Harnessing big heterogeneous data to evaluate the potential impact of HIV responses among key populations in Sub Saharan Africa: The Boloka Data Repository Initiative

Refilwe Nancy Phaswana-Mafuya, Edith Phalane, Amrita Rao, Kalai Willis, Katherine Rucinski, Alida Voet, Amal Abdulrahman, Claris Siyamayambo, Betty Sebati, Mohlago Seloka, Musa Jaiteh, Lucia Olifant, Katharine Journeay, Haley Sisel, Xiaoming Li, Bankole Olatosi, Hikmet Neset, Prashant Duhoon, Francois Wolmarans, Shiferaw Yeganew, Lifutso Motsieloa, Mashudu Rampilo, Stefan Baral

Submitted to: JMIR Research Protocols
on: June 25, 2024

Table of Contents

Harnessing big heterogeneous data to evaluate the potential impact of HIV responses among key populations in Sub Saharan Africa: The Boloka Data Repository Initiative

Refilwe Nancy Phaswana-Mafuya[1]; Edith Phalane[1]; Amrita Rao[2]; Kalai Willis[2]; Katherine Rucinski[2]; Alida Voet[2]; Amal Abdulrahman[2]; Claris Siyamayambo[1]; Betty Sebati[1]; Mohlago Seloka[1]; Musa Jaiteh[1]; Lucia Olifant[1]; Katharine Journeay[2]; Haley Sisel[2]; Xiaoming Li[3]; Bankole Olatosi[3]; Hikmet Neset[4]; Prashant Duhoon[4]; Francois Wolmarans[5]; Shiferaw Yeganew[6]; Lifutso Motsieloa[7]; Mashudu Rampilo[7]; Stefan Baral[2]

[1]South African Medical Research Council/University of Johannesburg (SAMRC/UJ) Pan African Centre for Epidemics Research (PACER) Extramural Unit University of Johannesburg Johannesburg ZA
[2]Key Populations Program, Center for Public Health and Human Rights, Johns Hopkins Bloomberg School of Public Health Johns Hopkins University Baltimore US
[3]Big Data Health Science Center, Arnold School of Public Health University of South Carolina Columbia US
[4]Engineering and Computing, Integrated Information Technology University of South Carolina Columbia US
[5]Technology Architecture & Planning University of Johannesburg Johannesburg ZA
[6]Department of Statistics, Faculty of Science University of Johannesburg Johannesburg ZA
[7]South African National AIDS Council (SANAC) Pretoria ZA

Corresponding Author:
Refilwe Nancy Phaswana-Mafuya
South African Medical Research Council/University of Johannesburg (SAMRC/UJ) Pan African Centre for Epidemics Research (PACER) Extramural Unit
University of Johannesburg
40 Bunting Road
Auckland Park
Johannesburg
ZA

Abstract

Background: Key Populations (KPs), including gay men and other men who have sex with men, female sex workers, transgender persons, people who use drugs, and incarcerated persons, have a higher risk of human immunodeficiency virus (HIV) acquisition and transmission than the general population. Currently, there is no centralized data repository where KP HIV surveillance and programming data is gathered and stored in South Africa. Data on KPs are being collected on a smaller scale by numerous stakeholders and managed in silos; hence, there exists an opportunity to harness these variety of data sources for evaluating the potential impact of HIV responses among KPs in South Africa.

Objective: To leverage and harness Big Heterogeneous Data on HIV among KPs and rigorously harmonize as well as analyze it to inform a targeted HIV response for greater impact in Sub-Saharan Africa (SSA).

Methods: The Boloka data repository initiative has five main stages. There will be engagement of a wide range of stakeholders throughout South Africa to develop meaningful partnerships to both facilitate acquisition of data and to generate ideas around a base structure for the Boloka data repository (stage 1). Through these engagements, different data types including programmatic, and research data will be collated (stage 2). The data will be filtered and screened to enable high quality analyses (stage 3). The collated data will be stored in the Boloka data repository (stage 4). The established Boloka data repository will be made accessible for stakeholders and authorized users (stage 5).

Results: The stakeholder analysis process is underway and the Excel project tracking tool listing potential stakeholders, their contact details, and their level of engagement with data (e.g., district, provincial, and national) is in place. A series of engagements have been carried out with the respective organizations; the engagements are at different stages. The research team has identified and begun meeting with numerous data stakeholders and developed meaningful research-practice partnerships to facilitate collaboration and data sharing; several collaborative agreements are underway

Conclusions: A truly “complete” data infrastructure that systematically and rigorously integrates empiric, contextual, observational, and programmatic data for KPs will not only improve our understanding of local epidemics but will also improve HIV interventions and policies. Further, it will inform future research directions, and become an incredible institutional mechanism for epidemiological and public health training.

