GUEST LEADER

Check for updates

AUTHORS:
Keymanthri Moodley[1,2] (iD)
Stuart Rennie[2,3] (iD)

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

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

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

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 missive 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-intense 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. Petrozzino C. Who pays for ethical debt in AI? AI Ethics. 2021;1:205–208. https://doi.org/10.1007/s43681-020-00030-3