A Roadmap for Building Data Science Capacity for Health Discovery and Innovation in Africa

Joseph Beyene [1]*, Solomon W. Harrar [2], Mekibib Altaye [3], Tessema Astatkie [4], Tadesse Awoke [5], Ziv Shkedy [6] and Tesfaye B. Mersha [3]*

[1] Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada, [2] Dr. Bing Zhang Department of Statistics, University of Kentucky, Lexington, KY, United States, [3] Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, United States, [4] Faculty of Agriculture, Dalhousie University, Truro, NS, Canada, [5] Department of Epidemiology and Biostatistics, University of Gondar, Gondar, Ethiopia, [6] I-BioStat, Hasselt University, Diepenbeek, Belgium

Technological advances now make it possible to generate diverse, complex and varying sizes of data in a wide range of applications from business to engineering to medicine. In the health sciences, in particular, data are being produced at an unprecedented rate across the full spectrum of scientific inquiry spanning basic biology, clinical medicine, public health and health care systems. Leveraging these data can accelerate scientific advances, health discovery and innovations. However, data are just the raw material required to generate new knowledge, not knowledge on its own, as a pile of bricks would not be mistaken for a building. In order to solve complex scientific problems, appropriate methods, tools and technologies must be integrated with domain knowledge expertise to generate and analyze big data. This integrated interdisciplinary approach is what has become to be widely known as data science. Although the discipline of data science has been rapidly evolving over the past couple of decades in resource-rich countries, the situation is bleak in resource-limited settings such as most countries in Africa primarily due to lack of well-trained data scientists. In this paper, we highlight a roadmap for building capacity in health data science in Africa to help spur health discovery and innovation, and propose a sustainable potential solution consisting of three key activities: a graduate-level training, faculty development, and stakeholder engagement. We also outline potential challenges and mitigating strategies.

Keywords: big data, health informatics, capacity building, knowledge discovery, data science, Africa, training, stakeholder

INTRODUCTION

NIH's Strategic Plan for Data Science released in June 2018 (1) defines data science as the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data (2). The constant evolution of technology in our digital world generated a growing need to discover knowledge and support decision in near real time from large volume of data sets. The versatility, diversity, and connectivity of data capturing devices available today allow data to be generated and stored at increasingly high speed. From national health systems to data collected at rural clinics to the most advanced

high-throughput sequencing technologies data are central to our ability to improve health. As data are becoming deeper and richer with new sources of data generated from new technologies and sensors (e.g., social media, geospatial data, mobile phones, wearables, electronic medical records, bioimaging, and genomics), our ability to harness and leverage useful knowledge from these data are critical to accelerate discoveries and innovations that can impact public health (3). Properly harnessed data can provide insights and drive discovery that will accelerate biomedical advances, improve patient outcomes, and reduce costs.

Health Data Science is crucial because traditional study design and analytical approaches are inadequate to tackle challenges posed by the unprecedent volume of large and unstructured datasets. New knowledge generated through the power of data science could enhance precision and patient-focused medicine, cost-effective drug discovery, improvement in patient outcome and delivery of care, as well as support policy makers. Potential applications of data science are increasingly being reported in a wide range of health areas including child health (4), mental health (5), critical care (6), laboratory medicine (7), clinical pharmacology and drug development (8), non-communicable diseases (9–11), physical medicine and rehabilitation (12), and infectious diseases such as COVID-19 pandemic (13).

Some argue that the world's most valuable resource is no longer oil, but data (14). But as oil needs to be refined, data must be properly and optimally analyzed so that it can be transformed into new knowledge. The key question then becomes, how could the availability of data be harnessed so that health innovations and breakthroughs can be achieved and help alleviate sufferings of individuals, communities and society at large, as well as reduce the economic burden on healthcare systems? Addressing this question is urgent in the context of Africa. Like the rest of the world, recent technological advances are enabling African researchers to collect voluminous and complex data at an unprecedented rate on a wide range of health conditions and domains including biomedical, clinical, public health and health systems. However, the ability to harness these data and generate new knowledge is lagging in Africa due to lack of well-trained data scientists.