(JMIR Preprints 25/06/2024:63583)
DOI: https://doi.org/10.2196/preprints.63583

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Protocol Paper

Refilwe Nancy Phaswana-Mafuya[1,2,3], Edith Phalane[1,2], Amrita Rao[4], Kalai Willis[4], K Rucinski[3], K. Alida Voet[4], Amal Abdulrahman[4], Claris Siyamayambo[1,2], Betty Sebati[1,2], Mohlago Seloka[1,2], Musa Jaiteh[1,2], Lucia Olifant[1,2], Katharine S. Journeay[4], Haley I. Sisel[4], Xiaoming LI[5], Bankole Olatosi[5], Hikmet Neset[6], Prashant Duhoon[6], Francois Wolmarans[7], Shiferaw Yeganew[8], Lifutso Motsieloa[9], Mashudu Rampilo[9], Stefan D. Baral[4]

[1] South African Medical Research Council/University of Johannesburg (SAMRC/UJ) Pan African Centre for Epidemics Research (PACER) Extramural Unit

[2] Department of Environmental Health, Faculty of Health Sciences, University of Johannesburg

[3]Arnold School of Public Health, Department of Health Services Policy Management, University of South Carolina, Columbia SC

[4]Key Populations Program, Center for Public Health and Human Rights, Johns Hopkins Bloomberg School of Public Health

[5]Big Data Health Science Center, Arnold School of Public Health, University of South Carolina, Columbia SC

[6]Engineering and Computing, Integrated Information Technology, University of South Carolina, Columbia SC

[7]Technology Architecture & Planning, University of Johannesburg

[8]Department of Statistics, Faculty of Science, University of Johannesburg

[9] South African National AIDS Council (SANAC)

Corresponding author:

Prof Refilwe Nancy Phaswana-Mafuya

SAMRC/UJ PACER Extramural Unit

Email: refilwep@uj.ac.za

PO Box 524, Auckland Park, 2006

Harnessing big heterogeneous data to evaluate the potential impact of HIV responses among key populations in Sub Saharan Africa: The Boloka Data Repository Initiative

Abstract

Background: Key Populations (KPs), including gay men and other men who have sex with men, female sex workers, transgender persons, people who use drugs, and incarcerated persons, have a higher risk of human immunodeficiency virus (HIV) acquisition and transmission than the general population. Currently, there is no centralized data repository where KP HIV surveillance and programming data is gathered and stored in South Africa. Data on KPs are being collected on a smaller scale by numerous stakeholders and managed in silos; hence, there exists an opportunity to harness these variety of data sources for evaluating the potential impact of HIV responses among KPs in South Africa.

Objective: To leverage and harness Big Heterogeneous Data on HIV among KPs and rigorously harmonize as well as analyze it to inform a targeted HIV response for greater impact in Sub-Saharan Africa (SSA).

Methods: The Boloka data repository initiative has five main stages. There will be engagement of a wide range of stakeholders throughout South Africa to develop meaningful partnerships to both facilitate acquisition of data and to generate ideas around a base structure for the Boloka data repository (stage 1). Through these engagements, different data types including programmatic, and research data will be collated (stage 2). The data will be filtered and screened to enable high quality analyses (stage 3). The collated data will be stored in the Boloka data repository (stage 4). The established Boloka data repository will be made accessible for stakeholders and authorized users (stage 5).

Results: The stakeholder analysis process is underway and the Excel project tracking tool listing potential stakeholders, their contact details, and their level of engagement with data (e.g., district, provincial, and national) is in place. A series of engagements have been carried out with the respective organizations; the engagements are at different stages. The research team has identified and begun meeting with numerous data stakeholders and developed meaningful research-practice partnerships to facilitate collaboration and data sharing; several collaborative agreements are underway

Conclusions: A truly “complete” data infrastructure that systematically and rigorously integrates empiric, contextual, observational, and programmatic data for KPs will not only improve our understanding of local epidemics but will also improve HIV interventions and policies. Further, it

will inform future research directions, and become an incredible institutional mechanism for epidemiological and public health training.

Keywords: Key populations; HIV, Sub-Saharan Africa; Big heterogenous data; Data repository

Introduction

In 2023, the world is at a critical juncture in the HIV response, counting down seven years towards the global goal of ending AIDS as a public health threat by 2030 [1,2]. Despite the investment and focus on addressing HIV/AIDS, it remains a significant public health threat with persistent prevention and treatment challenges globally and in South Africa, as reflected in the South African National Strategic Plan, 2023-2028 [3]. South Africa has the largest HIV epidemic in the world, with about eight million people living with HIV (PLHIV), which represents approximately 1 in 5 of the estimated 38.4 million PLHIV globally in 2022 [1,4]. While there has been a steady decline of 30.5% (n = 198,311) new infections in the last five years, the country still has an unacceptably high HIV incidence (South African National HIV, Prevalence, Incidence, Behaviour and Communication Survey [3-5]. South Africa has the largest HIV treatment program in the world to meet the treatment needs of the highest proportion of PLHIV [1].

