⁶ The GDPR also applies to bodies and entities outside the EU that process data of data subjects who are in the EU (Article 3, GDPR).¹⁶ As the GDPR is an EU Regulation, it applies in all EU member states without the need for any further implementing or enabling legislation to be passed in those member states (an EU regulation is law once passed and published in the official journal).¹⁶ As the UK was a member of the EU at that time, the GDPR applied in the UK.

The GDPR's stated aim is to harmonise data privacy laws across Europe (Article 1, GDPR).¹⁶ The GDPR sets out the conditions for the lawful processing of data in Article 6 and lists the conditions for the lawful consent of the data subject to the processing of personal data in Article 7.¹⁶ Article 8 makes provision for special conditions in the processing of children's data, and Article 9 provides special conditions for the processing of special categories of data.¹⁶ Article 9(1) prohibits the processing of information related to personal data that reveals the data subject's "racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".¹⁶ These conditions have important implications for the processing of health data on mobile phones.

Article 9(2) provides for circumstances under which the prohibition on the processing of data mentioned in sub-article 9(1) does not apply.¹⁶ These exclusions, inter alia, include instances where "the data subject has given explicit consent to the processing of those personal data for one or more specified purposes"; if the processing is "necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis", or "the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional" (Article 9(2)).¹⁶

In addition, according to sub-article 9(4), member states may "maintain or introduce further conditions, including limitations", in respect of the "processing of genetic data, biometric data or data concerning health" (Article 9(4)).¹⁶

However, after 31 December 2020, at the end of the Brexit transition period, the GDPR ceased to apply directly in the UK but was incorporated into UK domestic law under section 3 of the *European Union (Withdrawal) Act 2018* as well as the *Data Protection Act 2018* (DPA 2018), successor to the DPA 1998.³³ The UK now is considered a "third country" in terms of the GDPR; nevertheless, as mentioned above, the UK's DPA 2018 enacted

the GDPR's requirements into UK law, and closely corresponds to the GDPR.¹⁶ In addition, as from 1 January 2021, the Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019 (DPPEC Regulations) that amended the DPA 2018³⁵ came into effect.³⁶ The DPPEC Regulations amend both the GDPR and the DPA 2018 and turn it into the UK's new data protection framework (UK-GDPR).¹⁶

The UK-GDPR broadly is the same as the GDPR in terms of its substantive requirements; however, as the UK no longer is a member of the EU, it provides for an alternative enforcement mechanism.¹⁶ An Information Commissioner's Office is set up as the new UK-specific supervisory body by the DPA 2018.³⁵ This is an independent body which reports directly to Parliament. The jurisdiction, functions, and powers of the Information Commissioner's Office are set out in the DPA 2018.³⁵

Data privacy in the context of mobile phones in the UK is regulated further by the Privacy and Electronic Communications (EC Directive) Regulations 2003 (PECR)³⁷ which implement the requirements of Directive 2002/58/EC (as amended by Directive 2009/136/EC)³⁷ which provides a specific set of privacy rules for the processing of personal information by the telecommunications sector.³⁴ Unlike the GDPR, the PECR remains in force in the UK despite the UK's departure from the EU. Therefore, three main instruments or pieces of legislation constitute the UK-GDPR: the DPA 2018, the PECR, and the DPPEC Regulations.^{16,35,37}

In keeping with regulations in the EU and other parts of the world, the UK-GDPR contains provisions to ensure the protection of personal data. These include the requirement that personal data be "processed lawfully and fairly"; that such processing should be based on the data subject's consent or, if consent is absent, that it be based on another specified legal basis; it grants the data subject the right to obtain information about the processing of personal data and to demand that inaccurate personal data be rectified; it confers appropriate functions on the Information Commissioner's Office (see above), endowing that Office with the responsibility to monitor and enforce the provisions of the UK-GDPR.¹⁶

Importantly, the DPA 2018 adopts the definitions of the (EU's) GDPR, such as "personal data" meaning "any information relating to an identified or identifiable living individual"; "processing" meaning "an operation or set of operations which is performed on information, such as collection, recording, storage, disclosure, combination etc"; "data subject" as a "living individual to whom personal data relates", and so on.¹⁶

On 28 June 2021 the EU adopted an adequacy decision for the UK.³⁸ This means that entities in the UK that process personal data from data subjects in the EU can do so in the same way as they did previously until June 2025.³⁸

On the face of it, the UK-GDPR framework constitutes a solid mechanism that protects individual privacy, including in relation to personal data being processed on mobile phones.¹⁶ However, research by Kollnig et al.³⁹ suggests that "there has been limited change in the behaviour of cell phone apps regarding third-party tracking and the collection and sharing of behavioural data about individuals". They state that this circumstance is a significant and ubiquitous privacy threat in mobile apps and that there exists limited empirical evidence about the efficacy of the existing EU and UK privacy protection frameworks. Specifically, Kollnig et al.³⁹ found that "there has been limited change in the presence of third-party tracking in apps, and that the concentration of tracking capabilities among a few large gatekeeper companies persists". The authors found that the GDPR has had little effect on third-party tracking across apps on the UK Google Play Store (and hence, neither has the UK-GDPR)^{16,39}

A 2021 literature review by Steven Furnell, commissioned by the UK government, revealed that although, on the face of it, the UK has a watertight data privacy framework, the reality is not as clearcut as it seems.⁴⁰ Furnell found that mobile phone app stores have "varying approaches with correspondingly variable levels of information and clarity"⁴⁰. This variability is observed in terms of both the presence and content of their privacy and other policies, as well as in relation to supporting users' understanding of these policies when downloading specific apps. This is particularly apparent when observing the presence and clarity of messaging about app permissions and in the handling of personal data. Some stores provide

details that are comprehensive whereas others provide "nothing that most users would find meaningful"⁴⁰.

In the light of an Australian study which found that there are significant shortcomings in relation to privacy, and inconsistent privacy practices in health-related mobile phone apps⁴¹, one is left wondering whether the same can be said for the UK.

Conclusions and recommendations

In exploring the use of mobile phones in health care, this article provides an overview of the complex mobile phone landscape and identifies various legal concerns relating to the processing of personal information on mobile phones. Despite the existence of data protection legislation in most countries, the shortcomings in relation to the protection of personal information in health-related mobile phone apps identified in Australia probably are relevant everywhere.

The increased availability and use of health and fitness apps on mobile phones provide various benefits to users and HCPs. However, the risk of unlawful data processing on mobile phones still exists despite the presence of general data protection legislation. The protection of privacy on mobile phones is a challenge given a complex landscape with various role players. The most common legal basis for the processing of personal data remains the consent of the data subject. Yet operating systems and app developers often use longwinded and opaque language upon seeking consent or providing information about the purpose of data processing. This practice is of particular concern in South Africa given the low literacy level in certain population groups.

A multi-disciplinary approach – in combination with the development of clear guidance for HCPs, healthcare institutions, patients, and the manufacturers of digital devices – will address the various ethical and legal issues in digital health care. Furthermore, it is recommended that guidelines for the protection of personal data on mobile phone apps are developed based on the principles of lawfulness, fairness, and transparency. A reliance on these principles is important, not only in South Africa but everywhere. The development of legislation for the use of AI in healthcare services is recommended to further strengthen the protection of privacy and personal data in healthcare services in South Africa.

The collection, use and sharing of mobility and location data in health care in South Africa presents a scenario with significant benefits and risks. Adequate legal protection is essential to ensure that the data are collected, used and shared in a responsible and ethical manner that respects individual rights and privacy. A comprehensive legal framework that includes data protection regulations, ethical guidelines and oversight mechanisms is a necessary requirement to address the complex issues surrounding mobility and location data in health care. Such a framework should account for the unique cultural and societal contexts in South Africa. It is an imperative that policymakers, healthcare providers, and other stakeholders work together to develop and to implement an effective legal framework that protects the rights of individuals while promoting the responsible use of mobility and location data to improve healthcare outcomes. Only in doing so, can South Africa fully leverage the potential in these technologies to improve the delivery of health care and ensure that individual privacy and rights are safeguarded.

Acknowledgements

Research reported in this publication was supported by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Competing interests

We have no competing interests to declare.

Authors' contributions

D.B.: Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work; drafted/



revised the work critically for important intellectual content. A.G.N.M.: Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work; drafted/revised the work critically for important intellectual content. N.C.: Drafted/revised the work critically for important intellectual content. All authors approved the final version and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.





References

- Jiang Y, Mosquera L, Jiang B, Kong L, El Emam K. Measuring re-identification risk using a synthetic estimator to enable data sharing. *PLoS ONE*. 2022;17, e0269097. <https://doi.org/10.1371/journal.pone.0269097>
- Jiang D, Shi G. Research on data security and privacy protection of wearable equipment in healthcare. *J Healthc Eng*. 2021;2021, Art. # 6656204. <https://doi.org/10.1155/2021/6656204>
- Alnajrani HM, Norman AA, Ahmed BH. Privacy and data protection in mobile cloud computing: A systematic mapping study. *PLoS ONE*. 2020;15, e0234312. <https://doi.org/10.1371/journal.pone.0234312>
- Carrillo MA, Kroeger A, Cardenas Sanchez R, Diaz Monsalve S, Runge-Ranzinger S. The use of mobile phones for the prevention and control of arboviral diseases: A scoping review. *BMC Public Health*. 2021;21(1):110. <https://doi.org/10.1186/s12889-020-10126-4>
- Working Party. Working document 02/2013 providing guidance on obtaining consent for cookies. European Union; 2013.
- World Intellectual Property Organization (WIPO). A guide to data protection in mobile applications Geneva: WIPO; 2021. Available from: <https://www.wipo.int/export/sites/www/ip-development/en/agenda/docs/wipo-guide-data-protection-mobile-apps.pdf>
- Republic of South Africa. Protection of Personal Information Act 4 of 2013. Republic of South Africa; 2013.
- European Union Agency for Cybersecurity (ENISA). Privacy and data protection in mobile applications: A study on the app development ecosystem and the technical implementation of GDPR [document on the Internet]. c2017 [cited 2022 Sep 16]. Available from: https://www.enisa.europa.eu/publications/privacy-and-data-protection-in-mobile-applications/at_download/fullReport
- Kamarinou D, Millard C, Turton F. Responsibilities of controllers and processors of personal data in clouds. In: Millard C, editor. *Cloud computing law*. 2nd ed. Oxford: Oxford University Press; 2021. p. 294–339. <https://doi.org/10.1093/oso/9780198716662.003.0009>
- Johnson G, Runge J, Seufert E. Privacy-centric digital advertising: Implications for research. *Customer Needs and Solutions* 2022;9:49–54. <https://doi.org/10.1007/s40547-022-00125-4>
- Fowler GA. I tried to read all my app privacy policies. It was 1 million words. *Washington Post*. 2022 May 31. Available from: <https://www.washingtonpost.com/technology/2022/05/31/abolish-privacy-policies/>
- Republic of South Africa. Consumer Protection Act 68 of 2008. *Government Gazette*. Volume 526 Number 32186. South African Government; 2009 [cited 2022 Sep 16]. Available from: https://www.gov.za/sites/default/files/32186_467.pdf
- Rath DK, Kumar A. Information privacy concern at individual, group, organization and societal level – a literature review. *Vilakshan – XIMB Journal of Management*. 2021;18:171–186. <https://doi.org/10.1108/XJM-08-2020-0096>
- Seifert A, Hofer M, Allemann M. Mobile data collection: Smart, but not (yet) smart enough. *Front Neurosci*. 2018; 12, Art. #971. <https://doi.org/10.3389/fnins.2018.00971>
- Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, editors. *Artificial intelligence in healthcare*. Cambridge, MA: Academic Press; 2020. p. 295–336. <https://doi.org/10.1016/B978-0-12-818438-7.00012-5>
- The European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons. The General Data Protection Regulation (GDPR). European Union; 2016.
- South African Government. National Health Act 61 of 2003. Republic of South Africa; 2004.
- Ventola CL. Mobile devices and apps for health care professionals: Uses and benefits. *P T*. 2014;39:356–364.
- SimilarWeb. Top Apps Ranking [webpage on the Internet]. No date [updated 2023 Mar 12; cited 2023 Mar 14]. Available from: <https://www.similarweb.com/apps/top/apple/store-rank/za/health-fitness/top-free/iphone/>
- Apple App Store. App Store Preview: Health & fitness [webpage on the Internet]. No date [cited 2023 Mar 14]. Available from: <https://apps.apple.com/us/charts/iphone/health-fitness-apps/6013>
- Google Play. Top charts [webpage on the Internet]. No date [cited 2023 Mar 14]. Available from: <https://play.google.com/store/apps>
- Mulder T. Health apps, their privacy policies and the GDPR. *Eur J Law Technol*. 2019;10(1):1–21.
- Feroz A, Jabeen R, Saleem S. Using mobile phones to improve community health workers performance in low-and-middle-income countries. *BMC Public Health*. 2020;20, Art. #49. <https://doi.org/10.1186/s12889-020-8173-3>
- Kwon H, An S, Lee H-Y, Cha WC, Kim S, Cho M, et al. Review of smart hospital services in real healthcare environments. *Healthc Inform Res*. 2022;28:3–15. <https://doi.org/10.4258/hir.2022.28.1.3>
- Mittelstadt B. The impact of artificial intelligence on the doctor-patient relationship. Strasbourg: Council of Europe; 2021. <https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859>
- Silven AV, Van Peet PG, Boers SN, Tabak M, De Groot A, Hendriks D, et al. Clarifying responsibility: Professional digital health in the doctor-patient relationship, recommendations for physicians based on a multi-stakeholder dialogue in the Netherlands. *BMC Health Serv Res*. 2022;22, Art. #129. <https://doi.org/10.1186/s12913-021-07316-0>
- The Data Protection Commission. Data Protection Commission announces decision in WhatsApp inquiry [media release]. 2021 September 02 [cited 2022 Sep 16]. Available from: <https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-announces-decision-whatsapp-inquiry>
- Khuluvhe M. Adult illiteracy in South Africa. Pretoria: Department of Higher Education and Training; 2022. Available from: https://www.dhet.gov.za/Planning%20Monitoring%20and%20Evaluation%20Coordination/Fact%20Sheet%20-%20Adult%20illiteracy%20in%20South%20Africa_March%202022.pdf
- Melicher W, Kurilova D, Segreti SM, Kalvani P, Shay R, Ur B, et al. Usability and security of text passwords on mobile devices. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*; 2016 May 7–12; San Jose, CA, USA. New York: Association for Computing Machinery; 2016. p. 527–539. <https://doi.org/10.1145/2858036.2858384>
- European Data Protection Supervisor. Guidelines on the protection of personal data processed by mobile applications provided by European Union institutions [document on the Internet]. c2016 [cited 2022 Sep 16]. Available from: https://edps.europa.eu/sites/default/files/publication/16-11-07_guidelines_mobile_apps_en.pdf
- Council of Europe. Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine (ETS No. 164). Strasbourg: Council of Europe; 1997. Available from: www.coe.int/en/web/bioethics/oviedo-convention
- Council of Europe. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Strasbourg: Council of Europe; 1981. Available from: <https://rm.coe.int/1680078b37>
- European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons. Data Protection Act 1998. United Kingdom, 1998 [cited 2022 Sep 16]. Available from: <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>
- European Parliament and of the Council of 24 October 1995. Data Protection Directive 95/46/EC. European Union; 1995 [cited 2022 Sep 16]. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046>
- European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons. The Data Protection Act 2018. United Kingdom; 2018 [cited 2022 Sep 16]. Available from: https://www.legislation.gov.uk/ukpga/2018/12/pdfs/ukpga_20180012_en.pdf



36. European Parliament and of the Council of the Council of 27 April 2016 on the Protection of Natural Persons. The Data Protection, Privacy and Electronic Communications (Amendment etc) (EU Exit) Regulations 2019. United Kingdom; 2019 [cited 2022 Sep 16]. Available from: <https://www.legislation.gov.uk/ukdsi/2019/9780111177594/contents>
 37. European Parliament and of the Council of 11 December 2003. The Privacy and Electronic Communications (EC Directive) Regulations 2003. United Kingdom; 2003 No. 2426 [cited 2022 Sep 16]. Available from: <https://www.legislation.gov.uk/ukdsi/2003/2426/contents/made>
 38. Information Commissioner's Office. Data protection and the EU in detail. Data protection at the end of the transition period [document on the Internet]. c2019 [cited 2022 Sep 16]. Available from: <https://ico.org.uk/media/for-organisations/dp-at-the-end-of-the-transition-period/data-protection-and-the-eu-in-detail-1-0.pdf>
 39. Kollnig K, Binns R, Van Kleek M, Lyngs U, Zhao J, Tinsman C, et al. Before and after GDPR: Tracking in mobile apps. *Internet Policy Rev.* 2021;10. <https://doi.org/10.14763/2021.4.1611>
 40. Furnell S. Closed consultation: Literature review on security and privacy policies in apps and app stores [webpage on the Internet]. c2022 [cited 2022 Sep 16]. Available from: <https://www.gov.uk/government/consultations/app-security-and-privacy-interventions/literature-review-on-security-and-privacy-policies-in-apps-and-app-stores>
 41. Tangari G, Ikram M, Ijaz K, Kaafar MA, Berkovsky S. Mobile health and privacy: Cross sectional study. *BMJ.* 2021;373, Art. #1248. <https://doi.org/10.1136/bmj.n1248>
-

**AUTHORS:**

Nezerith Cengiz¹ 
 Siti M. Kabanda¹ 
 Tonya M. Esterhuizen² 
 Keymanthri Moodley¹ 

AFFILIATIONS:

¹Centre for Medical Ethics and Law, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

²Division of Epidemiology and Biostatistics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

CORRESPONDENCE TO:

Nezerith Cengiz

EMAIL:

ncengiz@sun.ac.za

DATES:

Received: 30 Sep. 2022

Revised: 01 Feb. 2023

Accepted: 02 Mar. 2023

Published: 30 May 2023

HOW TO CITE:

Cengiz N, Kabanda SM, Esterhuizen TM, Moodley K. Exploring perspectives of research ethics committee members on the governance of big data in sub-Saharan Africa. *S Afr J Sci.* 2023;119(5/6), Art. #14905. <https://doi.org/10.17159/sajs.2023/14905>

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:

big data, data governance, data regulation, research ethics committees, sub-Saharan Africa

FUNDING:

US National Institutes of Health (1U01MH127704-01)



Exploring perspectives of research ethics committee members on the governance of big data in sub-Saharan Africa

Interest in the governance of big data is growing exponentially. However, finding the right balance between making large volumes of data accessible, and safeguarding privacy, preventing data misuse, determining authorship and protecting intellectual property remain challenging. In sub-Saharan Africa (SSA), research ethics committees (RECs) play an important role in reviewing data-intense research protocols. However, this regulatory role must be embedded in a context of robust governance. There is currently a paucity of published literature on how big data are regulated in SSA and if the capacity to review protocols is sufficient. The aim of this study was to provide a broad overview of REC members' awareness and perceptions of big data governance in SSA. A descriptive cross-sectional survey was conducted from April to July 2022. We invited 300 REC members to participate in our online survey via Research Electronic Data Capture (REDCap). A total of 140 REC members, representing 34 SSA countries, completed the online survey. Awareness of data governance laws, policies and guidelines was variable across the subcontinent. A quarter of respondents (25%) indicated that national regulations on the trans-border flow of research data are inadequate. Institutional policies on research data protection were also regarded as being inadequate. Most respondents (64%) believed that they lacked experience in reviewing data-intense protocols. Data governance and regulation in SSA need to be strengthened at both national and institutional levels. There is a strong need for capacity development in the review of data-intense research protocols on the subcontinent.

Significance:

This is the first empirical survey in SSA in which awareness and perspectives of REC members have been explored specifically relating to the review of data-intense research protocols. Big data have raised new ethics and legal challenges, and this survey provides a broad overview of these challenges in SSA. Our study confirms that knowledge and awareness of legislative frameworks and ethics guidance in SSA vary considerably where big data are concerned. The research results could be useful for a range of stakeholders, including RECs, data scientists, researchers, research and academic institutions, government decision-makers and artificial intelligence (AI) coders.

Background

The abundance of health and research data that exists today has enormous potential to unlock future advances in science – a prospect discussed for decades by researchers and policymakers.¹ Recently, the potential of big data to solve some of the world's most challenging problems has become more apparent. 'Big data' refers to large volumes of a variety of raw data processed at high speed and frequency.² The sharing of research data is of increasing interest, with many funders advocating for, or even requiring researchers to share data sets as a condition of funding to maximise their utility and value.³ Understandably, sharing research data is regarded as a best practice by the World Health Organization (WHO).^{4,5}

Despite the benefits of data sharing, finding the right balance between making data accessible and safeguarding privacy, preventing data misuse, determining authorship and protecting intellectual property is challenging.^{4,6,7} This challenge has been reported to be greater in low- and middle-income countries (LMICs) such as in sub-Saharan Africa (SSA) because of the gap that exists in decision-making between data producers and data users.^{4,7} Some SSA countries have introduced data protection regulations in response to the recent digital revolution.

South Africa is one of the countries that has sought to enforce data governance via the *Protection of Personal Information Act (POPIA), Act No. 4 of 2013*, which came into force on 1 July 2020.⁸ However, legal and ethics frameworks to guide data sharing and protect the interests of data donors on the subcontinent appear to vary considerably in their structure, terms, procedures and authority.⁹

Data protection has also become concerning in the context of the cross-border transfer of human biological materials (HBMs) and data.¹⁰ In response to this, Material Transfer Agreements (MTAs) and Data Transfer Agreements (DTAs) have evolved to contractually govern the transfer of biological materials and data between parties to protect the interests of stakeholders.¹¹ A DTA is a legal contract governing the transfer of deidentified human subject data, or identifiable human subject data in cases where a respondent has given voluntary, informed and electronic consent.¹² DTAs are required when data owned by one institution are transferred to another institution for the continuation of research efforts. A DTA sets out the related protection, rights and obligations of both parties and delineates the specific purpose(s) for which the data may be used. This facilitates the cross-border transfer of data.^{11,12} In some countries, there is an additional requirement to inform the relevant national data protection authority about the cross-border transfer of data.

