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

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 genome sequence data from South Africa and Nigeria can be used to advance understanding of viral dynamics in Africa. Project 2 examines the effectiveness 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. Project 3 investigates how to use geographic, mobility and demographic factors to forecast disease spread.

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 interventions.
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. 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 materials based on the FAIR principles for scientific data management and stewardship.

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. 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.](https://doi.org/10.17159/sajs.2023/14659)
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](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.

**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