Although Africa comprises 15% of the world's population, it bears 25% of the global disease burden (15). Africa's population is expected to double by 2050 as the rate of growth is higher than any other continent including Asia and Latin America. The burden of disease both with respect to communicable and non-communicable diseases is striking across the African continent (9–11, 16, 17). The role data science played at a global level in combatting the Covid-19 pandemic–from infectious disease modeling approaches to risk prediction for various subgroups of populations worldwide–cannot be understated. Covid-19 affected almost every nation on earth, but other infectious diseases like malaria, tuberculosis, HIV/AIDS and Ebola continue to be major causes of mortality and suffering in Africa (13).

Billions of dollars have been committed to combat these and many other communicable diseases by various global funding agencies and there is a wealth of data collected over several decades. The whopping increase in population coupled with disproportionate global disease burden requires local talent to investigate context-specific risk factors, discover new knowledge, and produce relevant and timely evidence to impact health practice and policy appropriate to the culture, aspirations and developmental goals of people and governments in the region.
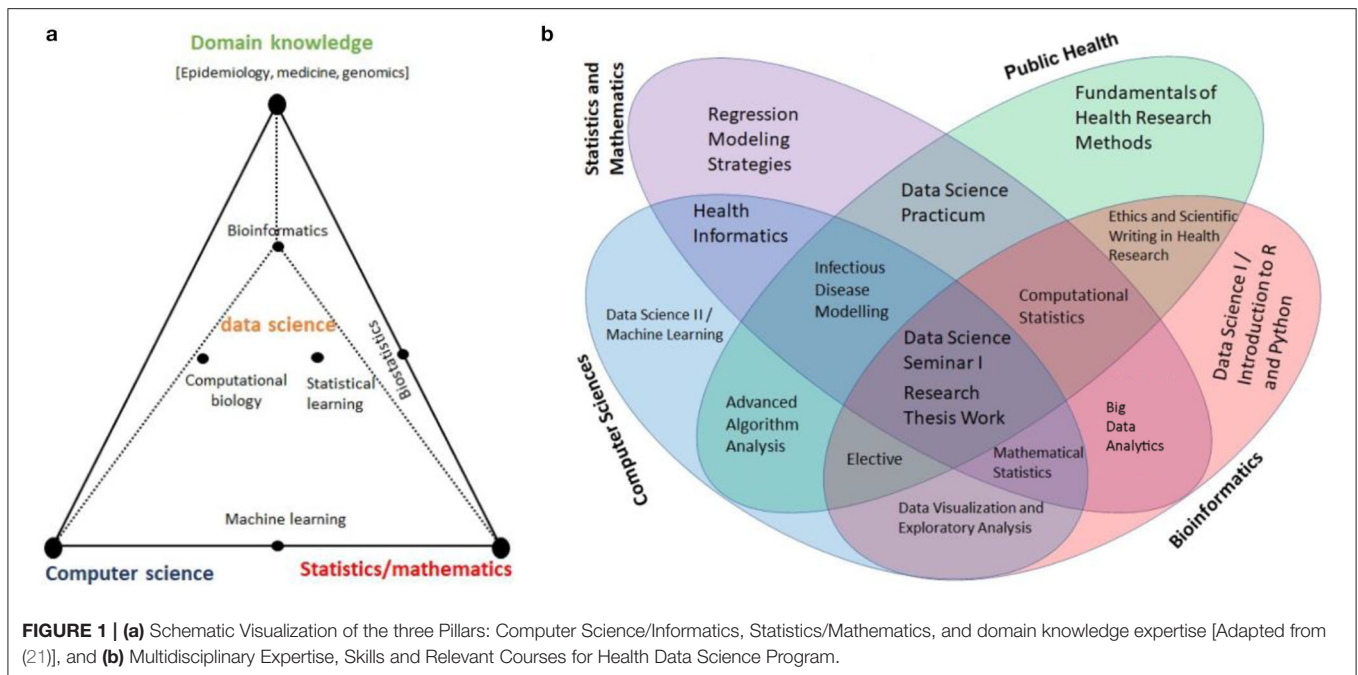
Data science has important implications in achieving the United Nation's Sustainable Development Goals (SDGs) (18). The SDGs highlight that achieving health and well-being for all requires harnessing data and creating new knowledge and innovations in health and other sectors (19). For example, pattern recognition methodologies and tools allow identifying a segment of the population that might be at high risk for developing chronic and non-communicable diseases. Spatial-temporal data science approaches are crucial in detecting "hot spots" and trajectories over time of emerging and re-emerging health problems including communicable diseases across communities and regions. In this paper, we propose a roadmap for developing a strong health data science program in the African context through problem-based graduate-level training activities, faculty development and stakeholder engagement.