Research ethics committees (RECs) have traditionally been established to protect the rights of research participants. However, they also play an important role in reviewing data-intensive research protocols where data protection and data sharing are important.¹³ The recent pandemic has placed increasing demands on RECs as research engaging with big data and artificial intelligence (AI) was accelerated. Many scholars have been deliberating on the role of RECs in reviewing data-intensive research protocols, and have found that developed countries such as Switzerland², the UK¹⁴ and Australia¹⁵ lack the expertise or skills to review such studies. Big data research should be differently legislated and considered as it poses greater or unique risks and implications than flows of samples. Conventional informed consent is not ideal for protecting participants in big data research.² Other examples of the implications of big data research include anonymisation, algorithmic bias, data protection, data storage and data reuse. In many countries in SSA, biological samples are regulated in legislation via MTAs and in guidelines.¹⁶ However, data, and particularly big data, are excluded. The rapid flow of large volumes of data carries benefits to science, but also many risks to personal information protection and governance, and should be regulated.

The data ecosystem is becoming increasingly complex. Apart from RECs, Data Access Committees (DACs) have emerged as another governance mechanism to manage the controlled access of data.¹³ A DAC comprises a group of individuals who have the responsibility of reviewing and assessing research data access requests.¹³ They may serve as part of an REC or may be an independent committee in an institution or country with the aim of promoting the benefits of data access, whilst minimising potential harm to data respondents or donors.¹³

Data governance is understood as the practice of safeguarding valuable information from exploitation, compromise and loss or theft. It is largely executed through regulatory and legal data protection frameworks.¹⁷⁻¹⁹ These frameworks govern how certain data types are collected, processed and shared. This secures the privacy, availability and integrity of data through frameworks that set out how sensitive data, in particular, and privacy should be managed via the provision of tools and policies that restrict the unauthorised access, use and/or transfer of data.¹⁷⁻¹⁹ Examples of personal identifiable data include names, photographs, email addresses, bank account details, the Internet Protocol (IP) addresses of personal computers and biometric data.¹⁷

It is important to note that data protection laws may differ across various countries, thereby causing an inequality and disparity in the degree of data protection. Some of these countries have stricter rules that apply, which may require notification or approval by the data protection authority and/or special conditions, as well as consent from the data subject as a requirement for the cross-border transfer of data.²⁰

In South Africa, the National Health Research Ethics Council (NHREC) developed a national guideline, 'Ethics in Health Research: Principles, Processes and Structures', in 2015 to ensure that research is conducted responsibly and ethically in South Africa.²¹ The NHREC emphasises the importance of recognising the values, beliefs and attitudes of data donors.²¹

The guidance document recommends the responsible management of data collection, informed consent, the protection of vulnerable populations, the permissible secondary use of data, and the non-maleficent use of genetic and genomic research.²¹ However, these guidelines are not specific to big data collection, and improved recommendations are required to meet international standards of data management.^{21,22}

Being cognisant of the challenges in the big data ecosystem in LMICs, we aimed to determine REC members' perceptions of data governance in SSA and to describe related challenges. This study is part of a bigger project exploring the ethical, legal and social implications of big data and AI in SSA.

To date, there are no published studies from SSA that have explored the perspectives of REC members on data governance or on the review of data-intensive research protocols. Consequently, it is unclear how REC members on the subcontinent navigate governance structures and processes, and review such protocols. This study offers a novel contribution to the empirical literature in SSA as it aimed to explore these perspectives.

Methods

Study design and sampling

A descriptive cross-sectional survey with both quantitative and qualitative components, involving 140 REC members representing 34 SSA countries, was conducted from April to July 2022. Our aim was to recruit at least one representative from each of the 49 SSA countries. The study population was selected based on membership of a private, institutional or national REC in SSA.

Respondents were invited to participate in an online survey through a web-based application, Research Electronic Data Capture (REDCap). We recruited our sample of REC members through a purposive selection of professional networks of the Stellenbosch University's Centre for Medical Ethics and Law across SSA, and employed a snowballing technique to recruit further respondents.²³ All respondents participated in their personal capacities and provided online consent prior to their completion of the survey.

The survey instruments

The questionnaire was developed based on a review of the literature and consultation with experts in research ethics. A final draft of the questionnaire was developed using REDCap. This online questionnaire was piloted with six REC members from Stellenbosch University to assess its legibility, eliminate ambiguous questions, address repetition and identify any missing information. This was to ensure the face validity of the data collection tool.

The piloted version of the questionnaire consisted of 20 closed-ended questions, of which four were conditional questions that required respondents to meet a certain condition to be asked the following question. These questions were used to establish baseline data regarding the existence of research data-sharing frameworks and guidelines in SSA, the level of awareness of these frameworks and guidelines by REC members, and perspectives regarding existing legal and ethical challenges. In the questionnaire, we distinguished between the institutional and national governance of research data protection and the trans-border flow of research data to take into account the SSA countries without national governance laws. These were divergent across some institutions and countries.

The data collection tool was developed in English and translated into French and Portuguese to cater for Francophone and Lusophone countries.

Data analysis

Survey responses were exported from REDCap into the Statistical Package for Social Sciences (SPSS) version 28 for analysis. Frequencies and percentages were used to describe responses to the closed questions. A trained researcher analysed the answers from the open-ended questions manually by identifying recurring responses.

Ethical aspects

Research integrity was maintained throughout the study, and participation in this research remained entirely voluntary. This survey was a minimal-risk study as the questionnaires involved a factual enquiry with educated, empowered respondents who had the full capacity to consent or decline participation. We approached members in their individual capacities, and respondents consented in their personal capacities. Ethics approval was granted by the Health Research Ethics Committee of the Faculty of Medicine and Health Sciences (reference no: N22/03/028) at Stellenbosch University, South Africa.

Results

Demographic information

A total of 300 individuals were invited to participate in the research study and 140 completed the online survey, yielding an overall response rate of 47% (140/300). The total number of respondents represented 34 of the 49 SSA countries (Figure 1).

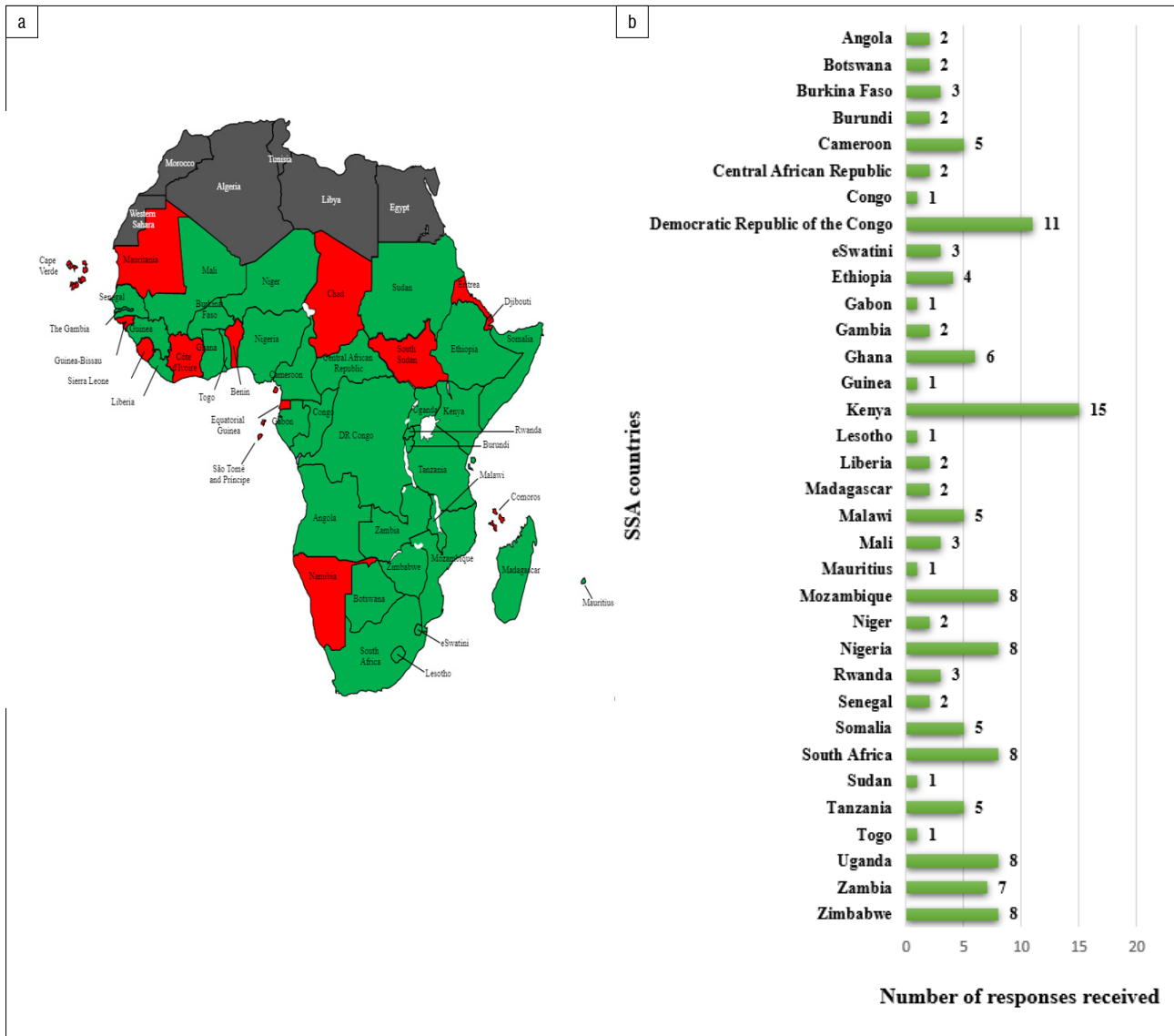


Figure 1: (a) Sub-Saharan Africa (SSA) and (b) the representation of responses received across SSA countries.

More than half the respondents (63%) self-identified as male (88/140), whilst 46% of the respondents (64/140) were PhD graduates, and 41% (58/140) were master's degree graduates (Table 1). Of the respondents, 80% (112/140) had served in the capacity of an REC member and most responses (69%) came from those who had served on an institutional REC (96/140).

Awareness of current laws and policies on research data protection

Just over half the respondents (59%; 82/140) indicated that their country had laws on research data protection (Table 2). Less than half (48%; 67/140) indicated that their country had restrictions and/or prohibitions regarding the trans-border flow of research data. We validated whether respondents responded correctly when reporting on the existence of legislation in their respective countries (Table 3). Of 107 respondents, 76% (81/107) showed concordance, whilst 24% (26/107) showed discordance. For this calculation, we excluded the 33 'unsure' responses. The validity, estimated at 76% in the study, was based on this one question.

Most respondents (69%; 96/140) indicated that their institutions had policies on research data protection, and 50% (70/140) specified that restrictions and/or prohibitions for the trans-border flow of research

data were also in place. Interestingly, just over a third (34%) of the respondents (48/140) mentioned that their affiliated institutions had no restrictions for the trans-border flow of research data.

Perceptions of the current laws and policies on research data protection and transfer

Respondents were asked to indicate how much they agreed or disagreed (on a six-point scale) with statements about the adequacy of their country's laws and institutional policies on research data protection (Table 4). Of the respondents, 45% (63/140) expressed the view that their country's current laws on research data protection were adequate, whereas 19% (27/140) disagreed. Of those who disagreed, 9% (12/140) disagreed strongly. Similarly, 40% (56/140) of respondents perceived their national restrictions and prohibitions on the trans-border flow of research data to be adequate. Of those who agreed, only 7% (10/140) agreed strongly. Just over half (51%) of all respondents (72/140) perceived their institutional policies on research data protection to be adequate.

On the other hand, a quarter (25%) of the respondents (35/140) indicated that their national restrictions and prohibitions on the trans-border flow of research data were inadequate. Slightly fewer (21%; 29/140) felt that their institutional policies on research data protection were also inadequate.

Table 1: Characteristics of survey respondents ($N = 140$)

Characteristic	<i>n</i> (%)
Gender	
Male	88 (63)
Female	51 (36)
Other	1 (1)
Education/qualification	
Bachelor's degree	7 (5)
Honours degree	6 (4)
Master's degree	58 (41)
Doctoral degree	64 (46)
Other	5 (4)
Type of REC	
National	37 (26)
Institutional	96 (67)
Private	7 (5)
Position/role	
Chair	19 (14)
Co-chair	4 (3)
Vice-chair	5 (4)
Member	112 (80)
Years of experience	
Less than 2 years	20 (14)
2–4 years	19 (14)
4–6 years	29 (21)
6–8 years	13 (9)
8 or more years	59 (42)

Table 2: Respondents' awareness of laws and policies on research data protection ($N = 140$)

Any law on research data protection in the country?	<i>n</i> (%)
Yes	82 (59)
No	24 (17)
Unsure	34 (24)
Restrictions/prohibitions placed on the trans-border flow of research data in the country?	
Yes	67 (48)
No	34 (24)
Unsure	39 (28)
Any policy on research data protection at the institution?	
Yes	96 (69)
No	26 (19)
Unsure	18 (13)
Restrictions/prohibitions placed on the trans-border flow of research data at the institution?	
Yes	70 (50)
No	48 (34)
Unsure	22 (16)

Table 3: Validation of responses received ($N = 140$)

Responses	Existing privacy laws		Total
	Yes	No	
Yes	67/107 (63%)	16/107 (15%)	83/107 (78%)
No	10/107 (9%)	14/107 (13%)	24/107 (22%)
Total	77/107 (72%)	30/107 (28%)	107 (100%)

Table 4: Respondents' perceptions of data-related laws or policies ($N = 140$)

Adequate law on research data protection within the respondent's country	<i>n</i> (%)
None (no law or policy)	25 (18)
Disagree strongly	12 (9)
Disagree somewhat	15 (11)
Unsure	25 (18)
Agree somewhat	46 (33)
Agree strongly	17 (12)
Adequate restrictions or prohibitions on the trans-border flow of research data at country level	
None (no law or policy)	16 (11)
Disagree strongly	17 (12)
Disagree somewhat	18 (13)
Unsure	33 (24)
Agree somewhat	46 (33)
Agree strongly	10 (7)
Adequate institutional-level policy on research data protection	
None (no law or policy)	17 (12)
Disagree strongly	8 (6)
Disagree somewhat	21 (15)
Unsure	22 (16)
Agree somewhat	54 (39)
Agree strongly	18 (13)
Adequate institutional-level restrictions or prohibitions on the trans-border flow of research data	
None (no law or policy)	22 (16)
Disagree strongly	12 (9)
Disagree somewhat	18 (13)
Unsure	31 (22)
Agree somewhat	46 (33)
Agree strongly	11 (8)

Transfer agreements

Awareness of MTAs and DTAs was generally good, but around 20% of respondents (28/140) were uncertain of the existence of such agreements. Just over a third (36%; 50/140) indicated that their institutions had a

separate DTA in place. Most respondents (74%; 103/140) indicated that their REC was required to review DTAs and MTAs. Only 13% (18/140) indicated that their REC did not review these documents (Table 5).

Table 5: Protection of research data or HBM (N = 140)

Separate DTA available at the respondent's institution	n (%)
Yes	50 (36)
No	90 (64)
Separate MTA available at the respondent's institution	
Yes	74 (53)
No	66 (47)
Combined DTA and MTA available at the respondent's institution	
Yes	33 (24)
No	107 (76)
My institution has appropriate regulatory policies in place	
None	16 (11)
Disagree strongly	8 (6)
Disagree somewhat	6 (4)
Unsure	19 (14)
Agree somewhat	48 (34)
Agree strongly	43 (31)
My institution has appropriate ethics guidance in place	
None	11 (8)
Disagree strongly	3 (2)
Disagree somewhat	5 (4)
Unsure	12 (9)
Agree somewhat	44 (31)
Agree strongly	65(46)

HBM, human biological material; DTA, Data Transfer Agreement; MTA, Material Transfer Agreement

Most respondents (64%; 89/140) indicated that they lacked experience in reviewing data-intense protocols that involve data sharing, as up to 50% of all protocols that they reviewed did not relate to data at all, whilst only 14% of respondents (19/140) indicated that more than half of their reviewed protocols related purely to large data sets or big data.

Support systems for REC members

Respondents were asked to indicate the ease of accessing their country's data regulatory body for consultation. Over a third (38%) of respondents (53/140) indicated that they could easily do so, whereas 25% (35/140) disagreed. A portion of respondents (12%; 17/140) indicated that no data regulatory body existed within their country.

A minority of respondents 14% (20/140) indicated that they had received no training on how to review protocols involving data sharing. A fifth (21%) of respondents (30/140) indicated that their institution did not have appropriate regulatory policies on the protection of research data and/or HBMs. Likewise, 14% of respondents (19/140) indicated that their institution did not have appropriate ethics guidance on the protection of research data and/or HBMs (Table 4).

Challenges with data governance

Just over a third (36%) of respondents (51/140) indicated that they faced challenges in their countries regarding the development of

legal frameworks or guidance for research data protection. Only 59% of respondents (82/140) reported having current national laws on data protection. The reasons provided were based on poor resources available within these countries, coupled with a lack of capacity to focus on the development of legislation:

The lack of law is the main challenge to be recorded in SSA. [Country 1]

Specific guidance/law for research data protection is not developed at country level. Laws and [the] Constitution address issues related to data protection in fragmented ways. [Country 2]

Respondents raised a lack of adequately trained legal and ethical experts as another challenge:

The legal experts who develop legal frameworks or guidance for research data protection have not been trained in research ethics. As such, the current legal frameworks for research data protection lack ethical input. Secondly, the current legal frameworks are very restrictive because the regulators are rigid and do not want to move with the signs of the times. [Country 3]

Lack of legal and ethics experts to develop the frameworks...Lack of trained personnel in this field.... [Country 4]

The lack of awareness regarding research ethics and related issues was raised as an issue:

There is a shortage of knowledge amongst clinician practitioners involved in research requiring the implications of the Protection of Personal Information Act. [Country 5]

Respondents also identified the lack of clear DTAs for many countries in SSA as a hindrance to good data governance:

We need to come up with a clear DTA. [Country 6]

Addressing issues related to data in collaborative research. Issues of consent for secondary use of data – use of data for other research not included in the original protocol for which informed consent was provided. [Country 7]

The majority of respondents (66%; 93/140) revealed that they experience some level of difficulty in reviewing data sharing related protocols (Figure 2).

Suggested improvements

Most respondents (71%; 99/140) expressed the view that data sharing for research could be better regulated at their institution. Respondents emphasised a need for the development of institutional policies with clear guidelines for implementation and adequate processes for the follow-up of research protocols. Suggestions around the potential development of DACs within institutions emerged as an idea for the better regulation of data sharing within research.

More than half the respondents (64%; 89/140) indicated that their institutions did not have DACs to handle data-related issues in research. These findings further highlight the need for a DAC as it relates to institutional regulation.

This should start from drafting laws and policies that specifically govern/regulate specimen and data sharing. Research institutions can then draw from these to develop their standard operating procedures or guidelines. External research partners can develop capacity in this area through funding [the] training of IRB members involved in the review of protocols that involve samples and data sharing. [Country 8]

By establishing Data Access Ethics Subcommittees to function under RECs, or better still, provision of training to RECs so that they can play the regulatory role. [Country 7]

Many respondents suggested the development of comprehensive DTAs to improve regulation at a national level. Qualitative responses highlighted the importance of local and international collaboration and the increased need for support to researchers.

The need to raise awareness through education among research stakeholders, including IRB members, researchers, communities, as well as respondents about the benefits and risks of data sharing. This empowerment will encourage research stakeholders to appreciate the need for [the] regulation of samples and data sharing to avoid unethical practices in sample and data sharing like exploitation and harm to individual respondents and communities where the research is conducted. [Country 8]

We need to support researchers to understand the bigger value of data and appreciate [the] value of engaging in data agreements with collaborating institution, which business they have been leaving to the regulator. [Country 9]

Discussion

Historically, RECs have been tasked with reviewing classic clinical trials and other research protocols with limited data sets.²⁴ Robust governance frameworks exist globally and in SSA to guide this type of research review.²⁵ Likewise, a reasonable amount of capacity development has occurred in research ethics review in SSA.²⁵ Big data have raised new ethics and legal challenges²⁶, and our results provide a broad overview of these challenges in SSA. To our knowledge, this is the first empirical survey in SSA in which awareness and perspectives of REC members have been explored specifically as they relate to the review of data-intensive research.

There are governance challenges relating to data protection in research as not all countries in SSA have a legal framework to regulate the use of big data in research. Instead, there is a spectrum of legal regulation, ranging from the strict, comprehensive protection of data to no legal frameworks at all.²⁷⁻²⁹ Likewise, research ethics policies and guidelines suffer the same level of variability across the subcontinent where big data are concerned.²⁵

Our study confirms this variability as knowledge and awareness of legislative frameworks and ethics guidance in SSA vary considerably. Only 58% of the REC members surveyed indicated that laws existed at a national level, with the remainder indicating no knowledge or uncertainty about the existence of such laws. More specifically, a quarter (24%) of REC members were uncertain about whether such frameworks existed within their respective countries or institutions.

Most concerning is the apparent lack of legislative frameworks for the cross-border transfer of big data on the subcontinent and out of Africa to other parts of the world. This is important because of the historical concern with data and samples leaving SSA in an unregulated manner, which raises concerns about exploitative research practices.³⁰⁻³² Although just under two-thirds of respondents were unaware of laws relating to data-intensive research, only half were aware of laws relating to the cross-border transfer of data. This suggests that research data may be crossing borders without agreements or export permits in place. This is supported by Labuschaigne et al.³³ who reported that HBMs may be leaving South Africa without export permits or MTAs during collaborative research. Mwaka and Munabi³⁴, who undertook a similar study on perceptions and experiences on the transfer of HBMs in international collaborative research in Uganda, reported that the development of an MTA and its implementation lacked transparency.

This concern is reflected at a more granular level as knowledge or awareness of DTAs and DACs demonstrate. Our findings reflect this, as 13% of respondents indicated that some countries and/or institutions do not have DTAs or MTAs in place to regulate the national or trans-border sharing of data. While MTAs were more common than DTAs, a fifth of the respondents were not even certain whether such transfer agreements existed within their affiliated institutions. Notably, although our findings indicate the absence of DTAs or MTAs at some institutions within SSA, most respondents (74%) indicated that their RECs were still responsible for reviewing these legal documents together with data sharing-related research protocols when required. This raises concern about the quality of review being conducted on the DTAs and MTAs submitted to RECs. Respondents perceived the development of comprehensive DTAs focused on safeguarding the privacy, anonymity and confidentiality of research participants as an effective resolution. Respondents emphasised that these DTAs should be stringent, with importance placed on institutions instigating mechanisms to improve regulatory compliance. Suggestions included consultation with legal experts in the development of new DTAs, or improvements to current DTAs to ensure that they are aligned to existing laws or regulations. The implementation of access control systems that concentrate on standard criteria for data use and propositions may reduce the likelihood of data misuse, and may legally complement data transfer across borders.

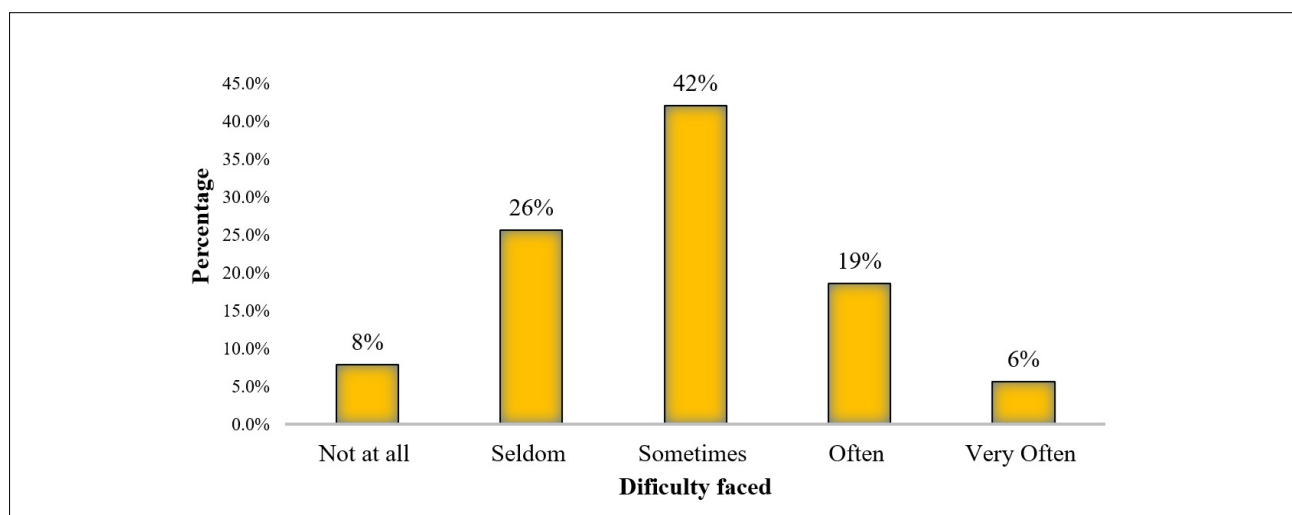


Figure 2: Difficulty in reviewing data-sharing protocols.

Some respondents were of the view that their country's laws were fragmented and consequently exacerbated ethical challenges, thus needing to be harmonised. This was echoed in the responses indicating that data sharing for research could be better regulated both within their institutions (70%) and nationally (71%). Suggestions to develop policies with clear frameworks or stringent standard operating procedures on data sharing emerged, along with improving awareness and access to adequate training on protocol review, data sharing, processing and protection. Likewise, over a third of respondents were not aware of the restrictions placed on the trans-border flow of research data at their institutions.

Many challenges exist in data governance in SSA. The lack of legal and ethics expertise within RECs was recognised as a challenge in adequately reviewing research protocols that related to big data, research transfer agreements and in developing frameworks and policies. Some respondents reported that their institutions do not have ethics (11%) and regulatory (8%) guidance in place for the protection of research data or HBMs, whilst others reported being unsure about whether such ethics (14%) and regulatory (9%) guidance were utilised within their institutions. These findings are comparable with the systematic review conducted by Barchi and Little²⁸, who found that 29 of the 49 SSA countries (59%) had some form of national ethics guidance. Barchi and Little concluded that SSA countries that still lacked regulatory guidance on research data or HBMs would require extensive health-system strengthening in ethics governance before they could be fully engaged in the modern research enterprise.²⁸

Respondents reported the development of adequate legal frameworks or ethics guidance and policies for research data protection within their respective countries as a pressing challenge. A lack of resources was identified as a common reason for this as respondents expressed an increased need for resources, such as training, to efficiently develop and maintain legislative frameworks for data protection in SSA.

Although some of the epistemic gaps presented with RECs could be addressed, some of the committees' responsibilities may be seen as falling outside their mandate and scope of function. This drew attention to the question of who should review such documents when an epistemological challenge exists amongst RECs. Some authors have argued that such responsibility is incompatible with RECs' legislative oversight role and that a legal body is better suited to review such legal documents.¹¹

The current lack of training available in the field of data science for REC members to better handle the ethical, legal and social implications of big data-related research highlights the need to proactively educate and train²⁶ SSA research-based institutions to foster and empower the formation of DACs^{13,35}. While most respondents confirmed that their institutions lacked DACs to handle data-related issues in research, such committees could play a significant role in the data governance ecosystem.^{13,35} The suggestion to form institutional DACs emerged from our study results; however, respondents also indicated that difficulty may be encountered in establishing these committees with members of sufficient and diverse knowledge, skills and experience.

Training needs were evident across the subcontinent. REC members recognised a deficit in their experience and expertise pertaining to the review of research protocols involving big data and related research transfer agreements. This is evident in the large cohort of respondents (64%) that were not often exposed to research protocols that related purely to large data sets or big data as they clearly indicated that the bulk of all research protocols reviewed did not relate to data sharing at all. This finding was further strengthened by the third (32%) of respondents in our study who explicitly stated that they had not received any training on reviewing protocols involving data use and data sharing. Interestingly, 23% of respondents expressed uncertainty on whether they engage with data sharing related research protocols as a result of not entirely understanding what data sharing and big data essentially encompass. This training deficit is not unique to SSA. Ferretti et al.² found that REC members in Switzerland faced similar challenges in adequately reviewing protocols involving big data research due to an existing lack of expertise and experience in the field.^{2,36} In Australia, Pysar et al.¹⁵ revealed that genomic confidence scores in reviewing related research protocols were low amongst REC members

that were less experienced, and had less exposure and training in the field. Hence, most participants (76%) in this study indicated that non-genetics experts that serve on RECs require additional training and/or resources on big data research. Equipping RECs with basic epistemological advantages, in the form of skills and knowledge in big data, would allow them to better fulfil their roles in effectively reviewing data-sharing protocols.

Pisa et al.³⁷ proposed addressing funding issues, strengthening data management systems, providing training and conducting workshops to strengthen regulatory capacity. This will reduce and mitigate instances of data exploitation or harm encountered by research participants and data subjects.

Study limitations

A notable limitation to be acknowledged when interpreting the results of this study is the predominance of responses from some SSA countries compared to other countries (indicated in Figure 1). This may be due to a higher number of RECs in these countries, more active research sites and the fact that it was easier to locate active email contacts from representatives of these SSA countries. These findings were also from a relatively small survey. Potential participants without reliable internet access may have been unintentionally excluded from participation given the internet-based nature of the survey. Because these results were confined to the SSA context, and 15 of the SSA countries did not participate in our survey, we may not have been able to represent the entire continuum of variability present within the SSA region. However, given the absence of empirical studies on the awareness and perspectives of REC members in SSA, these limitations do not pose a major threat to our survey's exploratory aim. Our qualitative research may address some of these limitations and will be published separately.

Overall, our highest number of survey responses was obtained from the Democratic Republic of the Congo, Kenya, Mozambique, Nigeria, South Africa and Uganda. This may be because most of these countries (South Africa, Nigeria, Kenya and Uganda)³⁸ are ranked as the most research-intense countries in SSA by research output in the fields of public health, and environmental and occupational health³⁹⁻⁴¹. The increased research activities in these SSA countries may be associated with increased cross-border data transfer.

South Africa and Kenya are the most stringent in their data export protection. For data to be transferred out of these countries, the data transfer must be purposeful, consent must be obtained from data subjects, and the data processor must verify to the data commissioner that the third-party recipient's jurisdiction is bound by appropriate safeguards for the security and protection of the data.⁴² Yet, our results did not entirely reflect this, as not all responses from Kenya appeared to be in agreement, indicating a divide. Likewise, a divide was observed in the aggregated results from Nigeria, although the country is very research active. This may be because the country's moderately rigid data export protection does not require third-party recipients of data to be bound by adequate data protection laws or agreements in cases where consent is acquired, or where the transfer meets an exception.^{29,38} For South Africa, the highest-ranked SSA country by research output in public health, and environmental and occupational health³⁸, our results reveal consensus amongst respondents regarding cross-border data transfers, which may be due to awareness of POPIA^{29,43}.

Conclusion

In this study, we intended to provide a broad overview of REC members' awareness and perceptions on data governance in SSA and related legal and ethical challenges. Our results uncovered valuable insights and offer a novel contribution to the empirical literature in SSA on big data. Our findings indicate variability in data governance and regulation in SSA, as well as variability in REC members' perceptions of the adequacy of their national laws and institutional policies. Suboptimal awareness of the existence of data protection laws or the lack thereof amongst REC members in the sample was concerning. This will impact negatively on how data-intense protocols are reviewed. There is a unanimous expressed need for the training of REC members on the African continent. Established RECs across SSA would benefit from the reformation of practices and oversight mechanisms, expertise and regulations to better cater for the

big data research context. Transparent, robust and standardised data governance may promote shared ethical values to conduct research with big data on the subcontinent. Data governance within SSA continues to be inadequately supported by legislative and enforcement frameworks.

Acknowledgements

Research reported in this publication was supported by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank all respondents for their invaluable time in participating in the study. We also thank Ms Qunita Brown (MA) for her assistance in analysing qualitative responses from the questionnaire. We furthermore acknowledge the following individuals for their guidance during the early stages of the project: Aneeka Domingo, Theresa Burgess, Shenuka Singh, Meagan Jacobs-Alfred, Sharon Kling, Gonasagrie Nair and Emmanuel Obasa.

Competing interests

We have no competing interests to declare.

Authors' contributions

K.M. conceptualised the study, reviewed and edited the survey instrument and protocol, and reviewed and edited the manuscript. N.C. developed and submitted the protocol for ethics committee review, designed the online survey instrument, was involved in participant recruitment, led the development of the manuscript and was responsible for data capturing. S.M.K. undertook the data cleaning and analysis, and contributed to the discussion of the results. T.M.E. provided expert advice on statistical analysis. All authors read and approved the final manuscript.

References

- Borgman CL. The conundrum of sharing research data. *J Am Soc Inf Sci Tec*. 2012;63(6):1059–1078. <https://doi.org/10.1002/asi.22634>
- Ferretti A, Ienca M, Velarde MR, Hurst S, Vayena E. The challenges of big data for research ethics committees: A qualitative Swiss study. *J Empir Res Hum Res Ethic*. 2022;17(1–2):129–143. <https://doi.org/10.1177/15562646211053538>
- Walport M, Brest P. Sharing research data to improve public health. *Lancet*. 2011;377(9765):537–539. [https://doi.org/10.1016/S0140-6736\(10\)62234-9](https://doi.org/10.1016/S0140-6736(10)62234-9)
- Kaewkungwal J, Adams P, Sattabongkot J, Lie RK, Wendler D. Issues and challenges associated with data-sharing in LMICs: Perspectives of researchers in Thailand. *Am J Trop Med Hyg*. 2020;103(1):528–536. <https://doi.org/10.4269/ajtmh.19-0651>
- Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine, C, James, A, et al. Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *Ann Intern Med*. 2017;167(1):63–65. <https://doi.org/10.7326/M17-1028>
- Can Panhuis WG, Paul P, Emerson C, Grefenstette, J, Wilder, R, Herbst, AJ, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health*. 2014;14, Art. #1144. <https://doi.org/10.1186/1471-2458-14-1144>
- Alter GC, Vardigan M. Addressing global data sharing challenges. *J Empir Res Hum Res Ethics*. 2015;10(3):317–323. <https://doi.org/10.1177/1556264615591561>
- Staunton C, Tschigg K, Sherman G. Data protection, data management, and data sharing: Stakeholder perspectives on the protection of personal health information in South Africa. *PLoS ONE*. 2021;16(12), e0260341. <https://doi.org/10.1371/journal.pone.0260341>
- Knoppers BM, Harris JR, Budin-Ljøsne I, Dove ES. A human rights approach to an international code of conduct for genomic and clinical data sharing. *Hum Genet*. 2014;133:895–903. <https://doi.org/10.1007/s00439-014-1432-6>
- Chalmers D, Nicol D, Nicolás P, Zeps N. A role for research ethics committees in exchanges of human biospecimens through material transfer agreements. *J Bioeth Inq*. 2014;11:301–306. <https://doi.org/10.1007/s11673-014-9552-1>
- Thaldar DW, Botes M, Nienaber A. South Africa's new standard material transfer agreement: Proposals for improvement and pointers for implementation. *BMC Med Ethics*. 2020;21, Art. #85. <https://doi.org/10.1186/s12910-020-00526-x>
- Polanin JR, Terzian M. A data-sharing agreement helps to increase researchers' willingness to share primary data: Results from a randomized controlled trial. *J Clin Epidemiol*. 2019;106:60–69. <https://doi.org/10.1016/j.jclinepi.2018.10.006>
- Cheah PY, Piasecki J. Data access committees. *BMC Med Ethics*. 2020;21, Art. #12. <https://doi.org/10.1186/s12910-020-0453-z>
- Sellers C, Samuel G, Derrick G. Reasoning "uncharted territory": Notions of expertise within ethics review panels assessing research use of social media. *J Empir Res Hum Res Ethics*. 2020;15(1):28–39. <https://doi.org/10.1177/1556264619837088>
- Pysar R, Wallingford CK, Boyle J, Campbell SB, Eckstein L, McWhirter R, et al. Australian human research ethics committee members' confidence in reviewing genomic research applications. *Eur J Hum Genet*. 2021;29:1811–1818. <https://doi.org/10.1038/s41431-021-00951-5>
- Ballantyne A. Adjusting the focus: A public health ethics approach to data research. *Bioethics*. 2019;33(3):357–366. <https://doi.org/10.1111/bioe.12551>
- Odusote A. Data misuse, data theft and data protection in Nigeria: A call for a more robust and more effective legislation. *Beijing Law Rev*. 2021;12(4):1284–1298. <https://doi.org/10.4236/blr.2021.124066>
- Jang-Jaccard J, Nepal S. A survey of emerging threats in cybersecurity. *J Comput Syst Sci*. 2014;80(5):973–993. <https://doi.org/10.1016/j.jcss.2014.02.005>
- Ducato R. Data protection, scientific research, and the role of information. *Comput Law Secur Rev*. 2020;37, Art. #105412. <https://doi.org/10.1016/j.clsr.2020.105412>
- Bubela T, Guebert J, Mishra A. Use and misuse of material transfer agreements: Lessons in proportionality from research, repositories, and litigation. *PLoS Biol*. 2015;13, e1002060. <https://doi.org/10.1371/journal.pbio.1002060>
- Department of Health Republic of South Africa. Ethics in health research principles, processes and structures [document on the Internet]. c2015 [cited 2022 Aug 01]. Available from: [https://www.ru.ac.za/media/rhodesuniversity/content/ethics/documents/nationalguidelines/DOH_\(2015\)_Ethics_in_health_research_Principles,_processes_and_structures.pdf](https://www.ru.ac.za/media/rhodesuniversity/content/ethics/documents/nationalguidelines/DOH_(2015)_Ethics_in_health_research_Principles,_processes_and_structures.pdf)
- Knight J. The need for improved ethics guidelines in a changing research landscape. *S Afr J Sci*. 2019;115(11/12), Art. #6349. <https://doi.org/10.17159/sajs.2019/6349>
- Naderifar M, Goli H, Ghaljaie F. Snowball sampling: A purposeful method of sampling in qualitative research. *Stride Dev Med Educ*. 2017;14(3), e67670. <https://doi.org/10.5812/sdme.67670>
- Tusino S, Furfaro M. Rethinking the role of research ethics committees in the light of Regulation (EU) No 536/2014 on clinical trials and the COVID-19 pandemic. *Br J Clin Pharmacol*. 2022;88(1):40–46. <https://doi.org/10.1111/bcp.14871>
- Nabyonga-Orem J, Asamani JA, Makanga M. The state of health research governance in Africa: What do we know and how can we improve? *Health Res Policy Syst*. 2021;19, Art. #11. <https://doi.org/10.1186/s12961-020-00676-9>
- Ferretti A, Ienca M, Sheehan M, Blasimme A, Dove ES, Farsides B, et al. Ethics review of big data research: What should stay and what should be reformed? *BMC Med Ethics*. 2021;22, Art. #51. <https://doi.org/10.1186/s12910-021-00616-4>
- Townsend B. The lawful sharing of health research data in South Africa and beyond. *Inf Commun Technol Law*. 2022;31(1):17–34. <https://doi.org/10.1080/13600834.2021.1918905>
- Barchi F, Little MT. National ethics guidance in sub-Saharan Africa on the collection and use of human biological specimens: A systematic review. *BMC Med Ethics*. 2016;17, Art. #64. <https://doi.org/10.1186/s12910-016-0146-9>
- Brand D, Singh JA, Nienaber McKay AG, Cengiz N, Moodley K. Data sharing governance in sub-Saharan Africa during public health emergencies: Gaps and guidance. *S Afr J Sci*. 2022;118(11/12), Art. #13892. <https://doi.org/10.17159/sajs.2022/13892>
- Moodley K. Research imperialism resurfaces in South Africa in the midst of the COVID-19 pandemic – this time, via a digital portal. *S Afr Med J*. 2020;110(11):1068–1069. <https://doi.org/10.7196/SAMJ.2020.v110i11.15285>



31. Moodley K, Singh S. "It's all about trust": Reflections of researchers on the complexity and controversy surrounding biobanking in South Africa. *BMC Med Ethics*. 2016;17, Art. #57. <https://doi.org/10.1186/s12910-016-0140-2>
32. Singh S, Moodley K. Stakeholder perspectives on the ethico-legal dimensions of biobanking in South Africa. *BMC Med Ethics*. 2021;22, Art. #84. <https://doi.org/10.1186/s12910-021-00645-z>
33. Labuschaigne M, Dhai A, Mahomed S, Behrens K, Nienaber A, Moodley K, et al. Protecting participants in health research: The South African Material Transfer Agreement. *S Afr Med J*. 2019; 109(5):353–356. <https://doi.org/10.7196/SAMJ.2019.v109i5.13803>
34. Mwaka ES, Munabi IG. Trans-border transfer of human biological materials in collaborative biobanking research: Perceptions and experiences of researchers in Uganda [preprint]. *medRxiv*. 2022. <https://doi.org/10.1101/2022.04.01.22273073>
35. Kaye J, Hawkins N. Data sharing policy design for consortia: Challenges for sustainability. *Genome Med*. 2014;6, Art. #4. <https://doi.org/10.1186/gm523>
36. Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. *PLoS ONE*. 2018;13(10), e0204937. <https://doi.org/10.1371/journal.pone.0204937>
37. Pisa M, Dixon P, Nwankwo U. Why data protection matters for development: The case for strengthening inclusion and regulatory capacity. Center for Global Development; 2021. Available from: <https://www.cgdev.org/sites/default/files/why-data-protection-matters-for-development.pdf>
38. Scimago Journal and Country Rank. Country rankings: Public health, environment, and occupational health [webpage on the Internet]. No date [cited 2022 Aug 01]. Available from: <https://www.scimagojr.com/countryrank.php?region=Africa&category=2739>
39. Lucas-Dominguez R, Alonso-Arroyo A, Vidal-Infer A, Aleixandre-Benavent R. The sharing of research data facing the COVID-19 pandemic. *Scientometrics*. 2021;126:4975–4990. <https://doi.org/10.1007/s11192-021-03971-6>
40. Moorthy V, Henao Restrepo AM, Preziosi M-P, Swaminathan S. Data sharing for novel coronavirus (COVID-19). *Bull World Health Organ*. 2020;98:150. <https://doi.org/10.2471/BLT.20.251561>
41. Capocasa M, Anagnostou P, Bisol GD. A light in the dark: Open access to medical literature and the COVID-19 pandemic. *Inf Res*. 2022;27(2), Art. #929. <https://doi.org/10.47989/irpaper929>
42. Suominen K, Vambell E. Alliance for e-trade development: Toward an African data transfer regime to enable MSMEs' cross-border ecommerce [document on the Internet]. c2021 [cited 2022 May 06]. Available from: https://www.allianceforetradedevelopment.org/_files/ugd/478c1a_72021e35a826441db0723642a79e65e5.pdf
43. Academy of Science of South Africa (ASSAf). POPIA Code of Conduct for Research [document on the Internet]. c2021 [cited 2022 May 03]. Available from: https://www.assaf.org.za/files/2020/POPIA%20CoC%20Research_Conceptnote_Letter%20003.pdf

**AUTHORS:**

Siti M. Kabanda¹
 Nezerith Cengiz¹
 Kanshukan Rajaratnam²
 Bruce W. Watson²
 Qunita Brown¹
 Tonya M. Esterhuizen³
 Keymanthri Moodley¹

AFFILIATIONS:

¹Centre for Medical Ethics and Law, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

²School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa

³Division of Epidemiology and Biostatistics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

CORRESPONDENCE TO:

Nezerith Cengiz

EMAIL:

ncengiz@sun.ac.za

DATES:

Received: 08 Nov. 2022

Revised: 04 May 2023

Accepted: 05 May 2023

Published: 30 May 2023

HOW TO CITE:

Kabanda SM, Cengiz N, Rajaratnam K, Watson BW, Brown Q, Esterhuizen TM, et al. Data sharing and data governance in sub-Saharan Africa: Perspectives from researchers and scientists engaged in data-intensive research. *S Afr J Sci.* 2023;119(5/6), Art. #15129. <https://doi.org/10.17159/sajs.2023/15129>

ARTICLE INCLUDES:

- Peer review
- [Supplementary material](#)

DATA AVAILABILITY:

- Open data set
- All data included
- On request from author(s)
- Not available
- Not applicable

EDITOR:

Floretta Boonzaier

KEYWORDS:

big data, data governance, sub-Saharan Africa, researchers, scientists, data transfer agreements, data sharing

FUNDING:

US National Institutes of Health (1U01MH127704-01)



Data sharing and data governance in sub-Saharan Africa: Perspectives from researchers and scientists engaged in data-intensive research

The data ecosystem is complex and involves multiple stakeholders. Researchers and scientists engaging in data-intensive research collect, analyse, store, manage and share large volumes of data. Consequently, capturing researchers' and scientists' views from multidisciplinary fields on data use, sharing and governance adds an important African perspective to emerging debates. We conducted a descriptive cross-sectional survey and received 160 responses from researchers and scientists representing 43 sub-Saharan African countries. Whilst most respondents were satisfied with institutional data storage processes, 40% indicated that their organisations or institutions did not have a formally established process for storing data beyond the life cycle of the project. Willingness to share data was generally high, but increased when data privacy was ensured. Robust governance frameworks increased the willingness to share, as did the regulation of access to data on shared platforms. Incentivising data sharing remains controversial. Respondents were satisfied with exchanging their data for co-authorship on publications (89.4%) and collaboration on projects (77.6%). However, respondents were split almost equally in terms of sharing their data for commercial gain. Regarding the process of managing data, 40.6% indicated that their organisations do not provide training on best practices for data management. This could be related to a lack of resources, chronic institutional under-investment, and suboptimal research training and mentorship in sub-Saharan Africa. The sustainability of data sharing may require ethical incentive structures to further encourage researchers and scientists. Tangible infrastructure to facilitate such sharing is a prerequisite. Capacity development in data governance for researchers and scientists is sorely needed.