BUILDING HEALTH DATA SCIENCE CAPACITY IN AFRICA

Graduate-level degree programs in Health Data Science are gradually emerging especially in European and North American Universities. African Universities, especially those in the East Africa region, lack Health Data Science programs. From a funding point of view, there are encouraging initiatives that are intended to promote the establishment of Health Data Science programs in Africa. One such initiative is the recent announcement by the U.S. National Institutes of Health (NIH) with a significant investment ($58 million) to catalyze data science and health research innovation in Africa (20).

Training Health Data Scientists

Building modern health data science capacity is feasible in many countries in Africa mainly due to the relatively less expensive infrastructural and training requirements compared to similar activities conducted in laboratory-intensive disciplines, increasing Internet penetration rates, as well as improved access to health-related public databases and open-source software. In recent years, data science programs for public health and biomedical data have started to emerge. Existing Health Data Science programs are mostly at the Masters level with the possibility of pursuing an interdisciplinary PhD degree. However, the standards for the composition of course work and research requirements for a PhD-level rigorous training are still under development in many universities. In launching Health Data Science program, the overarching goals should include: (a) train a cohort of students in Health Data Science that will have the skills to become independent investigators, research leaders, and research collaborators and contribute to Health Data Science research in Africa, (b) faculty development initiatives to strengthen and improve the curricula to match the rapidly changing technological advancement, and (c) participate and

FIGURE 1 | (a) Schematic Visualization of the three Pillars: Computer Science/Informatics, Statistics/Mathematics, and domain knowledge expertise [Adapted from (21)], and (b) Multidisciplinary Expertise, Skills and Relevant Courses for Health Data Science Program.

create expert hub and networking for groups focusing on data science training, which might cover topics such as core competencies, curriculum sharing, and supporting other similar programs. In addition, the curriculum should reflect the cultural, traditional and language contexts that are relevant to health problems facing Africa.

The main module of the Health Data Science Training Program should encompass three interdisciplinary areas: (a) Computer Science/Informatics, (b) Statistics/Mathematics, and (c) Domain knowledge experts (21). **Figure 1a** shows these three pillars along with some examples that combine skills from these focus areas, and **Figure 1b** displays diverse expertise and skills necessary for a successful Health Data Science training program. The program should include mentors representing varied disciplines with expertise from basic sciences to community-based applied research. There are many African universities that are qualified to host and offer training in Health Data Science. Universities with strong programs in Medical, Biomedical, Public Health, Statistics, Informatics, and Computer Science at MSc, PhD and MD level can serve as the primary hubs to advance training program in Health Data Science in Africa.

The Health Data Science graduate program curricula should prepare prospective students in Africa for careers involving the use of data to inform public-health decision making. The program should involve training in the design, analysis and reporting of health science data, using a blend of traditional and modern analytic and computational techniques. We propose a list of pertinent courses based on which a curricula for graduate programs at various levels can be developed. A set of courses that may be required for acquiring core competency in Health Data Science are listed in **Table 1A.** These courses

spanning health research methodology, statistics and informatics will allow trainees to learn critical skills important in research and application of Health Data Science. In addition to the core courses, a training program should also include a wide range of elective courses (**Table 1B**) that students can choose from depending on their interest and research focus.

Faculty Development

A diverse and accomplished group of faculty members from various disciplines are crucial for a successful Health Data Science program. The goals of training programs can be achieved and sustained only if institutions invest in the professional development of their faculty members so that they can have successful career in education, research and academic leadership. To address this problem, institutions should create a wide range of faculty development opportunities including customized short-term trainings and workshops for faculty and implement mentorship programs to enhance teaching, research, and service/leadership capacity, and in particular strengthen career development of junior faculty members.

For faculty members to maintain competence and learn about new and developing areas in data science, workshops and short-courses should be designed and offered on a regular basis. Similarly, creating learning opportunities that will enhance scientific writing skills of faculty members, in particular early career researchers, is crucial. The ability to publish research findings in peer-reviewed journals and securing research funding are critical for faculty members to establish and sustain a program of research. Therefore institutions should put a plan in place to help their faculty excel in these important academic activities. Other professional development opportunities may include essentials of supervision, mentoring and leadership foundations.