Significance:

Data sharing is necessary to advance science, yet there are many constraints. In this study, we explored factors that promote a willingness to share, as well as constraining factors. Seeking potential solutions to improve data sharing is a scientific and ethical imperative. The standardisation of basic data sharing and data transfer agreements, and the development of a Data Access Committee will strengthen data governance and facilitate responsible data sharing in sub-Saharan Africa. Funders, institutions, researchers and scientists ought to jointly contribute to fair and equitable data use and sharing during and beyond the life cycle of research projects.

Background

High-quality and accurate data generated via research have enormous transformative potential for evidence-based decision-making, together with data analytics that helps to improve the tracking of targets that have been put in place.^{1,2} Such advantages, which emanate from the digital revolution, are embodied as velocity, veracity and variability.³⁻⁵ The consideration of transparency, sharing, governance and management frameworks regarding big data become more challenging in the context of volume, velocity and variety. High-quality data create the foundation of science, regardless of volume (small data or big data), whilst also serving a vital role in informing sound decision-making for optimal action.⁶ As data become a focal point of innovative scientific discovery, data sharing by researchers and scientists has become a critical aspect of scientific advancement.⁷ Data sharing is described as the act of providing access by transferring data in a form that can be used by other individuals.^{6,8} Its prominence in current research debates is premised on open science, which is intended to make data and scientific research widely accessible.^{7,9} This is especially important given that most published articles are not available to people without a personal or institutional subscription, and most data are not made available on public repositories.¹⁰ As a result, the open science movement has the potential to revolutionise scientific research and improve its transparency and potential for collaboration.^{10,11} Additionally, this encourages researchers and scientists to share their data with others, which can lead to numerous benefits, such as increased scientific reproducibility, robustness and new opportunities for collaboration, thereby enriching the potential to inform interventions or policy decisions.^{7,9} Various initiatives, such as the Transparency and Openness Promotion (TOP) guidelines and the findable, accessible, interoperable and reusable (FAIR) principles, offer guidance for the improved clarity and reproducibility of research.^{10,12} By funding replication studies and recognising and crediting their efforts, researchers can be incentivised to engage in open science practices that can promote transparency, collaboration and innovation in scientific research.^{10,13} Various stakeholders, publishers, funders, custodians of data repositories, tertiary and research institutions, and librarians play a pivotal role in developing structures and systems that support and promote data sharing.^{14,15}

Data-sharing policies, such as the Bermuda Principles, the Fort Lauderdale Principles and the Global Alliance for Genomics and Health, expose key principles on open access to genome sequence data¹⁶⁻¹⁸ with the aim of accelerating advances in science by supporting the free and unrestricted use of such data¹⁹. The adoption of access policies for publicly funded research has replaced the previous divisive lack of consensus amongst funding agencies and research institutions.¹⁹

Despite the benefits of data sharing in open science, many researchers and scientists remain reluctant to share their data. This stance is driven by privacy or intellectual property concerns, the historical misuse of data, and concerns of being scooped.²⁰⁻²² Kim et al.²³ conducted a survey-based descriptive study on the data-sharing attitudes and practices of researchers in Korean government research institutes. From their work, the most common reasons for withholding data included time- and effort-intensive requirements to organise data, followed by concerns about data ownership and lack of reward or recognition for data sharing.²³ Additionally, Kim et al.²³ found that respondents had concerns about sharing data that contained sensitive information or where there were potential errors within the data. The degree to which scientists or researchers share or withhold data is not solely a personal choice, as institutional and national factors greatly impact data sharing. For instance, in the context of laws, regulations and policies, restrictions may apply to data sources that are copyrighted and may prohibit the publication of certain types of data (i.e. medical records).^{7,24}

Furthermore, data transfer agreements (DTAs) govern the transfer of identifiable human participant data, where voluntary and informed consent have been obtained from participants.^{25,26} Both material transfer agreements (MTAs) and DTAs contractually govern biological material and data transfer between parties to safeguard the interests of stakeholders.^{25,26} These contractual agreements outline the specific purpose(s) for which the data may be used, as well as the related protections, rights and obligations of stakeholders and collaborators. Despite the important role that MTAs and DTAs play in bio-sample and data governance, these agreements are occasionally perceived as an impediment to data sharing, given their complexity and associated bureaucracy.²⁷ As a result, it is important to develop strategies and policies to promote effective data sharing, whilst simultaneously maintaining privacy and confidentiality. Although data-sharing practices vary across fields, data-sharing perceptions and experiences can be similar.²⁸ In a study conducted by Pujol Priego et al.²⁸, researchers in physics, astronomy, life sciences and computer science recognised the benefits of having access to others' data. However, when compared to physics and astronomy researchers, many researchers in life sciences were less eager to share their data. The reluctance to share data in life sciences could depend on ethical and cultural limitations, especially amongst scientists who work with human participants.^{7,29} The difference in perceptions and practices of data sharing across scientific fields is highly determinative in the fields of life sciences, astronomy and physics due to their long-standing tradition of engagement with large volumes of data compared to other fields.²⁸ Nonetheless, most researchers and scientists worldwide have a positive attitude towards data sharing⁷, yet those in low- and middle-income countries (LMICs) face more challenges in this regard.

Various studies illustrate these challenges in LMICs, particularly in sub-Saharan Africa (SSA).³⁰ A study by Bangani and Moyo³¹ found that limited resources increased the reluctance to share data amongst South African researchers. A lack of funding and financial investment in physical infrastructure (i.e. power and the Internet) are contributing factors to the challenges in data availability and accessibility.²¹ Similarly, a Zimbabwean study discovered that persistent power challenges may be a factor in data sharing.³² These struggles are exacerbated by the current inequities in the global research community, which largely excludes researchers from LMICs from actively participating in the progression of science, where they are often relegated to the role of data generator, instead of published author.³³ It is important that researchers and scientists are provided with the necessary resources and government support to reinforce their data-sharing processes.

Furthermore, Skelly and Chiware³⁴ proposed that future policies define the roles of international research funders, journal publishers and inter-institutional and country collaborators to ensure equitable data custodianship in African-generated research. Data sharing is an important component of scientific investigation that should always strive to uphold the rights and interests of all stakeholders.³⁴ This underscores the need for organisations and institutions to have data governance mechanisms in place, such as data management plans and policies that encapsulate ethical data-sharing practices.^{35,36}

Whilst the focus of this paper is not on big data from commercial endeavours, one must note that data regulations govern both commercial and non-commercial big data. Although the difference between commercial and research big data lies in the motive for collecting and analysing data, where private information is involved, both commercial and research entities must treat data with care to ensure good governance.³⁷ The Organisation for Economic Cooperation and Development (OECD) refers to data governance as the:

*...diverse arrangements, including technical, policy, regulatory or institutional provisions, that affect data and their cycle (creation, collection, storage, use, protection, access, sharing and deletion) across policy domains and organisational and national borders.*³⁸

For the purposes of this paper, we define data governance as frameworks and policies that regulate data use, collection, storage or management, protection and sharing. Whilst some SSA countries have such frameworks in place, others still lag behind.^{35,39-41}

One concern is that some countries may be transferring or sharing data without the existence of legislation, institutional policies or frameworks and good data management standards.³⁵ Good data governance supports the generation of high-quality data and the preservation of control over data. South Africa's *Protection of Personal Information Act* (POPIA)⁴² is an example of a firm privacy and security law as it closely resembles Europe's General Data Protection Regulation (GDPR).⁴³ In addition, Data Access Committees (DACs) have been shown to play an essential role in improving data governance within the context of research as they are able to approve or disapprove data access requests after deliberation and consideration of the potential benefit and harm to the individuals from whom the data were sourced, their communities, researchers and other stakeholders.⁴⁴

Considering the big data revolution in the African region, continental researchers and scientists must reflect on data governance and regulation, and what it means to establish effective support systems for the management of large data sets.³⁴ Whilst a growing body of global research has explored the practices and perceptions of researchers and scientists related to data governance and data protection policies and frameworks, there are limited studies on this phenomenon across SSA. Our study, therefore, aimed to address this gap by investigating the perceptions and experiences of researchers and scientists on data governance and data protection policies in SSA. In this paper, we present and discuss our major findings from data use and reuse, data practices, data management support, data sharing and data protection. Finally, we offer recommendations to strengthen data governance and facilitate responsible data sharing in SSA.

Methods

Study design and sampling

We conducted a descriptive cross-sectional online survey with both quantitative and qualitative components with 160 researchers and scientists representing 43 SSA countries from June 2022 to September 2022. The population was selected based on the profession of the participants as a researcher or scientist involved in data-intensive research in SSA. We recruited our sample through a purposive selection of the professional networks of Stellenbosch University's Centre for Medical Ethics and Law across SSA and used a snowballing technique for further recruitment. We also identified potential participants through a desktop search based on their profession. The survey was directly emailed to those who fit the field of study, and they were invited to participate in their personal capacity. The European and Developing Countries Clinical Trials Partnership research network and Stellenbosch University's Faculty of Medicine and Health Sciences' Marketing and Communications newsletters were useful platforms to invite researchers and scientists to participate in the survey. Respondents were invited to anonymously participate in an online survey through Research Electronic Data Capture (REDCap). All respondents provided voluntary electronic consent.

Survey instruments

The questionnaire was designed electronically using REDCap following a review of the current literature related to data sharing and data governance amongst scientists and researchers, and in consultation with experts in the field of big data research (see the [Supplementary material](#)). The face validity of the survey instrument was assessed by piloting the questionnaire with six data scientists and researchers. Minor amendments were made to produce the final version of the questionnaire before its circulation amongst respondents. These amendments included improving the language to enhance the ease of understanding and restructuring ambiguous questions. The questionnaire consisted of 16 closed-ended questions and three open-ended questions addressing demographic characteristics, respondents' perspectives on data use and reuse, data management, data sharing and the use of others' data. Regarding the open-ended qualitative aspect of the study, three questions were asked to briefly explore respondents' thoughts on data protection steps, data use agreements and any additional comments they wished to add. The data collection tool was developed in English and further translated and localised into French and Portuguese by an academic institution's language centre to cater for African Francophone and Lusophone countries. Data were collected through REDCap using mostly pre-defined categorical responses that did not require cleaning. The age category (not reported in our study) was missing in 91 (57%) of the respondents. This field was the only one that was not completed by all respondents. All 160 responses received were included in the analysis.

Data analysis

Data were exported from REDCap to Statistical Package for Social Sciences (SPSS) (version 28) for analysis. Descriptive statistics were used to describe quantitative data using frequencies and/or percentages in tables and bar graphs. For the meaningful interpretation of the survey responses, questions presented on a five-point Likert scale as strongly disagree, disagree somewhat, neither agree nor disagree, agree somewhat and strongly agree were collapsed into three simpler categories: disagree, neither agree nor disagree, and agree.

In terms of the qualitative component of the study, a trained researcher manually analysed the responses to the open-ended questions using thematic analysis. The researcher first familiarised herself with the responses before identifying and creating codes. Thereafter, she identified patterns or recurring responses in the data. Quotations extracted from the data are included in the paper to illustrate findings from the participants' perspectives. A manual method of analysis was employed due to the small volume of qualitative data that emerged from the three open-ended questions.⁴⁵

Ethical aspects

Research integrity was maintained throughout the study and participation in the research remained entirely voluntary. This survey was a minimal-risk study as the questionnaires involved a factual enquiry with educated and empowered respondents who had full capacity to consent or decline participation. The sample was approached in their individual capacities and respondents consented in their personal capacities. Ethics approval was granted by the Health Research Ethics Committee of the Faculty of Medicine and Health Sciences (reference no: N22/03/028) at Stellenbosch University, South Africa.

Results

Demographic information

In total, 160 individuals responded and completed the online survey. The respondents represented 43 of the 49 SSA countries, with 16 countries having at least one respondent (Figure 1).

Most respondents (68.8%) identified as male and were highly educated, with 60% having completed a doctorate, 52.5% being employed within academia and more than two-thirds (79.5%) self-identifying as researchers or scientists (Table 1).

Data use and reuse

Most respondents reported generating their own data (76.3%) and described the sort of data that they worked with most often as research and academic data (58.8%), public health data (55%) or clinical health service data (37.5%) (Table 2).

Table 1: Characteristics of survey respondents (*N* = 160)

Characteristic	<i>n</i> (%)
Job title	
Business analyst	6 (3.8)
Data scientist	13 (8.1)
Developer	1 (0.6)
Researcher	116 (72.5)
Other	24 (15.0)
Gender	
Male	110 (68.8)
Female	50 (31.3)
Education/qualification	
Bachelor's degree	7 (4.4)
Honours degree	3 (1.9)
Master's degree	50 (31.3)
Doctoral degree	96 (60)
Other	4 (2.5)
Employment by sector	
Academia	84 (52.5)
Government or public sector	33 (20.6)
Commercial	2 (1.3)
Not-for-profit organisation	37 (23.1)
Other	4 (2.5)

Table 2: Data use among respondents (*N* = 160)

Which term/s best describe/s the type of data you use? (Multiple selections applicable)	<i>n</i> (%)
Public health data	88 (55)
Clinical health services data	60 (37.5)
Research and academic data	94 (58.8)
Environmental data	19 (11.9)
Behavioural and socio-economic data	24 (15)
Health capabilities data	23 (14.4)
Information and communication technologies industry data	15 (9.4)
Individual and group data	20 (12.5)
Non-health data	9 (5.6)
Experimental	23 (14.4)
Interviews	28 (17.5)
Observational	29 (18.1)
Other	3 (1.9)
Do you own or generate the data you work with?	
Yes	122 (76.3)
No	38 (23.8)

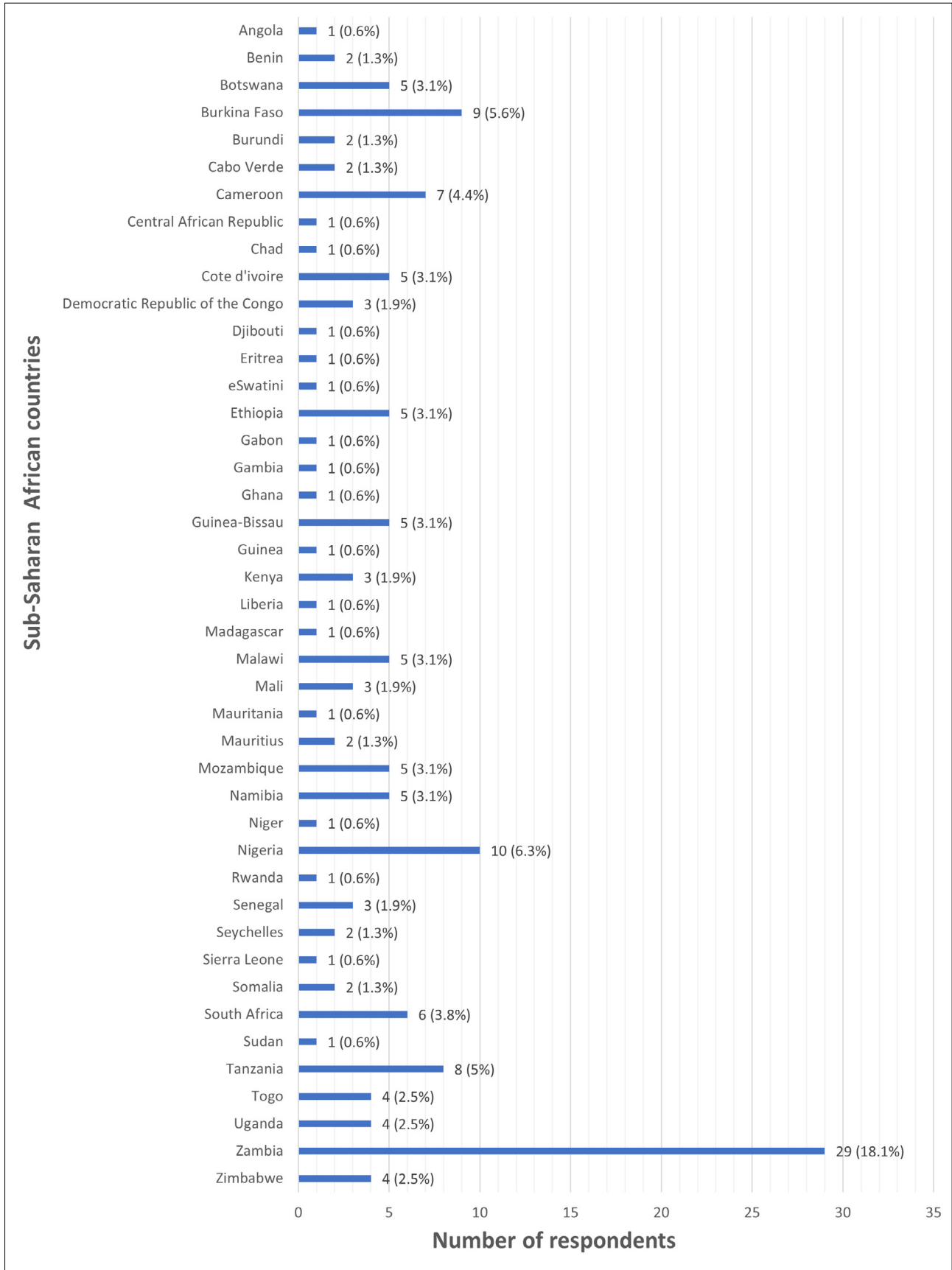


Figure 1: Number of respondents across the different sub-Saharan African countries.

Regarding the reuse of data, a great number of respondents (88.1%) perceived the lack of access to data generated by other researchers and scientists or institutions as an impediment to scientific progress, and 71.9% reported facing limitations in answering scientific questions as a result thereof (Figure 2).

Data practices

Data practices focused on the satisfaction rate of respondents' processes used in collecting, searching for and storing their data. Most respondents reported satisfaction with their institutional processes for long- and

short-term data storage (66.2% and 80%, respectively) (Figure 3). Data governance covers an important aspect of collecting and identifying data. Most respondents were satisfied with their current processes for the initial part of the research and data life cycle, which included searching for their data (76.9%) and collecting their data (82.5%). Respondents also reported satisfaction with the data tools used for the preparation of documentation (69.4%) and metadata (59.4%).

Just over a third (38.8%) of respondents indicated that most of their data were shared informally via emails and file-sharing or storage services such as Dropbox, OneDrive and Google Drive (Figure 4).

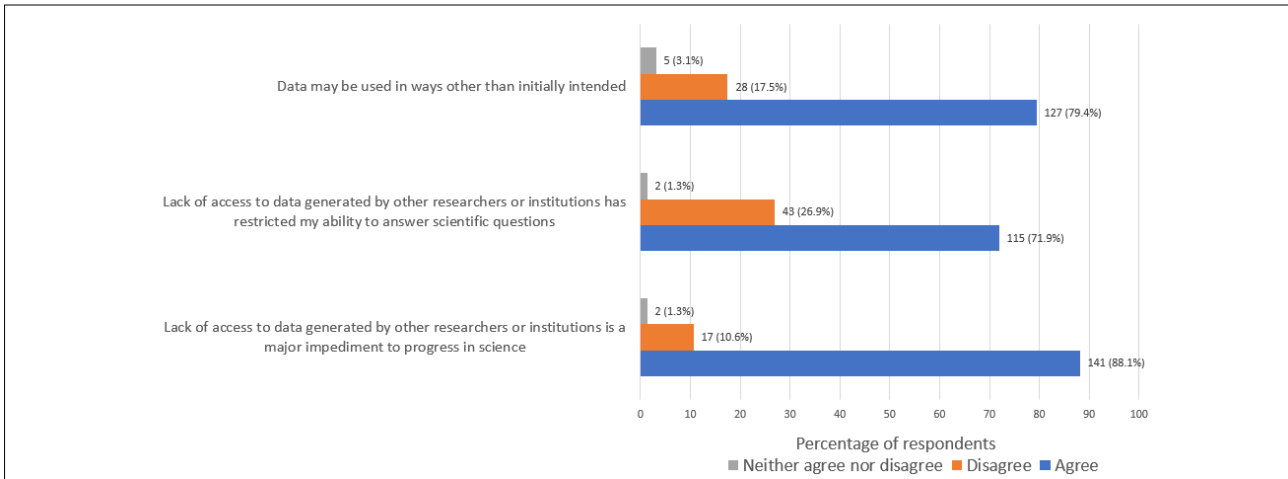


Figure 2: Perspectives on the reuse of data.

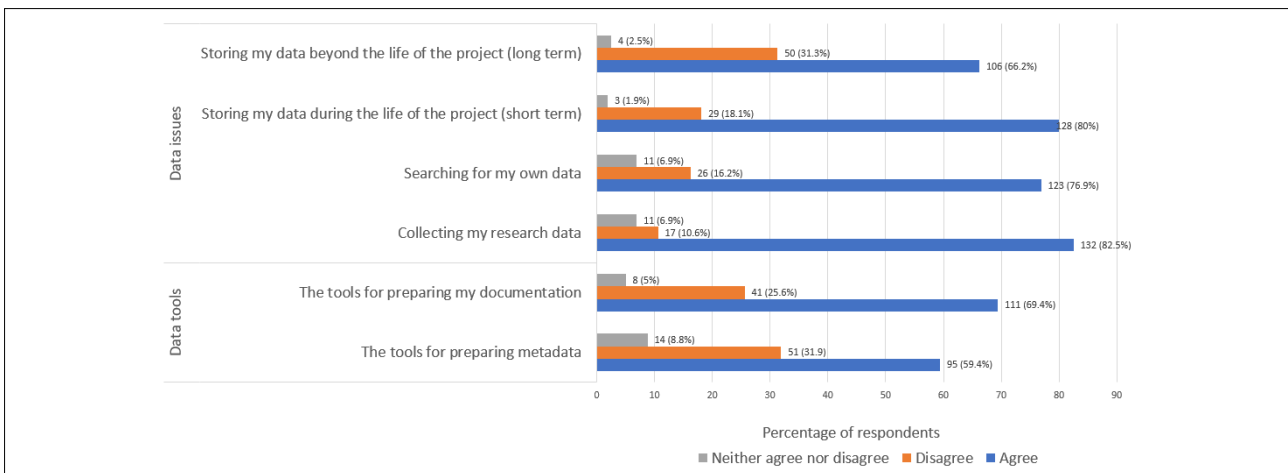


Figure 3: Satisfaction with data practices.

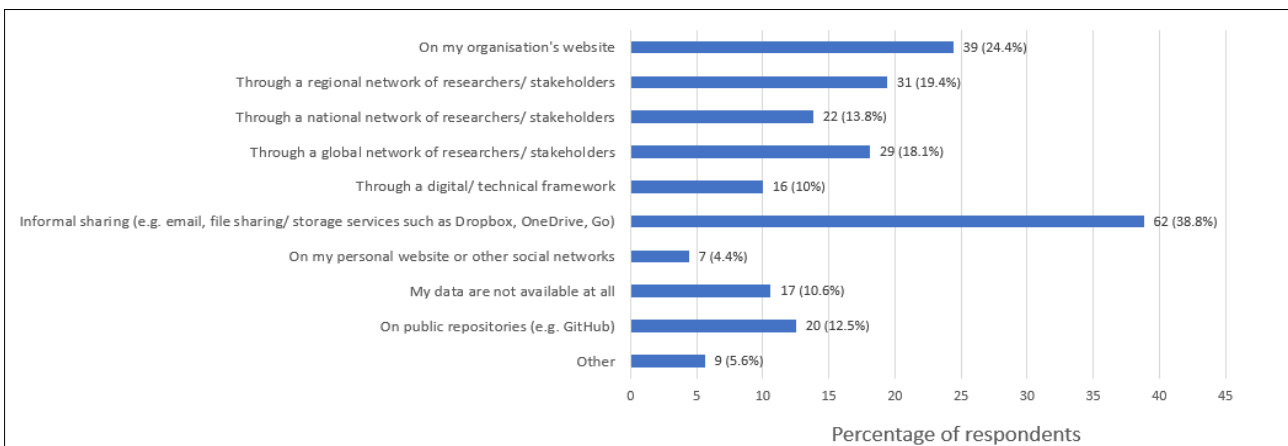


Figure 4: Data-sharing practices.

Data management support

Our survey questions on data management support assessed the satisfaction rate of respondents concerning the level of support provided by their organisations in managing their data during and beyond the research project's planning stage. Most respondents (75.7%) expressed satisfaction with the processes for managing their data, and 64.4% were satisfied with their institutional data management and/or governance plans (Table 3). The agreement rate for institutional or organisational support for data analysis during the life cycle of the project was higher over the short term (63.7%) than over the long term (53.1%).

Over half the respondents reported receiving the necessary tools and technical support for data management during (63.1%) and beyond (55%) the life cycle of the project. Just under half the respondents (40.6%) indicated receiving no training on practices for data management from their organisations or projects. Our results indicate that the provision of funds to support data management during the life cycle of a research project is higher (54.4%) than support beyond the life cycle of the research project (51.8%). These findings highlight the need for organisations or institutions to provide support or fund research data management and related infrastructure for researchers and scientists.

Data sharing

The lack of available frameworks for the mandatory sharing of data was found to be the most prominent reason (41.9%) for researchers and scientists across SSA countries to not make their data electronically available. This was followed by insufficient funds to make data available (31.9%) and not having the right to make the data available (26.9%) (Figure 5).

Almost all respondents (91.9%) agreed that they would use data sets of other researchers and scientists if these were easily accessible, and they would be willing to reciprocate (Table 4). Interestingly, most respondents (83.8%) reported a willingness to deposit some, but not all their data, into a public data repository lacking restrictions. This reported willingness to make data available increased when privacy and ethical conditions were applied (88.2%), as well as when there were conditions on governance and regulation on access (88.2%). This finding emphasises the importance of appropriate policies and governance mechanisms for data repositories to promote data sharing among scientists and researchers.⁴⁶

Furthermore, most respondents were satisfied with exchanging their data for co-authorship on publications (89.4%) and the opportunity to collaborate on projects (77.6%).

Table 3: Organisational involvement in data issues (N = 160)

	Agree n (%)	Disagree n (%)	Neither agree nor disagree n (%)
I am satisfied with the process of managing my data	121 (75.7)	30 (18.7)	9 (5.6)
I am satisfied with my institution's data management and/or governance plan	103 (64.4)	47 (29.4)	10 (3.3)
My organisation or project has a formal established process for supporting data analysis during the life of the project (short term)	102 (63.7)	50 (31.3)	8 (5)
My organisation or project has a formal established process for supporting data analysis beyond the life of the project (long term)	85 (53.1)	66 (41.3)	9 (5.6)
My organisation or project has a formal established process for storing data beyond the life of the project (long term)	87 (54.4)	65 (40.7)	8 (5)
My organisation or project has a formal established process for managing data during the life of the project (short term)	99 (61.9)	52 (32.6)	9 (5.6)
My organisation or project provides the necessary tools and technical support for data management during the life of the project (short term)	101 (63.1)	52 (32.6)	7 (4.4)
My organisation or project provides the necessary tools and technical support for data management beyond the life of the project (long term)	88 (55)	64 (40)	8 (5)
My organisation or project provides training on best practices for data management	86 (53.8)	65 (40.6)	9 (5.6)
My organisation or project provides the necessary funds to support data management during the life of a research project (short term)	87 (54.4)	65 (40.6)	8 (5)
My organisation or project provides the necessary funds to support data management beyond the life of the project (long term)	71 (44.4)	83 (51.8)	6 (3.8)

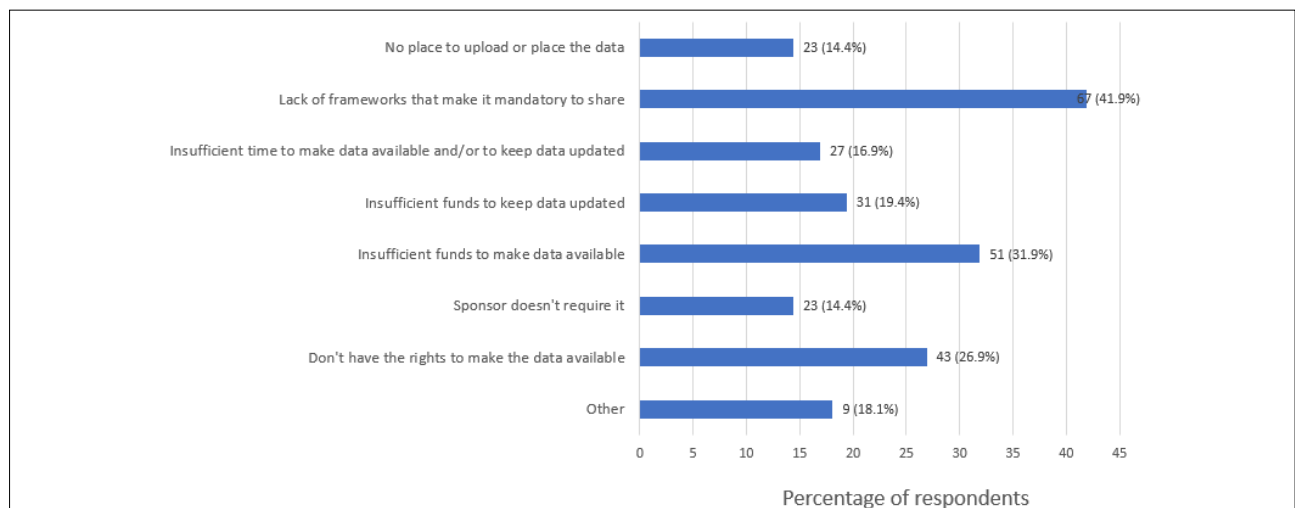


Figure 5: Reasons for not making data electronically available.

Almost all respondents (94.4%) agreed on the importance of having their data cited by other researchers and scientists. Just over half the respondents (52.5%) were satisfied with exchanging their data for royalties, while others (41.3%) agreed to exchanging their data for commercialisation purposes (Table 5). Regarding their perspectives on using and sharing others' data, the majority of respondents were satisfied (95.6%) with following ethical principles when using data from other researchers and scientists (Table 6). Most respondents were satisfied with offering co-authorship on publications in exchange for using other researchers' and scientists' data (77.5%) and the opportunity to collaborate on the project when using their data (93.1%). Over half the respondents (53.1%) disagreed with paying profits to other researchers and scientists to commercialise their data. Nearly two-thirds of the respondents (65.6%) were not keen on commercialising their data without profits (Table 7).

Data protection

Through open-ended questions, respondents were asked about their data protection practices during the sharing of data. Most respondents reported not following any particular data protection steps, whilst others followed technologically based safety measures. Of those who indicated the use of protective measures, encryption, password-protected devices and Internet security (backups and firewalls) were included.

Electronic data: secure platforms/protocols are used, data is encrypted, tools may have multilayer verification steps and PINs. Preceded by training in human subjects' protection, ethics in research.
[Country 2]

Table 4: Conditions for data sharing ($N = 160$)

	Agree <i>n</i> (%)	Disagree <i>n</i> (%)	Neither agree nor disagree <i>n</i> (%)
I would use other researchers' data sets if their data sets were easily accessible	147 (91.9)	9 (5.7)	4 (2.5)
I would equally reciprocate data sharing when data are shared with me	147 (91.9)	10 (6.2)	3 (1.9)
I would be willing to place at least some of my data into a public data repository with no restrictions	134 (83.8)	20 (12.6)	6 (3.8)
I would be willing to place all my data into a public data repository with no restrictions	88 (55)	65 (40.6)	7 (4.4)
I would be willing to make my data available if I could place privacy and ethical conditions on access	141 (88.2)	12 (7.5)	7 (4.4)
I would be more likely to make my data available if I could place conditions of governance and regulation on access	141 (88.2)	13 (8.1)	6 (3.8)
I would be willing to share data across a broad group of researchers who use data in different ways	139 (86.9)	18 (11.2)	3 (1.9)
It is important that my data are cited when used by other researchers	151 (94.4)	3 (1.9)	6 (3.8)
I am satisfied with exchanging my data knowing that secondary data will be retrieved and shared from my original data set, and then allowing those data to be shared	136 (85)	15 (9.4)	9 (5.6)
I am satisfied with exchanging my data if I know they will be used ethically	148 (92.5)	6 (3.7)	6 (3.8)
I am satisfied with exchanging my data for co-authorship on publications	139 (86.9)	16 (10.1)	5 (3.1)
I am satisfied with exchanging my data for formal acknowledgement in all disseminated work using those data	133 (83.1)	22 (13.8)	5 (3.1)
I am satisfied with exchanging my data for formal citation in all disseminated work using those data	143 (89.4)	13 (8.1)	4 (2.5)
I am satisfied with exchanging my data for the opportunity to collaborate on the project	140 (77.6)	14 (8.8)	6 (3.8)

Table 5: Conditions for data sharing related to commercialisation ($N = 160$)

	Agree <i>n</i> (%)	Disagree <i>n</i> (%)	Neither agree nor disagree <i>n</i> (%)
I am satisfied with exchanging my data for royalties	84 (52.5)	69 (43.1)	7 (4.4)
I am satisfied with exchanging my data for commercialisation purposes with profits	66 (41.3)	86 (53.8)	8 (5)
I am satisfied with exchanging my data for commercialisation purposes without profits	72 (45)	80 (50.1)	8 (5)
I am satisfied with exchanging my data for the recovery of a portion of the costs of data acquisition, retrieval or provision	88 (55.1)	62 (38.7)	10 (6.3)

Table 6: Using others' data ($N = 160$)

	Agree <i>n</i> (%)	Disagree <i>n</i> (%)	Neither agree nor disagree <i>n</i> (%)
I am satisfied with extracting secondary data from the primary data of other researchers and then share those data	132 (82.6)	25 (15.6)	3 (1.9)
I am satisfied with following ethical principles when using other researchers' data	153 (95.6)	5 (3.1)	2 (1.3)
I am satisfied with offering co-authorship on publications in exchange for using other researchers' data	124 (77.5)	32 (20.1)	4 (2.5)
I am satisfied with formally acknowledging other researchers in all disseminated work using their data	148 (92.6)	8 (5.1)	4 (2.5)
I am satisfied with formally citing other researchers in all disseminated work using their data	150 (93.8)	6 (3.8)	4 (2.5)
I am satisfied with offering other researchers the opportunity to collaborate on the project when using their data	149 (93.1)	8 (5)	3 (1.9)

Table 7: Using others' data related to commercialisation ($N = 160$)

	Agree <i>n</i> (%)	Disagree <i>n</i> (%)	Neither agree nor disagree <i>n</i> (%)
I am satisfied with paying royalties to use other researchers' data	75 (46.9)	80 (50)	5 (3.1)
I am satisfied with paying profits to other researchers to commercialise their data	66 (41.3)	85 (53.1)	9 (5.6)
I am satisfied with commercialising other researchers' data without paying them profits	48 (30)	105 (65.6)	7 (4.4)
I am satisfied with compensating a portion of the costs of data acquisition, retrieval or provision to other researchers when using their data	94 (58.8)	57 (35.6)	9 (5.6)

Confidentiality and anonymisation of information were other approaches supported by respondents.

The data should be protected confidentially to the benefit of both researchers and scientists and subjects as required in the scientific community. [Country 5]

Data management, access and sharing policies were also identified as vital in data protection.

The one requesting the data has to write a formal email or complete the form in the institution drive stating why he/she needs the data and then sign a form. Thereafter, after noting the reason why he/she needs the data, partial rights to access data can be granted. [Country 20]

DACs act as a gatekeeper for the data I generate. They review data access proposals and either grant or reject access based on the merit of the proposals. My data is accessed under the Fort Lauderdale rules of engagement, whereby there is a 2- to 3-year embargo for me to publish the data before public access is granted. [Country 22]

Respondents reported using various data agreements when sharing data to protect data ownership rights and/or the privacy or sensitivity of the data. These included memoranda of understanding (MoUs), non-disclosure agreements, DTAs and MTAs. In addition, data licensing agreements and copyright clauses were reported as important sources of data protection used. Some respondents indicated the frequent use of traditional ethics guidelines provided by their respective research ethics committees when ensuring data protection during data sharing. Whilst consent processes are vital to data sharing, another layer of protection is needed to ensure that data are adequately protected, such as pseudonymisation and encryption.⁴⁷

Consumers of data are required to sign non-disclosure agreements with confidentiality statements that they must adhere to when using protected data. [Country 7]

Respondents referred to DACs, the GDPR⁴³ and the Règlement Sanitaire International (RSI)⁴⁸ (International Health Regulation, 2005) for guidance regarding data protection. On the other hand, some respondents revealed that they do not use any data protection agreements.

Discussion

This study highlights the practices and perspectives of researchers and scientists in SSA countries regarding data sharing and data governance. Awareness of data protection policies and frameworks used in data governance was also explored. Respondents appeared relatively satisfied with their data storage processes, yet 40% indicated that their organisations or institutions did not have a formally established process for storing data beyond the life cycle of the project. There was less satisfaction with data management support; this challenge was experienced with respect to institutional support for data analysis, tools

and technical issues. Again, long-term support appeared to be lacking. This finding is similar to that of Tenopir et al.,^{6,7} who reported that short-term storage solutions provide researchers and scientists with a degree of closeness to their data during the project life cycle. We also found that more than half of the respondents were satisfied with the available tools used for documentation preparation, whilst over a third of the respondents were dissatisfied with the tools used for preparing their metadata. This correlates with the findings of another study⁷ in which respondents were also dissatisfied with the tools used for preparing their metadata. This could suggest that there is a need for adequate tools to assist SSA researchers and scientists to facilitate and enhance their use and management of data.

Although most respondents were satisfied with the process of managing their data, 40.6% disagreed that their organisation provides training on best practices for data management. This could be related to a lack of resources, chronic under-investment in universities and institutions and suboptimal research training and mentorship in SSA.^{49,50} This unmet need for training in data management has been previously documented.^{51,52} Integrating data management into research methods coursework was suggested as a possible approach for encouraging best practices amongst researchers and scientists.⁵³ With the growing adoption globally of big data, SSA researchers and scientists must be trained to harness big complex data sets to find solutions to scientific problems. Funding was another issue raised by respondents, with more than half indicating that their organisations did not provide the necessary funds to support data management beyond the life cycle of the project. These findings are similar to those of Tenopir et al.⁶ in which 59% of respondents indicated a lack of financial support for data management beyond the life cycle of the project. It will be crucial for organisations and institutions to invest and have sustainable funding for data management services in SSA. This has also been reported in other LMICs where the emphasis is on the importance of investment in data management.⁵⁴

Open science and the sharing of data are essential for the advancement of science, and are seen as an important part of economic growth in Africa, which is burdened with dual public health and economic crises.^{55,56} Furthermore, from an ethical perspective, data sharing is a significant way to recognise the altruism and generosity of participants (for example, those from clinical trials) because it increases the utility of the data they provide and thus the value of their contribution.⁵⁷ It was therefore important to explore the perspectives and practices of SSA researchers and scientists on data sharing. The majority of respondents reported that they had already shared their data. Lack of governance frameworks that make it mandatory to share data (41.9%) was one of the main reasons for not making data electronically available, followed by insufficient funds (31.9%). These reasons have also been reported as barriers to data sharing in LMICs, in African research institutions, as well as in institutions in Jordan.^{54,58} In the face of insufficient funding, Okafor et al.⁵⁹ emphasised the importance of funding to institutionalise open science in Africa. The fact that open science for Africa is seen as a potential route to increased funding opportunities is particularly noteworthy. Researchers and scientists in Africa can gain visibility and funding from a broader group of potential funders by openly sharing their research findings.

Most respondents had positive views of data sharing, but 40.6% indicated a need to restrict all their data when placed in a public data repository.

This could be because there are either ethical issues or concerns about commercialisation. Most of the respondents also agreed to sharing their data, provided that the condition for sharing is to receive proper citation credit, co-authorship and an opportunity to collaborate. The respondents did not differ much in their perspectives on using others' data. These findings support previous studies, where citation credit, co-authorship and an opportunity to collaborate were amongst the conditions and motivations for sharing and using others' data.^{6,34,60,61} In contrast, some studies reported that African counterparts seem to be largely motivated by altruistic means for data sharing, such as emphasising the public benefit or the good of sharing knowledge and data.^{34,62} Nevertheless, the findings could suggest that African countries are gradually becoming familiar with the significance of data sharing and its impact on their researchers' and scientists' careers, which is different from several years ago.⁶³ It would be useful for institutions or organisations to encourage data citation as a central data-sharing practice, and for researchers and scientists to be given co-authorship and collaboration in exchange for data sharing, taking authorship requirements into account.