TABLE 1 | Proposed Program for Acquiring Competency in Health Data Science (**1A.** Required Courses; **1B.** Elective Courses).

1A. Required Courses

Statistics/Mathematics	Computer Science	Public Health
• Regression Modeling Strategies • Mathematical Statistics • Computational Statistics • Data Visualization and Exploratory Analysis	• Introduction to Computing in Python • Machine Learning • Advanced Algorithm Analysis	• Health Research Methods • Ethics and Scientific Writing in Health Research • Health Informatics • Epidemics Models and Data Science • Infectious Disease Modeling

1B. Elective Courses

• Bayesian Statistics • Causal Inference • Spatial Analysis • Time Series Analysis • Linear Algebra • Numerical Optimization • Theory of Probability • Advanced Calculus	• Natural Language Processing • Big Data Analytics • Artificial Intelligence • Introduction to Cloud Computing	• Fundamentals of Epidemiology • Statistical Genetics • Epidemics Models and Data Science • Computational Biology

Engaging Stakeholders

For a data science training program to be successful, there is a very important and third component–stakeholder engagement. We distinguish two types of stakeholders: (1) domain-knowledge experts who are part and parcel of the data science research team (along with trainees and faculty members), and (2) individuals or organizations who have interest in new knowledge that data science teams generate. To elaborate further, health research will have the greatest impact in every day practice and policy if the model of collaborative research adopts "co-production" of knowledge involving various stakeholders. Similarly, community engagement across scientific disciplines and disseminating findings using various platforms including E-learning is crucial. This process comprises a close collaboration between subject-matter researchers, data science experts, and knowledge users. Health Data Science students should have opportunity to work closely with health researchers who can identify relevant clinical and other scientific problems and have the authority to implement research recommendations. Stakeholders include groups such as clinicians, biomedical researchers, public health experts, health policy makers, community leaders and specific health advocates and support groups. The various groups have unique expertise pertaining to the research topic of interest for their constituencies and knowledge of the context and potential for implementation.

In the collaboration and engagement framework, data science researchers bring methodological and analytical expertise to the collaboration. There are many potential benefits including better science, relevant and actionable research findings, increased use of evidence in policy and/or practice, and mutual learning. Some specific initiatives will include building relationships with knowledge users for implementing data science competence, engage local, national and regional policy makers related to epidemic modeling, as well as link students with people familiar with data hubs so that they can use their data science skills and address important health problems that is relevant to the stakeholders. In addition, the capacity building effort should explore opportunities for engaging data science experts of African origin in the diaspora to harness their teaching and research skills thereby converting brain drain to brain gain.

An E-learning platform could be developed on a minimum budget (22). While the value of courses to in-person attendees is clear and essential, inexpensive access to the internet allows reaching out to a large and geographically dispersed audience, expanding the impact beyond the attendees on campus. The development of an E-Learning platform will help create a collaborative network of users. It should be noted that by E-Learning system we do not mean a "distance learning system," rather a Web-based platform which offers materials to use in the class for "on campus" courses in Health Data Science. As was proven during the COVID-19 pandemic in many universities across the world, a good and functional E-learning system is essential to ensure the continuation and stability of any educational programs.

OTHER RELEVANT FACTORS, CHALLENGES AND MITIGATING STRATEGIES

Here, we briefly highlight additional relevant factors that will be needed to launch and sustain a successful Health Data Science training program. First, trainees should have access to real data sets for hands-on exercises, practicum, and thesis work. Broadly, data could be obtained from two sources: (i) primary data collected by subject-matter experts (biomedical scientists, clinicians, and public health researchers routinely collect primary data to address specific scientific questions), (ii) publicly available data. **Supplementary Table 1** (**Supplementary Material**) lists examples of publicly available datasets. Second, responsible conduct of research must be adhered by all involved in the data science program. Considerable attention should be given to ethics, regulatory issues, scientific rigor, transparency, reproducibility, unbiased and responsible dissemination of research findings, data protection and sharing following well-established guidelines such as the NIH guiding principles of ethical research (23). Third, concrete actions should be taken to promote diversity, equity and inclusivity (DEI) in recruitment, advancement, and retention.