It has been suggested that, in order to be eligible for co-authorship, a person must have made a significant contribution to the work (i.e. original acquisition, quality control and data curation) and be accountable for all aspects of the accuracy and integrity of the data provided, as well as ensure that the available data set adheres to the FAIR Guiding Principles.^{12,64} However, some studies have argued that co-authorship in exchange for data is a rather contentious issue, as it could be perceived as being potentially unethical.⁶⁵ In addition, Hood and Sutherland⁶⁶ further assert that author-type metrics, which are the gold standard for measuring scientific progression and success, are detrimental to scientific development. Hence, there is a need to develop different reward systems, whereby the output of data sets and data-index citations are collectively viewed as a measure of researcher growth and progression, instead of over-reliance on the number of publications or data-index citations. This shift in the reward system will greatly facilitate data sharing, especially in LMICs.⁶⁶

Interestingly, respondents had different perspectives on the commercialisation of shared data, with half not agreeing to exchange others' data for commercialisation purposes. These findings differ from those of a Malaysian study⁶⁷ which found that 90% of the surveyed researchers and scientists were interested in commercialising their research. Our respondents' views may have differed because some work with data (i.e. genetic information) that present significant dilemmas in the context of privacy and consent.⁶⁸ Most respondents indicated that they do not use any data protection steps when sharing data other than using technologically based safety measures (e.g. password protection or encryption methods). This is concerning as it suggests that researchers and scientists are still making use of suboptimal or mediocre data practices, placing their data at risk for misuse or theft, amongst other concerns.⁷ There is a need to encourage researchers and scientists in the African context to prioritise good data practices by storing and sharing data in repositories.⁷ This can be accomplished by changing researchers' negative perceptions around repositories by educating them on the standards and criteria of data repositories (increased security), as well as the benefits, such as adequately prepared metadata and the discoverability of the data.⁵⁷ Europe has adopted a common legal, governance, data quality and operability framework to facilitate access to and reuse of health data.⁶⁹

Another aspect of our findings was that respondents mentioned various data agreements they used when sharing data. These included DTAs, MTAs and MoUs. However, some of the respondents mentioned that they lacked such agreements. A common suggestion to improve these challenges included the development of DACs. Such committees balance issues of data ownership and foster data governance through their ability to approve or disapprove data access requests.^{44,69} This poses a question as to how SSA researchers and scientists share their data without the existence of policies or frameworks in their institutions or organisations. It is important to note that the lack of governance frameworks was the top reason respondents did not share their data. This has also been reported in the literature, where the lack of policy

and guideline frameworks at institutional and national levels is one of the reasons for African researchers and scientists not sharing their data.³⁴ About 18 SSA countries (including South Africa and Kenya) have a comprehensive data protection law that is currently in effect.⁷⁰ Considering the current advancements in digital technologies, SSA countries must implement data protection policies and frameworks that are a contextual fit, as this could provide assurances and confidence amongst researchers and scientists that measures are in place to secure their data sets during the sharing or transfer of data.

Furthermore, having policies or frameworks in place could encourage researchers and scientists in SSA to make their data electronically available. Despite the benefits of data sharing promoted by funders and journals, the volume of shared data remains low.⁷¹ Buy-in from and support for institutions or organisations to establish data-sharing policies that specify aims and data request procedures may be required. Cheah⁷¹ advised that the aims should be aligned with the institutional or organisational aims, as this would help researchers and scientists maximise the use of their data for primary and secondary analyses. In addition, having a data-sharing policy could put an institution or organisation in a better position when applying for funding and submitting manuscripts for publication.⁷¹ Nevertheless, there is a need for engagement or collaboration between researchers and scientists, their funders and institutions or organisations to find creative solutions that could enhance responsible and sustainable data governance.

Overall, the survey found that researchers and scientists were optimistic about data sharing, storage, data management support and reuse. Many researchers and scientists across SSA are using various types of data agreements and security measures during data sharing, whilst other researchers lack such tools, approaches and data protection policies and frameworks that promote safe data sharing. The study findings have been interpreted and discussed in light of the current available literature. When compared to the findings of previous global studies^{6,7,34,54,58,60,61}, our findings were similar and comparable in terms of data practices, data management support and data-sharing practices. However, some differences emerged in the perspectives of data sharing for commercialisation purposes.⁶⁷

Study limitations

The study is not without its limitations, which should be considered when interpreting the findings. There was a predominance of respondents from Zambia, Nigeria, Burkina Faso, Tanzania, Cameroon and South Africa in comparison with other SSA countries. This could be because email access was better in these countries. A consistent and salient finding across the comparison of responses from these six SSA countries with the highest number of responses was that most views were aligned – apart from some recurrent variations regarding organisational involvement in data activities and conditions of fair exchange. Based on previous experience with conducting research in SSA, obtaining a response to surveys is challenging, so we aimed to get a minimum of one response per country. Access to the Internet and email is inequitable in various settings in Africa.⁷² It is with significant effort that we were able to elicit responses from 43/49 SSA countries. The sample size was relatively small and may not represent all researchers and scientists in SSA countries. Future studies could include a larger sample across SSA countries so that the findings could be generalisable to the overall research population. However, data collection would take significantly longer than 4 months, given the challenges with responsiveness and Internet or email access that exist on the continent. Those respondents that did not complete the survey might have felt that the survey was too long. Despite these limitations, this study has provided a broad overview of important practices and perspectives on data governance amongst a sample of researchers and scientists in SSA, and has informed the qualitative phase of our study, in which we conducted in-depth interviews.

Conclusion

Data sharing is generally recognised as a public good that increases the diversity of research data. Most respondents demonstrated a positive attitude towards data sharing and were willing to share at least some of



their data, conditional upon robust governance with certain restrictions. In addition to funding, there is a need for the institutional support of data management, robust data protection legislation and appropriate policies to guide and promote data sharing in SSA countries. Given that DTAs vary between projects and countries, having standardised templates for DTAs and data use agreements would expedite sharing agreements between research collaborators. This will enable researchers and scientists, their funders, journals and institutions to collaborate and promote sustainable data sharing on the continent. In this context, sustainable data sharing includes providing ethical incentive structures for researchers and scientists who are willing to share their data, as well as tangible infrastructure to facilitate such sharing. Capacity development in data governance for researchers and scientists is sorely needed – and relevant knowledge transfer between SSA countries should be facilitated. Perceived and actual risks of commercialisation require further exploration.

Acknowledgements

Research reported in this publication was supported by the US National Institute of Mental Health of the US National Institutes of Health under award number U01MH127704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the respondents for their invaluable time in participating in the study. We also thank Prof. Juan Klopper for his guidance in developing the survey instrument. We furthermore acknowledge the following individuals for their guidance during the early stages of the project: Aneeka Domingo, Derrick Kourie, Gonasagrie Nair and Emmanuel Obasa.

Competing interests

We have no competing interests to declare.

Authors' contributions

S.M.K.: Made substantial contributions to the analysis and interpretation of data for the work; drafted/revised the work critically for important intellectual content. N.C.: Made substantial contributions to the design of the work, acquisition and interpretation of data for the work; drafted/revised the work critically for important intellectual content. K.R.: Drafted/revised the work critically for important intellectual content. B.W.W.: Drafted/revised the work critically for important intellectual content. Q.B.: Made substantial contributions to the analysis and interpretation of data for the work; drafted/revised the work critically for important intellectual content. T.M.E.: Made substantial contributions to the analysis of data for the work; drafted/revised the work critically for important intellectual content. K.M.: Made substantial contributions to the conception of the work; drafted/revised the work critically for important intellectual content. All authors approved the final version and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Guo H, Hackmann H, Gong K. Big data in support of the Sustainable Development Goals: A celebration of the establishment of the International Research Center of Big Data for Sustainable Development Goals (CBAS). *Big Earth Data*. 2021;5(3):259–262. <https://doi.org/10.1080/20964471.2021.1962621>
2. Hassani H, Huang X, MacFeely S, Entezarian MR. Big Data and the United Nations Sustainable Development Goals (UN SDGs) at a glance. *Big Data Cogn Comput*. 2021;5(3), Art. #28. <https://doi.org/10.3390/bdcc5030028>
3. De Mauro A, Greco M, Grimaldi M. A formal definition of Big Data based on its essential features. *Libr Rev*. 2016;65(3):122–135. <https://doi.org/10.1108/LR-06-2015-0061>
4. De Cnudde S, Martens D. Loyal to your city? A data mining analysis of a public service loyalty program. *Decis Support Syst*. 2015;73:74–84. <https://doi.org/10.1016/j.dss.2015.03.004>
5. Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. *SAGE Open Med*. 2020;8, Art. #2050312120934839. <https://doi.org/10.1177/2050312120934839>
6. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data sharing by scientists: Practices and perceptions. *PLoS ONE*. 2011;6(6):e21101. <https://doi.org/10.1371/journal.pone.0021101>
7. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS ONE*. 2020;15(3), e0229003. <https://doi.org/10.1371/journal.pone.0229003>
8. Pasquetto IV., Randles BM, Borgman CL. On the reuse of scientific data. *Data Sci J*. 2017;16(8). <https://doi.org/10.5334/dsj-2017-008>
9. Obiora OL, Olivier B, Shead DA, Withers A. Data sharing practices of health researchers in Africa: A scoping review protocol. *JBI Evid Synth*. 2022;20(2):681–688. <https://doi.org/10.11124/JBIES-20-00502>
10. Munafò MR, Nosek BA, Bishop DVM, Button, KS, Chambers CD, Percie Du Sert N, et al. A manifesto for reproducible science. *Nat Hum Behav*. 2017;1(1):1–9. <https://doi.org/10.1038/s41562-016-0021>
11. Klein O, Hardwicke TE, Aust F, Breuer J, Danielsson H, Mohr AH, et al. A practical guide for transparency in psychological science. *Collabra Psychol*. 2018;4(1). <https://doi.org/10.1525/collabra.158>
12. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):1–9. <https://doi.org/10.1038/sdata.2016.18>
13. Dienlin T, Johannes N, Bowman ND, Masur PK, Engesser S, Kümpel AS, et al. An agenda for open science in communication. *J Commun*. 2021;71(1):1–26. <https://doi.org/10.1093/joc/jqz052>
14. Kaye J, Terry SF, Juengst E, Coy S, Harris JR, Chalmers D, et al. Including all voices in international data-sharing governance. *Hum Genomics*. 2018;12(1):1–6. <https://doi.org/10.1186/s40246-018-0143-9>
15. MacMillan D. Data sharing and discovery: What librarians need to know. *J Acad Librariansh*. 2014;40(5):541–549. <https://doi.org/10.1016/j.acalib.2014.06.011>
16. Global Alliance for Genomics and Health. Framework for involving and engaging participants, patients and publics in genomics research and health implementation [document on the Internet]. c2021 [cited 2023 Apr 13]. Available from: https://www.ga4gh.org/wp-content/uploads/GA4GH_Engagement-policy_V1.0_July2021-1.pdf
17. Human Genome Project. The Bermuda Principles [webpage on the Internet]. c1996 [cited 2023 Apr 13]. Available from: https://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml
18. Wellcome Trust. Sharing data from large-scale biological research projects: A system of tripartite responsibility [document on the Internet]. c2003 [cited 2023 Apr 13]. Available from: <https://www.genome.gov/Pages/Research/WellcomeReport0303.pdf>
19. Amann RI, Baichoo S, Blencowe BJ, Bork P, Borodovsky M, Brooksbank C, et al. Toward unrestricted use of public genomic data. *Science*. 2019;363(6425):350–352. <https://doi.org/10.1126/science.aaw1280>
20. Gomes DGE, Pottier P, Crystal-Ornelas R, Hudgins EJ, Foroughirad V, Sánchez-Reyes LL, et al. Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc. R. Soc. B*. 2022;289(1987), Art. #20221113. <https://doi.org/10.1098/rspb.2022.1113>
21. Bezuidenhout L. To share or not to share: Incentivizing data sharing in life science communities. *Dev World Bioeth*. 2019;19(1):18–24. <https://doi.org/10.1111/dewb.12183>
22. Donaldson DR, Koepke JW. A focus groups study on data sharing and research data management. *Sci Data*. 2022;9(1), Art. #345. <https://doi.org/10.1038/s41597-022-01428-w>
23. Kim J, Hwang H, Jung Y, Cho S-N, Seo T-S. Data sharing attitudes and practices of researchers in Korean government research institutes: A survey-based descriptive study. *Sci Ed*. 2023;10(1), Art. #71. <https://doi.org/10.6087/kcse.299>
24. Zuiderwijk A, Shinde R, Jeng W. What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. *PLoS ONE*. 2020;15(9), e0239283. <https://doi.org/10.1371/journal.pone.0239283>



25. Polanin JR, Terzian M. A data-sharing agreement helps to increase researchers' willingness to share primary data: results from a randomized controlled trial. *J Clin Epidemiol*. 2019;106:60–69. <https://doi.org/10.1016/j.jclinepi.2018.10.006>
26. Chalmers D, Nicol D, Nicolás P, Zeps N. A role for research ethics committees in exchanges of human biospecimens through material transfer agreements. *J Bioeth Inq*. 2014;11:301–306. <https://doi.org/10.1007/s11673-014-9552-1>
27. Schaeffer V. The use of material transfer agreements in academia: A threat to open science or a cooperation tool? *Res Policy*. 2019;48(9), Art. #103824. <https://doi.org/10.1016/j.respol.2019.103824>
28. Pujol Priego L, Wareham J, Romasanta AKS. The puzzle of sharing scientific data. *Ind Innov*. 2022;29(2):219–250. <https://doi.org/10.1080/13662716.2022.2033178>
29. Berghmans S, Cousijn H, Deakin G, Meijer I, Mulligan A, Plume A, et al. Open data: The researcher perspective. Leiden University Centre for Science and Technology Studies, and Elsevier [document on the Internet]. c2017 [cited 2023 Apr 13]. Available from https://www.elsevier.com/_data/assets/pdf_file/0004/281920/Open-data-report.pdf
30. Abebe R, Aruleba K, Birhane A, Kingsley S, Obaido G, Remy SL, et al. Narratives and counternarratives on data sharing in Africa. *FAcT*. 2021;329–341. <https://doi.org/10.1145/3442188.3445897>
31. Bangani S, Moyo M. Data sharing practices among researchers at South African universities. *Data Sci J*. 2019;18, Art. #28. <https://doi.org/10.5334/dsj-2019-028>
32. Chinakidzwa M, Phiri M, Mashingaidze M. Research data sharing and re-use practices, perceptions and behaviours: Evidence from Zimbabwe. *J Afr Educ*. 2022;3(1), Art. #4. <https://doi.org/10.31920/2633-2930/2022/v3n1a4>
33. Evertsz N, Bull S, Pratt B. What constitutes equitable data sharing in global health research? A scoping review of the literature on low-income and middle-income country stakeholders' perspectives. *BMJ Glob Health*. 2023;8(3), e010157. <https://doi.org/10.1136/bmjgh-2022-010157>
34. Skelly L, Chiware ERT. African researchers do not think differently about open data. *Front Res Metr Anal*. 2022;7, Art. #950212. <https://doi.org/10.3389/frma.2022.950212>
35. Waitira N, Mutinda B, Cheah PY. Data management and sharing policy: The first step towards promoting data sharing. *BMC Med*. 2019;17(1), Art. #80. <https://doi.org/10.1186/s12916-019-1315-8>
36. Brand D, Singh JA, McKay AGN, Cengiz N, Moodley K. Data sharing governance in sub-Saharan Africa during public health emergencies: Gaps and guidance. *S Afr J Sci*. 2022;118(11–12), Art. 13892. <https://doi.org/10.17159/sajs.2022/13892>
37. Mittelstadt BD, Floridi L. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Sci Eng Ethics*. 2016;22(2):303–341. <https://doi.org/10.1007/s11948-015-9652-2>
38. Organisation for Economic Co-operation and Development (OECD). Why data governance matters. Data governance [webpage on the Internet]. c2022 [cited 2023 Apr 14]. Available from: <https://search.oecd.org/digital/data-governance/>
39. Akintola SO. Legal implications of data sharing in biobanking research in low-income settings: The Nigerian experience. *S Afr J Bioeth Law*. 2018;11(1):15–19. <https://doi.org/10.7196/SAJBL.2018.v11i1.00601>
40. Townsend B. The lawful sharing of health research data in South Africa and beyond. *Inf Commun Technol Law*. 2022;31(1):17–34. <https://doi.org/10.1080/13600834.2021.1918905>
41. Bezuidenhout L, Chakouya E. Hidden concerns of sharing research data by low/middle-income country scientists. *Global Bioethics*. 2018;29(1):39–54. <https://doi.org/10.1080/11287462.2018.1441780>
42. Parliament of the Republic of South Africa. Protection of Personal Information Act. Republic of South Africa [document on the Internet]. c2013 [cited 2023 Apr 14]. Available from: <https://www.gov.za/documents/protection-personal-information-act>
43. The European Parliament and the Council. The General Data Protection Regulation (GDPR). The European Union [webpage on the Internet]. c2016 [cited 2023 Apr 14]. Available from: [https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html#:~:text=Regulation%20\(EU\)%202016%2F679%20of%20the%20European%20Parliament%20and,\(OJ%20L%202019%2C%204.5](https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html#:~:text=Regulation%20(EU)%202016%2F679%20of%20the%20European%20Parliament%20and,(OJ%20L%202019%2C%204.5)
44. Piasecki J, Cheah PY. Ownership of individual-level health data, data sharing, and data governance. *BMC Med Ethics*. 2022;23(1):1–9. <https://doi.org/10.1186/s12910-022-00848-y>
45. Austin Z, Sutton J. Qualitative research: Getting started. *Can J Hosp Pharm*. 2014;67(6):436–440. <https://doi.org/10.4212/cjhp.v67i6.1406>
46. Devriendt T, Shabani M, Borry P. Policies to regulate data sharing of cohorts via data infrastructures: An interview study with funding agencies. *Int J Med Inform*. 2022;168, Art. #104900. <https://doi.org/10.1016/j.ijmedinf.2022.104900>
47. Viberg Johansson J, Bentzen HB, Mascalconi D. What ethical approaches are used by scientists when sharing health data? An interview study. *BMC Med Ethics*. 2022;23(1):41. <https://doi.org/10.1186/s12910-022-00779-8>
48. World Health Organization. Règlement Sanitaire International [document on the Internet]. c2005 [cited 2023 Apr 14]. Available from: <https://apps.who.int/iris/bitstream/handle/10665/246107/9789241580496-eng.pdf>
49. Izugbara CO, Kabiru CW, Amendah D, Dimbuene ZT, Donfouet HP, Atake EH, et al. 'It takes more than a fellowship program': Reflections on capacity strengthening for health systems research in sub-Saharan Africa. *BMC Health Serv Res*. 2017;17:1–5. <https://doi.org/10.1186/s12913-017-2638-9>
50. Zuk P, Sanchez CE, Kostick K, Torgerson L, Muñoz KA, Hsu R, Kalwani L, et al. Researcher perspectives on data sharing in deep brain stimulation. *Front Hum Neurosci*. 2020;14, Art. #578687. <https://doi.org/10.3389/fnhum.2020.578687>
51. Barone L, Williams J, Micklos D. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Comput Biol*. 2017;13(10), e1005755. <https://doi.org/10.1371/journal.pcbi.1005755>
52. Tenopir C, Allard S, Sinha P, Pollock D, Newman J, Dalton E, et al. Data management education from the perspective of science educators. *Int J Digit Curation*. 2016;11(1):232–251. <https://doi.org/10.2218/ijdc.v11i1.389>
53. Borghi JA, Van Gulick AE. Data management and sharing: Practices and perceptions of psychology researchers. *PLoS ONE*. 2021;16(5), e0252047. <https://doi.org/10.1371/journal.pone.0252047>
54. Kaewkungwal J, Adams P, Sattabongkot J, Lie RK, Wendler D. Issues and challenges associated with data-sharing in LMICs: Perspectives of researchers in Thailand. *Am J Trop Med Hyg*. 2020;103(1):528–536. <https://doi.org/10.4269/ajtmh.19-0651>
55. Ramsay M. African genomic data sharing and the struggle for equitable benefit. *Patterns*. 2022;3(1), Art. #100412. <https://doi.org/10.1016/j.patter.2021.100412>
56. Organisation for Economic Co-operation and Development. COVID-19 and Africa: Socio-economic implications and policy responses. OECD Policy Responses to Coronavirus (COVID-19) [webpage on the Internet]. c2020 [cited 2022 Oct 18]. Available from: <https://www.oecd.org/coronavirus/policy-responses/covid-19-and-africa-socio-economic-implications-and-policy-responses-96e1b282/>
57. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open* 2017;7(12), e018647. <https://doi.org/10.1136/bmjopen-2017-018647>
58. Al-Ebbini L, Khabour OF, Alzoubi KH, Alkaraki AK. Biomedical data sharing among researchers: A study from Jordan. *J Multidiscip Healthc*. 2020;13:1669–1676. <https://doi.org/10.2147/JMDH.S284294>
59. Okafor IA, Mbagwu SI, Chia T, Hasim Z, Udokanma EE, Chandran K. Institutionalizing open science in Africa: Limitations and prospects. *Front Res Metr Anal*. 2022;7, Art. #855198 <https://doi.org/10.3389/frma.2022.855198>
60. Mwangi KW, Mainye N, Ouso DO, Esoh K, Muraya AW, Mwangi CK, et al. Open science in Kenya: Where are we? *Front Res Metr Anal*. 2021;6, Art. #669675. <https://doi.org/10.3389/frma.2021.669675>
61. Hrynaskiewicz I, Harney J, Cadwallader L. A survey of researchers' needs and priorities for data sharing. *Data Sci J*. 2021;20(1). <https://doi.org/10.5334/dsj-2021-031>



62. Jao I, Kombe F, Mwalukore S, Bull S, Parker M, Kamuya D, et al. Involving research stakeholders in developing policy on sharing public health research data in Kenya: Views on fair process for informed consent, access oversight, and community engagement. *J Empir Res Hum Res Ethics* 2015;10(3):264–277. <https://doi.org/10.1177/1556264615592385>
63. National Academies of Sciences, Engineering, and Medicine. Sharing research data to improve public health in Africa: A workshop summary. O'Connell ME, Plewes TJ, rapporteurs. Committee on Population, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press; 2015. <https://doi.org/10.17226/21801>
64. Bierer BE, Crosas M, Pierce HH. Data authorship as an incentive to data sharing. *N Engl J Med*. 2017;376(17):1684–1687. <https://doi.org/10.1056/NEJMs1616595>
65. Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *J Clin Epidemiol*. 2022;150:33–41. <https://doi.org/10.1016/j.jclinepi.2022.05.019>
66. Hood ASC, Sutherland WJ. The data-index: An author-level metric that values impactful data and incentivizes data sharing. *Ecol Evol*. 2021;11(21):14344–14350. <https://doi.org/10.1002/ece3.8126>
67. Abdul Aziz NA, Ismail N, Hartono A. Strategies to enhance commercialisation activity: Researcher perspective. *Knowledge Management International Conference, Universitas Islam Indonesia*; 2021. p. 94–101.
68. Botes M, Slabbert MN, Olckers A. Data commercialisation in the South African health care context. *Potchefstroom Electron Law J*. 2021;24(1). <https://doi.org/10.17159/1727-3781/2021/v24i0a8577>
69. Shabani M. Will the European health data space change data sharing rules? *Science*. 2022;375(6587):1357–1359. <https://doi.org/10.1126/science.abn4874>
70. Bryant J. Africa in the information age: Challenges, opportunities, and strategies for data protection and digital rights. *Stan Tech L Rev*. 2020;24:389–439.
71. Cheah PY. Institutions must state policy on data sharing. *Nature*. 2019;565(7739):294. <https://doi.org/10.1038/d41586-019-00118-9>
72. Moodley K, Kabanda SM, Soldaat L, Kleinsmidt A, Obasa AE, Kling S. Clinical ethics committees in Africa: Lost in the shadow of RECs/IRBs? *BMC Med Ethics*. 2020;21(1):1–10. <https://doi.org/10.1186/s12910-020-00559-2>



Check for updates

AUTHORS:

Sunday O. Oladejo¹
Liam R. Watson^{1,2}
Bruce W. Watson¹
Kanshukan Rajaratnam¹
Maritha J. Kotze³
Douglas B. Kell^{4,5,6}
Etheresia Pretorius^{4,6}

AFFILIATIONS:

¹School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa

²David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

³Division of Chemical Pathology, Department of Pathology, National Health Laboratory Service, Tygerberg Hospital & Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

⁴Department of Biochemistry and Systems Biology, Faculty of Health and Life Sciences, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK

⁵The Novo Nordisk Foundation Centre for Biosustainability, Technical University of Denmark, Lyngby, Denmark

⁶Department 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



The Author(s). Published under a Creative Commons Attribution Licence.

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}

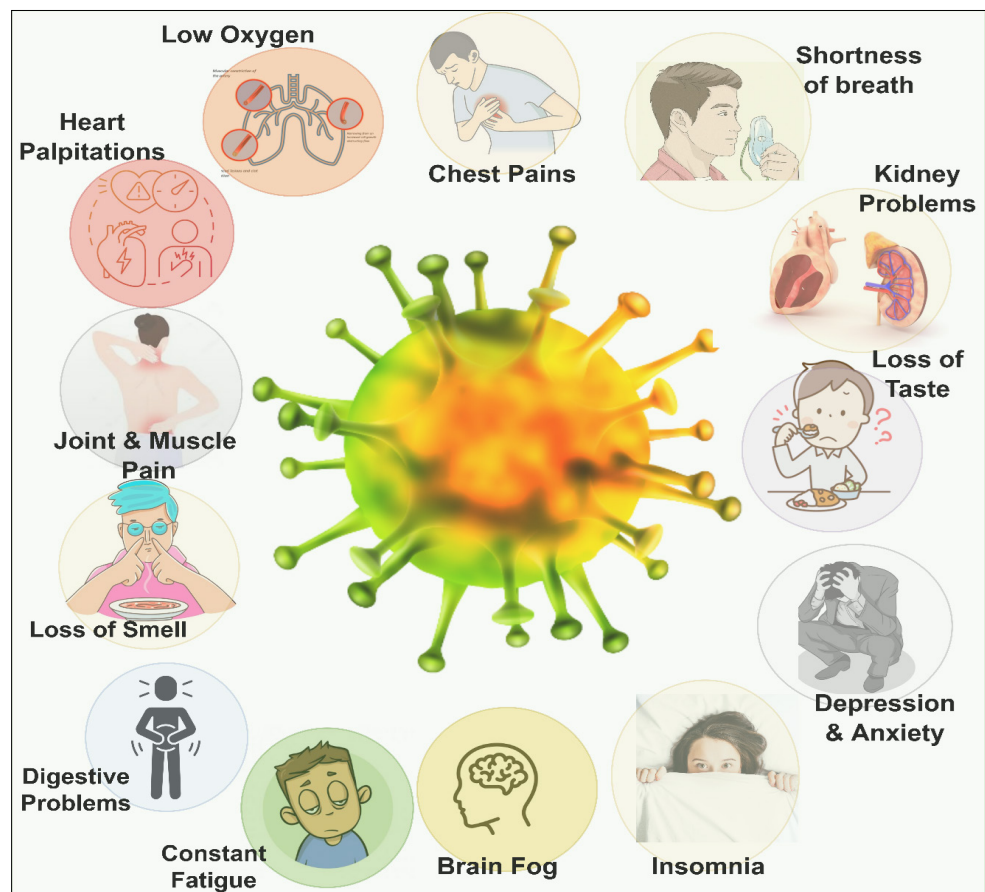


Figure 1: Illustration of the common Long COVID symptoms and complications reported in the literature.

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

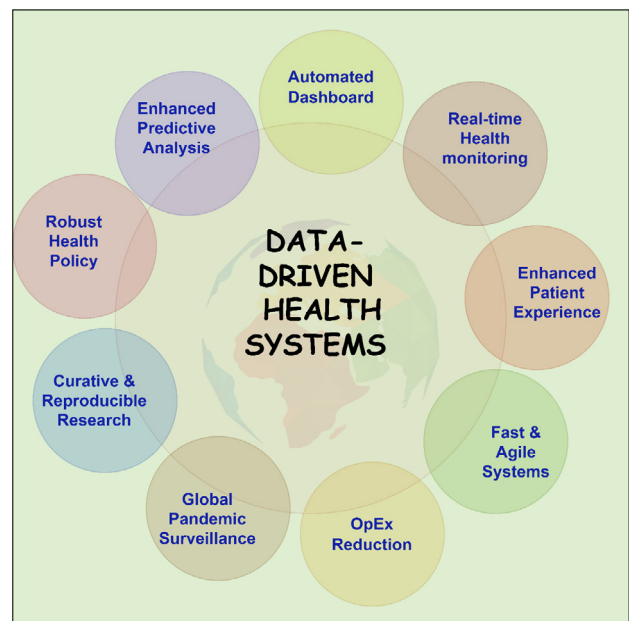


Figure 2: Benefits of adopting a data-driven framework for Long COVID management and healthcare systems in general.

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	–

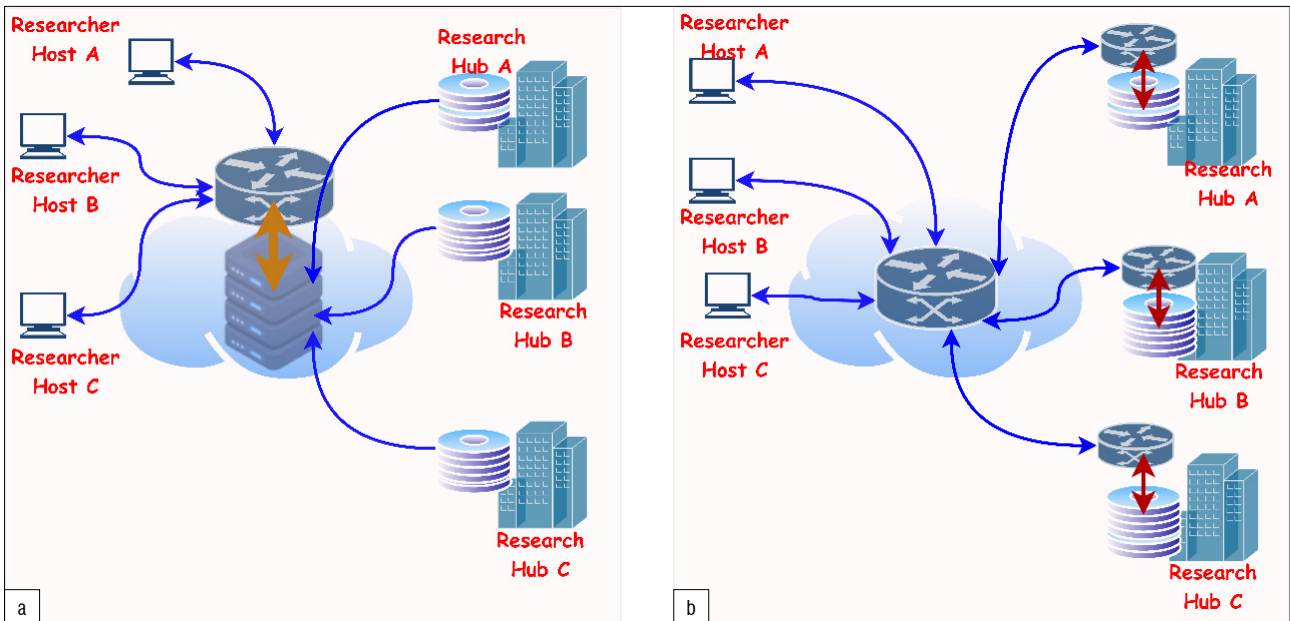


Figure 3: Illustration of the two main data-sharing strategies: (a) the centralised architecture and (b) the federated architecture.

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 discussion¹² 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 ethics⁴³, as illustrated in Figure 4. 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.

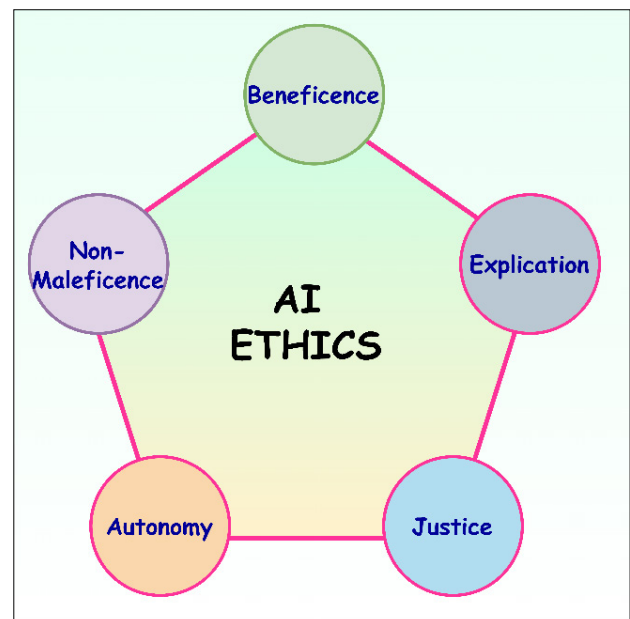


Figure 4: Pillars of artificial intelligence (AI) ethics.⁴³

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 famously 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 technologies⁶⁶ 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, international, 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

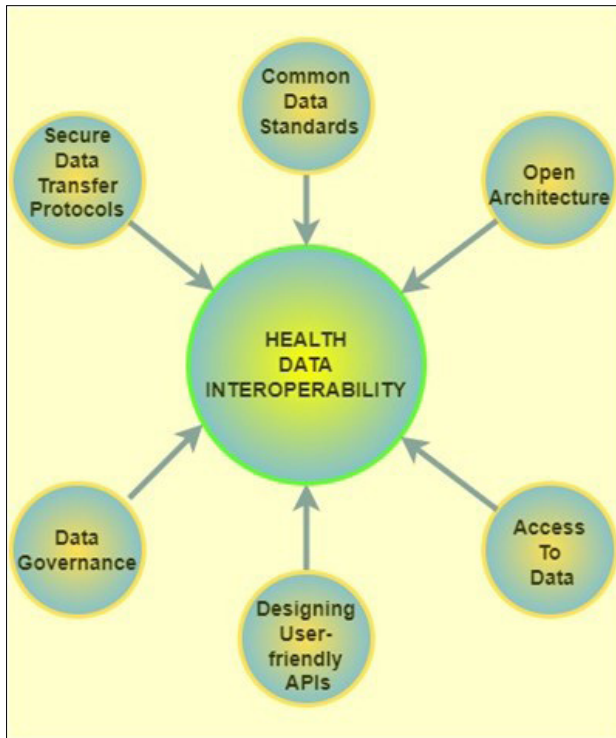


Figure 5: Key steps that contribute to the realisation of interoperability of health data.

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 adopters⁸², 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, Ravaut 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](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. Galaitis 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](https://doi.org/10.1016/S2589-7500(22)00048-6)
23. Vinod DN, Prabakaran 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, Menjivar 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/covid19schoolsinfectedsurveyquestionnairedataengland>
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/N5110C>
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%20.0%20million%20people,2022%20\(see%20Figure%201\)](https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/7july2022#:~:text=An%20estimated%20.0%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, Cows 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](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/10.7861/1h.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, Azimae 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 (UHP): 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.

**AUTHOR:**Marietjie Botes^{1*} **AFFILIATION:**¹SnT Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, Luxembourg, Luxembourg

*Current: Centre for Medical Ethics and Law, WHO Bioethics Collaborating Centre, Department of Medicine, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

CORRESPONDENCE TO:

Marietjie Botes

EMAIL:

wmbotes@sun.ac.za

DATES:**Received:** 24 Oct. 2022**Revised:** 30 Mar. 2023**Accepted:** 10 Apr. 2023**Published:** 30 May 2023**HOW TO CITE:**Botes M. Regulating scientific and technological uncertainty: The precautionary principle in the context of human genomics and AI. *S Afr J Sci.* 2023;119(5/6), Art. #15037. <https://doi.org/10.17159/sajs.2023/15037>**ARTICLE INCLUDES:**

- Peer review
- Supplementary material

DATA AVAILABILITY:

- Open data set
- All data included
- On request from author(s)
- Not available
- Not applicable

EDITOR:

Floretta Boonzaier

KEYWORDS:

precautionary principle, risk-based approach, human genomics, AI, regulation

FUNDING:

Luxembourg National Research Fund (IS/14717072), European Union's Horizon 2020 Innovative Training Networks (ID 956562), REMEDIS Project (INTER/FNRS/21/16554939/ REMEDIS)



Regulating scientific and technological uncertainty: The precautionary principle in the context of human genomics and AI

Considered in isolation, the ethical and societal challenges posed by genomics and artificial intelligence (AI) are profound and include issues relating to autonomy, privacy, equality, bias, discrimination, and the abuse of power, amongst others. When these two technologies are combined, the ethical, legal and societal issues increase substantially, become much more complex, and can be scaled enormously, which increases the impact. Adding to these complexities, both genomics and AI-enabled technologies are rife with scientific and technological uncertainties, which makes the regulation of these technologies not only challenging in itself, but also creates legal uncertainties. In science, the precautionary principle has been used globally to govern uncertainty, with the specific aim to prevent irreversible harm to human beings. The regulation of uncertainties in AI-enabled technologies is based on risk as set out in the AI Regulation that was recently proposed by the European Commission. However, when genomics and artificial intelligence are combined, not only do uncertainties double, but the current regulation of such uncertainties towards the safe use thereof for humans seems contradictory, considering the different approaches followed by science and technology in this regard. In this article, I explore the regulation of both scientific and technological uncertainties and argue that the application of the precautionary principle in the context of human genomics and AI seems to be the most effective way to regulate the uncertainties brought about by the combination of these two technologies.

Significance:

The significance of this article rests in the criteria framework proposed for the determination of the applicability of the precautionary principle and lessons learnt from the European Union's attempt to regulate artificial intelligence.

Introduction

Human genomics has the potential to provide an efficient and cost-effective means of preventing, diagnosing, and treating major diseases that burden populations and enables the tailoring of medicine to the specific needs of individuals. However, the exact impact of this rapidly evolving scientific field on diagnostic and therapeutic health services, and how it will affect societies, are still largely uncertain and subject to ongoing research. Since the completion of the draft human genome sequence more than 20 years ago, an extraordinary amount of genomic data has been generated, which will only increase in volume and complexity alongside the increase in genomic sequencing and related biological techniques. These circumstances force genomics researchers to turn to artificial intelligence (AI) and related machine learning (ML) based computational tools to help them extract, interpret, and analyse information from these valuable data sets into formats that can be used and translated into meaningful outcomes and effective treatments. Similar to human genomics, computer scientists are also continuously developing new techniques and technologies in their field of AI and ML, making it very dynamic, but also very complex and uncertain, which seems to be one of the most common and difficult problems to solve in AI-enabled technologies.¹

Regardless of the fact that the combination of genomics and AI/ML has only started fairly recently, some of the medical breakthroughs it envisions include

*examining people's faces with facial analysis AI programs to accurately identify genetic disorders; using ML techniques to identify the primary kind of cancer from a liquid biopsy; predicting how a certain kind of cancer will progress in a patient; identifying disease-causing genomic variants compared to benign variants using machine learning; and using deep learning to improve the function of gene editing tools such as CRISPR.*²

But despite the positive changes that these technologies promise, one cannot ignore that they are founded on rapidly developing and ever-evolving genomics and AI/ML technologies – fields that are both rife with scientific and technological uncertainties, and which uncertainty is merely exacerbated by their combined use, which in turn creates regulatory uncertainties.

Some of the most pressing ethical, legal, and societal issues associated with the combination of human genomics and AI/ML were presented by Farmer³ during the Global Alliance for Genomics and Health's (GA4GH) 10th Plenary Meeting in September 2022 and are summarised in Table 1. Although AI-powered genomics enhances the collection of data and the accuracy of genomic analysis, it still presents problems relating to missing data, bias, privacy, consent, and genetic discrimination in general. Due to its speed and ability to scale, AI has not only exacerbated these problems, but also added new ones such as interpretability, explainability, accountability, and enabling the ease with which more sensitive inferences can be drawn from genomic data – all whilst life sciences and big tech operates with critically different business models, incentives, cultures, and approaches to ethics.

© 2023. The Author(s). Published under a Creative Commons Attribution Licence.

Table 1: Ethical, legal, and social implications associated with human genomics and artificial intelligence / machine learning

Artificial intelligence / machine learning	Human genomics
Relies on mass data collection <ul style="list-style-type: none"> creates incentives to undermine privacy environmental impact of data storage 	Genomics data privacy <ul style="list-style-type: none"> the problem of secondary subjects genome data are hard to anonymise genomic data are particularly sensitive, and their value is hard to predict
Reliability	Reliability
Differential accuracy	Differential accuracy
Bias	Bias
	The genomic data double bind and 'double vulnerability'
Explainability and interpretability	
Accountability for AI decision-making	
Subjection to AI decision-making	
Ownership of and benefit from AI and its outputs	Ownership of and benefit from genomic data <ul style="list-style-type: none"> HeLa cells Public think they own their genomic data
	Cost and opportunity cost <ul style="list-style-type: none"> Question marks over the current value of genomic science The value of investment in genomics compared to other interventions or research

The aim of this paper is not to discuss the various ethical, legal, and social implications (ELSI) and related issues in detail, but to compare the precautionary principle that is widely used in genomic research with the risk-based approach embedded in proposed AI legislation, and to analyse the appropriate regulatory approach to govern scientific and technological uncertainties that will support scientific and technological innovation, without compromising the safety of people. Reference to the numerous ELSI with regard to the combined use of genomics and AI/ML serves to indicate the complexity of both of these large and emerging research fields, and how their inevitable combination adds to such complexity and uncertainty in their regulation.

Challenges posed by secondary findings in genomics

Genomic research often reveals 'unsolicited' or 'incidental' findings that may be important to the health, treatment, or future health of participants. While it is widely accepted that researchers have a moral obligation to disclose and report secondary findings to participants if there is effective treatment available for the specific health condition with an immediate onset, researchers are less widely considered to have a moral obligation to actively search for health-related findings, especially if it falls outside the scope of the research project.⁴ Koplin et al.⁴ argue that the only reason that genomic researchers are currently not morally obligated to actively search for secondary findings is because the present costs involved in

doing so still far outweigh likely benefits to the participants. However, by combining genomic research with AI/ML, researchers may soon acquire a moral obligation to actively search for secondary findings in the near future when the process of searching for such findings becomes more cost-effective, and serious harm to participants can actually be prevented through rapid improvements of technologies and treatments. But to what extent the benefits to participants must outweigh the costs associated with looking for secondary findings, to determine the moral duty of genomic researchers, is and may remain very uncertain. In an effort to provide guidance in this context, the American College of Medical Genetics published a list of medically actionable secondary findings that researchers must look for and report when doing clinical genome sequencing.⁵ But being non-binding recommendations, only some researchers strictly followed these suggestions, whilst others were reluctant to do so due to their concerns with the medical reality that only a small percentage of genetic variants associated with disease would actually result in participants manifesting with disease.⁶ Despite an updated list of medically actionable findings to return secondary findings, published by the American College of Medical Genetics, there is still no consensus among researchers, clinicians, and bioethicists about when, what, and how secondary findings must be sought or returned when found.⁷ In addition, a growing number of studies that investigate the preferences of the general public, patients, and research participants in this regard, including the impact on these groups of people upon receiving secondary findings, indicates that policies about the returning of secondary findings will be strongly influenced by increased public understanding of genomics and their subsequent preferences, alongside the views of experts.⁸

Further arguments on whether to report secondary findings trigger numerous ethical questions relating to autonomy, non-maleficence, and beneficence, – principles which are often contradictory to one another and in themselves inadequate to justify a fair and reasonable solution. In this regard, Saelaert et al.⁹ argue that the mandatory reporting of actionable secondary findings could even be interpreted as a "technological, soft paternalism" when participants' choices or access to their personal information are restricted by scientists, but may be ethically acceptable if the motives behind such restrictions are valid and the beneficial outcome for the participant is very likely. Subsequently, a patient's inability to make informed decisions relating to their future treatment, normative rationality, the efficacy of outcomes that may be beneficial to the patient, and how that beneficence should be determined, must be considered critically.