Although Health Data Science promises to advance health discovery, innovations and healthcare delivery in Africa, there are several potential challenges in building the needed

capacity. Health Data Science programs require significant investment in infrastructure and human resources. Data collection, storage, management and analysis of big data needs adequate computational facilities including hardware and software, which typically come at a high cost. To mitigate these challenges, public health institutuions, universities, government organizations such as Ministry of Health, and others in the public and private sectors should make efforts to prioritize the budgetary needs of Health Data Science academic programs in their resource allocations. Recruitment and retention of qualified faculty members is another potential challenge in building Health Data Science capacity in resource-limited settings. Brain drain remains to be Africa's largest hurdle for retention of skilled faculty members. The drain is largely driven by lack of incentive mechanisms to recognize the hard work of faculty members and inadequate compensations to off-set the consistently rising cost of living. To mitigate this problem, higher education, research and healthcare institutions in Africa should offer competetive salaries and benefits to their employees, including creating opportunities for professional development, promotion, and a conducive environment.

There are many advantages in strengthening capacity locally including the ability to scale up and reduce the possibility of brain drain. It is also vital that trainees be embedded within the environment where they can fully understand and appreciate unique challenges to the local context, culture and other norms. Data science is inherently team-based, so having trainees work closely with people who generate biomedical, clinical, and public health questions will allow them to learn about important problems, gain a broader understanding of the research enterprise, and become all-rounded data scientists. Training locally will also lead to a sustainable solution to the chronic lack of capacity we currently see in Africa. Other important considerations include the advantage trainees will have by not going too far from their community and family, as well as be part of a group that will establish a strong infrastructure and research culture in their home countries.

SUMMARY AND CONCLUSION

In this paper, we proposed a framework to guide how to build Health Data Science capacity in Africa using three major activities: (1) training health data scientists at a graduate level, (2) faculty development, and (3) stakeholder engagement. These three activities span the entire research process required for addressing health problems in an integrated and collaborative setting. The process encompasses key research components including asking relevant questions, planning appropriate study designs, developing measurement instruments, collecting data, conducting optimal analysis and dissemination of findings. The proposed Health Data Science training program follows a holistic and team-based approach to solving scientific problems related to health: competencies that trainees in the program will acquire, the key role faculty members will play in problem-solving while junior faculty are being empowered to develop their career, and engaging stakeholders to help

define important and context-based health problems as well as implement health innovations (24). Solving important problems requires fostering a collaborative environment and involving various team members at different steps of this cycle with a shared vision and dedication to discovering new knowledge and advancing health innovations (**Supplementary Figure 1**, **Supplementary Material**).

In conclusion, data science is a rapidly evolving multidisciplinary field which has an important role to play in health discovery and innovations. As technology continues to advance, and big and diverse data become common, the evolving field of data science has the potential to provide the opportunity to create a better future for human health by harnessing these data. Unfortunately, most African nations may not reap the benefits of data science due to lack of well-trained data scientists. This lack of capacity must be addressed urgently for Africa not to continue falling so far behind other parts of the world in this important and promising field of science. Data science requires combining rigorous study design with appropriate statistical inference and computational approaches. Scientific and/or clinical domain knowledge is a key ingredient to harness the potential of data science methods and tools. Finally, perhaps more than other multidisciplinary disciplines, critical skills including effective communication and other team science skills, as well as ethical and responsible conduct of research are necessary in Health Data Science.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

TM, JB, and SH conceptualized the study and drafted the manuscript. All authors critically reviewed the manuscript and approved the final manuscript as submitted.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.710961/full#supplementary-material

REFERENCES

1. National Institutes of Health (NIH) Strategic Plan For Data Science (2018). Available online at: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

2. Dunn MC, Bourne PE. Building the biomedical data science workforce. *PLoS Biol.* (2017) 15:e2003082. doi: 10.1371/journal.pbio.2003082

3. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. *Sci Transl Med.* (2015) 7:283rv283. doi: 10.1126/scitranslmed.aaa3487

4. Bennett TD, Callahan TJ, Feinstein JA, Ghosh D, Lakhani SA, Spaeder MC, et al. Data Science for Child Health. *J Pediatr.* (2019) 208:12–22. doi: 10.1016/j.jpeds.2018.12.041

5. Russ TC, Woelbert E, Davis KAS, Hafferty JD, Ibrahim Z, Inkste B, et al. How data science can advance mental health research. *Nat Hum Behav.* (2019) 3:24–32. doi: 10.1038/s41562-018-0470-9

6. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. *Chest.* (2018) 154:1239–48. doi: 10.1016/j.chest.2018.04.037

7. Gruson D, Helleputte T, Rousseau P, Gruson D. Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. *Clin Biochem.* (2019) 69:1–7. doi: 10.1016/j.clinbiochem.2019.04.013

8. Peck RW, Shah P, Vamvakas S, van der Graaf PH. Data science in clinical pharmacology and drug development for improving health outcomes in patients. *Clin Pharmacol Ther.* (2020) 107:683–6. doi: 10.1002/cpt.1803

9. Mapesi H, Paris DH. Non-communicable diseases on the rise in Sub-Saharan Africa, the Underappreciated threat of a dual disease burden. *Praxis.* (2019) 108:997–1005. doi: 10.1024/1661-8157/a003354

10. Kamau A, Mogeni P, Okiro EA, Snow RW, Bejon P. A systematic review of changing malaria disease burden in sub-Saharan Africa since 2000: comparing model predictions and empirical observations. *BMC Med.* (2020) 18:94. doi: 10.1186/s12916-020-01559-0

11. Gouda HN, Charlson F, Sorsdahl K, Ahmadzada S, Ferrari AJ, Erskine H, et al. Burden of non-communicable diseases in sub-Saharan Africa, 1990-2017: results from the global burden of disease study 2017. *Lancet Glob Health.* (2019) 7:e1375–87. doi: 10.1016/S2214-109X(19)30374-2

12. Ottenbacher KJ, Graham JE, Fisher SR. Data science in physical medicine and rehabilitation: opportunities and challenges. *Phys Med Rehabil Clin N Am.* (2019) 30:459–71. doi: 10.1016/j.pmr.2018.12.003

13. WHO. *Global Defence Against The Infectious Diseases Threat.* Geneva: World Health Organization.

14. Economist T. *The World's Most Valuable Resource Is No Longer Oil, But Data (2017).* Available online at: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (accessed March 10, 2021).

15. Simpkin V, Namubiru-Mwaura E, Clarke L, Mossialos E. Investing in health R&D: where we are, what limits us, and how to make progress in Africa. *BMJ Glob Health.* (2019) 4:e001047. doi: 10.1136/bmjgh-2018-001047

16. Arogundade FA, Omotoso BA, Adelakun A, Bamikefa T, Ezeugonwa R, Omosule B, et al. Burden of end-stage renal disease in sub-Saharan Africa. *Clin Nephrol.* (2020) 93:3–7. doi: 10.5414/CNP92S101

17. Mlotshwa BC, Mwesigwa S, Mboowa G, Williams L, Retshabile G, Kekitiinwa A, et al. The collaborative African genomics network training program: a trainee perspective on training the next generation of African scientists. *Genet Med.* (2017) 19:826–33. doi: 10.1038/gim.2016.177

18. United Nations. *Report Of The Secretary-General On Sdg Progress 2019 Special Edition.* Available online at: https://sustainabledevelopment.un.org/content/documents/24978Report_of_the_SG_on_SDG_Progress_2019.pdf United Nations, New York (accessed March 10, 2021).

19. Ezer D, Whitaker K. Data science for the scientific life cycle. *Elife.* 2019;8. doi: 10.7554/eLife.43979

20. NIH. NIH to invest $58M to catalyze data science and health research innovation in Africa. https://www.nih.gov/news-events/news-releases/nih-invest-58m-catalyze-data-science-health-research-innovation-africa News Released Monday, July 27, 2020 (accessed March 10, 2021).

21. Emmert-Streib F, Moutari S, Dehmer M. The process of analyzing data is the emergent feature of data science. *Front Genet.* (2016) 7:12. doi: 10.3389/fgene.2016.00012

22. Barteit S, Jahn A, Banda SS, Bärnighausen T, Bowa A, Chileshe G, et al. E-Learning for medical education in Sub-Saharan Africa and low-resource settings: viewpoint. *J Med Internet Res.* (2019) 21:e12449. doi: 10.2196/12449

23. NIH. *Guiding principles for ethical research. NIH clinical research trials and you.* Available online at: https://www.nih.gov/health-information/nih-clinical-research-trials-you/guiding-principles-ethical-research (accessed March 10, 2021).

24. Spiegelhalter D. *The Art of Statistics, Learning From Data, Pelican Books.* Penguin Random House