Even the seemingly simple act of recontacting participants after genetic and genomic research results have been reinterpreted is a complex issue involving a network of clinical and research laboratories, clinicians, and researchers across specialties. At present, the recontacting of participants necessitated by research findings occurs on an ad-hoc basis which may lead to information being provided only to those participants who can be easily located, or only in so far as research funding allows this to occur. To provide much needed guidance in this regard, the American Society of Human Genetics issued a position statement containing recommendations on how to operationalise the recontacting of participants, including when and how this should be done.¹⁰ Although these recommendations provide a set of principles researchers can use when they anticipate situations in which the return of study findings and the recontacting of participants may become appropriate, the operationalisation of these principles is still subject to institutional ethical review and the purview of advisory boards with regard to the practical implementation thereof.¹¹ In addition, these recommendations were issued in the midst of an evolving genomic and technological landscape with rapid changes occurring in IT, including AI/ML, which in turn will have significant influences on society's beliefs, values and approach to the implementation of these recommendations. Accordingly, recommendations and policies in this regard will have to be updated on a regular basis to keep pace with scientific and technological developments. It is in this context that the precautionary principle in genomic research finds its application to ensure the equitable and effective delivery of high-quality research results, including to those who participate in research.

For many of the above reasons, technological pessimists who fear the appearance of so-called ‘sorcerer’s apprentices’, advocate for stringent regulation of genomic activities; in contrast, technological optimists seem to have complete faith in the scientific progress and oppose regulation based on their argument that regulation acts only to stifle scientific progress. The precautionary principle poses a useful method of thinking to appease both the concerns of technological pessimists, whilst still allowing enough regulatory room for scientific innovation to thrive, specifically in circumstances in which genomic research activities and/or the application of cell and gene therapies poses uncertainty and potentially both success and risk. But, to consider the place and function of the precautionary principle in the combination of genomic science and AI/ML technologies, the extent and consequences of involving AI and ML in genomics must also be considered.

Challenges posed by AI/ML based computational technologies

Despite the potential that AI/ML holds for genomics and health care in general, some of the ethical issues associated with AI/ML, highlighted in a 2021 study by Stahl, specifically those most relevant to genomics, include

*cost to innovation, harm to physical integrity, lack of access to public services, lack of trust, security problems, lack of quality data, power asymmetries, negative impact on health, problems of integrity, lack of accuracy of data, lack of privacy, lack of transparency, potential for military use, lack of informed consent, bias and discrimination, unfairness, unequal power relations, misuse of personal data, potential for criminal and malicious use, loss of freedom and individual autonomy, contested ownership of data, reduction of human contact, problems of control and use of data and systems, lack of accuracy of predictive recommendations, lack of accuracy of non-individual recommendations, violation of fundamental human rights of end users, unintended, unforeseeable adverse impacts, prioritisation of the ‘wrong’ problems, negative impact on vulnerable groups, lack of accountability and liability, loss of human decision-making, and lack of access to and freedom of information.*¹²

Stahl’s¹² long list of ethical concerns not only shows us the uncertainty that AI/ML technologies bring along, but also cautions us not to reproduce, legitimise, and aggravate these concerns by unquestioningly implementing AI/ML in genomics.

In an effort to regulate some of these concerns, the European Union (EU) published a draft regulation for Artificial Intelligence (AI Regulation) on 21 April 2021, but none of the practical summaries, comments, or presentations contained in this draft deals with the fundamental question of how to regulate the above concerns and uncertainties brought about by AI/ML. Being fully aware of the uncertainties and risks that AI poses, the EU opted to introduce a risk-based approach for the regulation of risks associated with AI systems, based on three tiers: (1) unacceptable risk – which simply bans the use of any AI systems posing unacceptable risk; (2) high risk – which subjects high-risk AI systems to extensive technical, monitoring, compliance and transparency obligations; and (3) low risk – systems which are encouraged to self-regulate by implementing codes of conduct.¹³ Once the highest compliance risks to an organisation have been identified and the organisation manages to successfully reduce the identified risks with the prescribed compliance methods and tools, the AI system risk level can then be reduced to a lower one. From the perspective of using such a risk-based approach to protect data, the Article 29 Data Protection Working Party already stated in 2014 that the risk-based approach must span beyond a narrow “harm-based-approach” that only focuses on the prevention of damages, and that it should also take into account

*every potential as well as actual adverse effect, assessed on a very wide scale ranging from an impact on the person concerned by the processing in question to a general societal impact (e.g. loss of social trust).*¹⁴

The draft AI Regulation defines AI by referring to software systems that generate outputs for human-defined objectives (which explains its application in the field of genomics) as:

*...software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.*¹⁵

In Annex I of the draft Regulation, almost every technique currently known that relates to ML approaches, logic- and knowledge-based approaches, and statistical approaches is listed.¹⁵ This list was clearly intended to encapsulate a very broad spectrum of AI systems, but whilst doing so also includes “very unspecified objects”¹⁶. Legally speaking, this approach to regulate a very broad range of unspecified technologies with uncertain uses, outcomes, and consequences is extremely undesirable.

It seems that the European Commission tried to regulate risky *techniques* in general, instead of focusing on AI as a technology that employs some of these risky techniques. The effect thereof is that simple existing technologies such as the pocket calculator may be considered as AI in terms of the definition of AI and the techniques listed in Annex I. This situation will inevitably subject most, if not all, technologies using one or more of the techniques mentioned in the draft AI Regulation to stringent compliance regulations, and thereby possibly slow down the uptake, use, and implementation of technologies that do not pose serious technological risks. Rather, the goal of any AI act or regulation, as envisioned by the Article 29 Data Protection Working Party in 2014, should be to protect people against harmful inventions that threaten our fundamental rights, whilst avoid dampening innovation. Ironically, this is exactly the goal of the precautionary principle, but with one big difference: the precautionary principle is not codified in legislation.

The regulation of risks arising from uncertainties, especially those brought about by the combination of genomics and AI/ML, requires an approach from different perspectives because of the many unanswered ethical questions that remain, as discussed above. Accordingly, I will argue that an innovative technology should not only be considered and legislated with regard to its capabilities or its need to respect certain ethical principles, it must also be considered in light of the precautionary principle, having regard to possible irreversible damages, bias and inequity, privacy issues, and discrimination it may cause.

The precautionary principle

Legislation and associated regulations are not ideal tools that can provide immediate protection against pressing scientific or technological harms. These require a much longer and protracted process from drafting a bill to final enactment. In contrast to this process, and although no universally accepted definition of the precautionary principle exists, the precautionary principle is considered to enable decision-makers to adopt precautionary measures promptly when scientific evidence about an environmental or human health hazard is uncertain and the risks to human life and society are high.¹⁷

The precautionary principle has its origins in international environmental protection¹⁸, and was incorporated into almost all international treaties on environmental protection since the 1990s to the extent that France even incorporated this principle into its Constitution in 2005¹⁹, with Sweden, Belgium, the Netherlands and Australia formally incorporating it into their national environmental policies. This principle then became widely applied by states, in accordance with their national capabilities and where threats of “serious or irreversible damage, *lack of full scientific [and technological] certainty* shall not be used as a reason for postponing cost-effective measures to prevent environmental

degradation” (my addition and emphasis).²⁰ The precautionary approach is thus a broad epistemological, philosophical, and legal approach to innovations that pose a potential for causing harm when extensive scientific knowledge, and I will add technological knowledge, on the matter is lacking. It emphasises caution, pausing, and review before leaping into new innovations that may prove disastrous.

Whilst there is still no global consensus on the legal status of the precautionary principle in the context of international law, the European Union Court of Justice explicitly stated that:

*the precautionary principle can be defined as a general principle of Community law requiring the competent authorities to take appropriate measures to prevent specific potential risks to public health, safety and the environment, by giving precedence to the requirements related to the protection of those interests over economic interests.*²¹

And the European Commission is of the opinion that “this principle has been progressively consolidated in international environmental law, and so it has since become a full-fledged and general principle of international law”¹⁷.

Even though South Africa is a signatory to the Rio Declaration which imported the precautionary principle into South Africa’s policy frameworks, the precautionary principle has had limited national practical application, and I agree with Glazewski and Plit²² that the active implementation of this principle should be given serious consideration, especially considering South Africa’s national development agenda. South Africa’s National Development Plan 2030 states that “science and technology are fundamentally altering the way people live, connect, communicate and transact, with profound effects on economic growth and development” and the application of the precautionary principle will be fundamental to the furthering of “technological and scientific revolutions which underpin economic advances, improvements in health systems, education and infrastructure”²³. In addition, considering that the South African government considers Europe to “continue to be South Africa’s biggest trading partner for some years to come”²⁴, and Europe’s stance on the status of the precautionary principle as discussed above, it is advisable that this principle be implemented into scientific and technological developments sooner rather than later.

The scope and extent of the implementation of this principle will depend on prevailing social and political values and could be developed further in case law resulting from legal action, which makes this principle an ideal tool to dynamically regulate the uncertainties of emerging sciences and technologies in line with socio-political developments without amounting to legal uncertainty. A key variable in this regard is the degree of scientific or technological uncertainty that would likely mobilise authorities into action, having due regard to the severity and probability of the risks involved, the magnitude of the stakes, and the potential costs of action or inaction. However, I agree with Stirling²⁵ that although a science-based risk assessment offers a powerful method to determine strict states of risk, it is not applicable under conditions of uncertainty, ambiguity and ignorance and such reductive methods, in the absence of a strict state of risk, may prove to be irrational, unscientific, and potentially misleading. From a regulatory perspective, the quantification of risk, or a definitive expert judgement on safety, is of immense value for purposes of creating concrete legislation; but, unfortunately, this has no rational scientific basis. It is also expected that robust legislation must address long-term issues for effective governance, where robustness is a result of the accuracy of assessment results, not of their professed precision, hence the seemingly impossible task to regulate scientific and technological uncertainties via legislation. Stirling continues to explain that the reason that so-called “sound scientific” procedures often yield contrasting pictures of risk, is based on the specific framing of the analysis of answers delivered in risk assessments, which in turn can dramatically influence the framing of science for policy. It is in this context that the value of the precautionary principle becomes clear.

The precautionary principle is not, and has never been claimed to be, a definitive decision-making tool, nor a detailed protocol that can be used to determine risks and uncertainties, but it does provide a general, yet dynamic, normative guide towards effective policymaking in times of uncertainty where the benefit of any doubt should be tilted towards the protection of human health, specifically in the case of AI/ML-enabled genomics. This means that the implementation of the precautionary principle requires a level of scientific and technological motivation and persuasion on the side of scientists and technologists with regard to the gathering of evidence. In these circumstances the value of the precautionary principle manifests in the fact that none of these issues can be dealt with in a strict scientific way. Instead, the precautionary principle demands the incorporation of a broader range of non-reductive methods that include a wide variety of methods to reveal the normative and contestable basis for decisions, to regulate scientific and technological uncertainties.

Applying the precautionary principle

The main question is how to identify those cases that justify the application of the precautionary principle. In this regard, the 2005 report on the precautionary principle published by UNESCO’s World Commission on the Ethics of Scientific Knowledge and Technology, states that “when human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm”²⁶. An answer clearly stated in the precautionary principle itself, and defined specifically as a response to lack of scientific certainty when there is a threat of serious or irreversible harm. Morally unacceptable harm, according to this report, is harm that threatens human life or health, is effectively irreversible, inequitable to future generations, or is imposed without consideration of the human rights of those affected, with the caveat that the plausibility of such harm must be based on scientific analysis and subject to review.

In the context of the combined use of AI/ML and genomics, the following framework for criteria may, for example, be used as a screening process to identify scientific and technological uncertainty, and most importantly, their impact on human life and health to decide whether to apply the precautionary principle:

- risks and harms posed to public and/or individual health and life and physical integrity
- degree and type of scientific and technological uncertainty
- presence or absence of morally acceptable harm
- impact of genomic secondary finding disclosure on fundamental rights of individuals and/or community
- reliability, accuracy, and bias of AI/ML-enabled genomic predictions and predictive recommendations
- benefit to individuals and/or society
- scientific and technological doubts about quality, accuracy, applicability, and transparency of data
- power asymmetries
- general violation of fundamental human rights
- novel, unintended, unforeseeable, unprecedented, or adverse impacts
- clear violation of risk-based concentration thresholds or standards
- scientifically and technologically founded doubts on theory, model sufficiency, or applicability
- divergent individual or institutional perceptions of risk
- ethical, legal and social concerns, distributional issues or political mobilisation²⁷

When none of the criteria in this framework is triggered, the AI/ML-enabled genomics application in question does not need the application of the precautionary principle, in which event the case will be subject to conventional risk assessment. Only when uncertainty is prevalent will

it justify the initiation of a more precautionary approach. This hopefully shows how the precautionary principle does not present a blanket rejection of science, technology or even risk assessment, but rather triggers a careful consideration, measuring and approach towards the combination of genomics and AI/ML at different states of scientific and technological knowledge.

Conclusion

In proposing the AI Regulation, the European Commission tried to regulate the technological uncertainty brought about by AI, by attempting to introduce some legal certainty via further risk-based thresholds, in addition to existing legal requirements. The risk-based approach, introduced by this AI Regulation, functions on the assumption that only those AI systems that pose a high or moderate risk to fundamental rights will fall within the scope of the risk categories as set out in the AI Regulation, meaning that only those AI systems need to comply with the requirements of the proposed AI Regulation. However, although the regulation of risks and harms are preferable, this AI Regulation contains an overly broad definition of AI, as expanded upon in Annex I, which creates immense legal uncertainty as to what technologies actually fall within the ambit of the proposed regulation, over and above the technological uncertainties discussed above. Furthermore, if such a definition and expanded list of technologies is contained in a single AI act of any kind, any piece of legislation that tries to regulate any kind of software will find itself competing with the conditions set out in such an overarching act and possibly contain some contradictory clauses of its own. From a legal perspective, this scenario will only complicate the interpretation and application of legislation in scientific and technological fields which are already complex enough to govern due to rapid developments in these fields.

Thus, if we truly want to prepare ourselves for a functional, fair, reasonable, legal and ethical future in which the combination of genomic science and AI/ML serves human beings and contributes to the prospering of their existence, we should ensure that we capture scientific and technological techniques that may not be currently known, such as the combination of these technologies, instead of limiting ourselves exclusively to AI. Because absolute scientific and/or technological certainty, especially when combined, can never be achieved, the application of the precautionary principle in these circumstances can provide a dynamic framework that could help to achieve a better balance in genomics and AI/ML-based health outcomes and policies, whilst mitigating the difficulties presented by both scientific and technological uncertainty, before stringent regulations are enacted that may not be flexible enough to enable scientific and technological advancement. No single act, regulation, policy or guideline is enough to effectively protect fundamental rights and democracy and, most importantly, to avoid irreversible damage caused by the combined use of genomics and AI. The proposed AI Regulation, for example, does not deal with damages that may occur when applying AI to health care or in automated and opaque decisions, nor in the application of AI in the context of genomics. Hence the need to apply the precautionary principle in these circumstances to allow for the consideration of a vast array of governing instruments, ethical principles, and scientific and technological practicalities to allow for sustainable development in real time, whilst preserving the fundamental rights of both present and future generations.

Recommendations

South Africa has no formal policy documents relating to AI, nor has it entered bills to parliament for the regulation of AI. Instead, AI is regulated under existing legal principles as and when applicable. Rather than reinventing the wheel, South Africa can learn from the EU's attempts to regulate AI, and prevent many of the mistakes made in trying to govern AI per se. I therefore propose:

1. that any legislative effort in this regard must broaden the scope to rather regulate technologies, as opposed to limiting it to AI or ML exclusively, to allow for the long-term regulation of technologies, including those yet unknown;
2. to incorporate the precautionary principle into such legislation, much like the ethical principle of consent is now incorporated

into legislation globally, to allow for the consideration of a broader spectrum of consequences when dealing with scientific and technological uncertainties; and

3. considering the combined use of genomics and AI/ML, any legislative effort should also include non-digital technologies that may pose a threat to our fundamental rights, such as certain bio-technologies – this should prevent regulations to be treated in a sectorised way or without coordinated planning or with little technicality.

Acknowledgements

This work was funded by the Luxembourg National Research Fund (FNR) IS/14717072 'Deceptive Patterns Online (Deception)'; the European Union's Horizon 2020 Innovative Training Networks, Legality Attentive Data Scientists (LEADS) under grant agreement ID 956562; and REMEDIS Project INTER/FNRS/21/16554939/REMEDIS (Regulatory solutions to Mitigate DISinformation).

Competing interests

I have no competing interests to declare.

References

1. Wu J, Shang S. Managing uncertainty in AI-enabled decision making and achieving sustainability. *Sustainability*. 2020;12, Art. #8758. <https://doi.org/10.3390/su12218758>
2. National Human Genome Research Institute. Artificial intelligence, machine learning and genomics [webpage on the Internet]. No date [cited 2022 Oct 17]. Available from: <https://www.genome.gov/about-genomics/educational-resources/fact-sheets/artificial-intelligence-machine-learning-and-genomics>
3. Farmer H. Exploring the societal implications of AI and genomics. Paper presented at: Global Alliance for Genomics and Health (GA4GH) 10th Plenary Meeting; 2022 September 22–23; Barcelona, Spain. c2022 [cited 2022 Oct 17]. Available from: <https://assets.swoogo.com/uploads/2072198-633337284f00f.pdf>
4. Kopljin JJ, Savulescu J, Vears DF. Why genomics researchers are sometimes morally required to hunt for secondary findings. *BMC Med Ethics*. 2020;21, Art. #11. <https://doi.org/10.1186/s12910-020-0449-8>
5. US National Institutes of Health. ACMG Recommendations for reporting of incidental findings in clinical exome and genome sequencing [webpage on the Internet]. c2013 [cited 2022 Oct 17]. Available from: <https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/>
6. Mount Sinai Hospital. Most 'pathogenic' genetic variants have a low risk of causing disease. *Medical Xpress*. 25 January 2022 [cited 2022 Oct 17]. Available from: <https://medicalxpress.com/news/2022-01-pathogenic-genetic-variants-disease.html>
7. AlFayyad I, Al-Tannir M, Abu-Shaheen A, AlGhamdi S. To disclose, or not to disclose? Perspectives of clinical genomics professionals toward returning incidental findings from genomic research. *BMC Med Ethics*. 2021;22(1), Art. #101. <https://doi.org/10.1186/s12910-021-00670-y>
8. Purvis RS, Abraham TH, Long CR, Stewart MK, Warmack TS, McElfish PA. Qualitative study of participants' perceptions and preferences regarding research dissemination. *AJOB Empir Bioeth*. 2017;8(2):69–74. <https://doi.org/10.1080/23294515.2017.1310146>
9. Saelaert M, Mertes H, Moerenhout T, De Baere E, Devisch I. Ethical values supporting the disclosure of incidental and secondary findings in clinical genomic testing: A qualitative study. *BMC Med Ethics*. 2020;21(1), Art. #9. <https://doi.org/10.1186/s12910-020-0452-0>
10. Bombard Y, Brothers KB, Fitzgerald-Butt S, Vassy JL, Wagner JK, Levy HP. The responsibility to recontact research participants after reinterpretation of genetic and genomic research results. *American Society of Human Genetics Position Statement*. *Am J Hum Genet*. 2019;104(4):578–595. <https://doi.org/10.1016/j.ajhg.2019.02.025>
11. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng CM, Evans JP, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing. A policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017;19(2):249–255. <https://doi.org/10.1038/gim.2016.190>



12. Stahl BC, editor. Ethical issues of AI. In: Artificial intelligence for a better future. Cham: Springer; 2021. p. 35–53. https://doi.org/10.1007/978-3-030-69978-9_4
13. European Commission. Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. 2021/0106 (COD). Brussels: European Commission; 2021.
14. European Union. Article 29 Data Protection Working Party. Statement on the role of a risk-based approach in data protection legal frameworks. WP 2018. European Union; 2014.
15. European Commission. Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. 2021/0106 (COD), Article 3(1). Brussels: European Commission; 2021.
16. Dufour R, Koehof J, Van der Linden T, Smits J. AI or more? A risk-based approach to a technology-based society. Oxford Business Law Blog. c2021 [cited 2022 Oct 21]. Available from: <https://blogs.law.ox.ac.uk/business-law-blog/blog/2021/09/ai-or-more-risk-based-approach-technology-based-society>
17. Bourguignon D. The precautionary principle: Definitions, applications and governance. PE 573.876. European Parliamentary Research Service; 2015.
18. United Nations. Report of the United Nations Conference on Environment and Development. New York: United Nations; 1993.
19. Ewald F. Situation in France: The principle of precaution. In: Houdy P, Lahmani M, Marano F, editors. Nanoethics and nanotoxicology. Berlin: Springer; 2011. p. 483–494. https://doi.org/10.1007/978-3-642-20177-6_24
20. United Nations. Rio Declaration on Environment and Development. Principle 15. New York: United Nations; 1992.
21. Judgment in the case of Artegodan v. Commission of 26 November 2002 (T-74/00), paragraph 184.
22. Jan Glazewski J, Plit L. Towards the application of the precautionary principle in South African law. Stellenbosch Law Review. 2015;26(1).
23. The Presidency of the Republic of South Africa National Planning Commission. National Development Plan 2030. Our future – make it work. Pretoria: The Presidency; 2013.
24. McKinsey. Africa's path to growth. McKinsey Quarterly. c2010 [cited 2022 Oct 23]. Available from: <https://www.mckinsey.com/featured-insights/middle-east-and-africa/africas-path-to-growth-sector-by-sector>
25. Stirling A. Risk, precaution and science: Towards a more constructive policy debate. Talking point on the precautionary principle. EMBO Rep. 2007;8(4):309–315. <https://doi.org/10.1038/sj.embor.7400953>
26. UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology (COMEST). Report on the precautionary principle [document on the Internet]. c2005 [cited 2022 Oct 23]. Available from: <https://unesdoc.unesco.org/ark:/48223/pf0000139578>
27. Stirling A, Ely A, Renn O, Dreyer M, Borkhart K, Vos E, et al. A general framework for the precautionary and inclusive governance of food safety: Accounting for risks, uncertainties and ambiguities in the appraisal and management of food safety threats. Stuttgart: University of Stuttgart; 2006.