Key populations (KPs), including female sex workers (FSW) and their clients, gay men and other men who have sex with men (MSM), transgender people, people who use drugs (PWUD), and incarcerated persons face a disproportionate risk of HIV acquisition and onward transmission [6]. As a result of unmet prevention and treatment needs, 51% of new HIV infections are acquired by KPs and their sexual partners, despite making up approximately 1.5% of the total adult population in Sub-Saharan Africa (SSA) [1]. The estimated prevalence of HIV was 59.5% among FSW and 29.7% among MSM in South Africa in 2020 [3,4]. Given social network dynamics, the overall impact of unmet needs of KPs on onward transmission may be even greater. Other modelling studies have a higher risk of onward transmission due to the unmet prevention and treatment needs among KPs using the transmission population attributable fraction over time ($tPAF_t$) [7]. This demonstrates the need for the specificity of the HIV response in characterizing and addressing heterogeneity in onward transmission for a more significant impact in reduction of new infections. This disproportionate risk of HIV is driven by discrimination, stigma, and criminalization of behavior and/or identity [8,9]. These same factors alongside social network dynamics often make collecting data on HIV burden and specific HIV prevention and treatment needs challenging, due to distrust of

institutions and fear of poor treatment, arrest, or violence [10-12].

In South Africa, where there is robust HIV surveillance compared to other African countries, there is currently no specific mechanism or centralized system to gather and monitor KP data. The existing HIV surveillance systems include the District Health Information System (DHIS) and the Integrated Electronic Registers (Tier.Net), which both collect patient-level HIV data but have no identifiers for KPs, making it difficult to disaggregate data by these sub-populations [13]. The inability to disaggregate information among KPs can lead to misallocation of resources and services, perpetuating health inequities. This lack of adequate data can also lead to underestimation of the disproportionate risk of onward transmission and ultimately missed opportunities for targeted approaches that can lead to a significant reduction of new HIV cases [7,14,15]. Despite this lack of a centralized system, data on KPs are being collected on a smaller scale by numerous stakeholders, including program implementers, the government, academic partners, and others. There exists an opportunity to harness these different data sources for integration and in-depth analyses.

In generalized epidemic settings, there has been a tendency to focus the HIV response on the general population rather than KPs [16-18] The scientific justification for the project presented here is that continued reliance on non-specific population-based approaches to guide programs has limited the broader impact of the HIV response in a generalized epidemic setting like South Africa. At the same time, there are limited data to determine the extent to which a KP-tailored HIV response could reduce HIV incidence in South Africa [6,7,19-21]. A more specific HIV response will likely optimize utilization of limited resources [7,20]. The proposed work can support program-implementers, funders, and policymakers to make well-informed choices as to where, what, and on whom to prioritize, and how to deliver effective HIV control programs to maximize health benefits at a population level. An effective control of the HIV epidemic in South Africa requires a focus on KPs [22].

To address the identified gaps in the field, we are in the process of designing and developing a big data platform called “The Boloka data repository” and harmonization of disparate data from multiple data sources. The Boloka data repository seeks to store a diverse range of data including empiric systematically and rigorously, contextual, observational, and programmatic data, to improve understanding of HIV acquisition and transmission among KPs. Additionally, the Boloka data repository seeks to evaluate the potential impact of HIV responses in South Africa in the context of a

generalized epidemic setting. Boloka is a Sepedi, Sesotho, and Setswana word which means to *store or keep.* In this case, we will store or keep big heterogeneous data. These data, including HIV-related and relevant data for KPs from the year 2000 onward in South Africa, can be used to inform policy and programming. By harnessing big heterogeneous data, more data-driven, and ultimately more effective HIV response strategies can be generated.

In this protocol paper, we describe the process for developing the Boloka data repository for South Africa specifically. It is envisaged that once the Boloka data repository has been developed, it can be adapted to other countries throughout SSA. The current study has three specific objectives: First, to build a Boloka data repository to handle data on KPs. Second, to collate available HIV-related data for KPs in South Africa from 2000 onwards. Third, to make the data user-friendly for utilization by stakeholders and authorized users.

Methods

The Boloka data repository initiative will be executed in the South African Medical Research Council/ University of Johannesburg (SAMRC/UJ) Pan African Centre for Epidemics Research (PACER) Extramural Unit, which is an SAMRC extramural unit part of over 40 research centers and institutes, and 20 prestigious national research and industry-funded chairs that UJ is hosting. It will be one of the flagship initiatives that will be supported under the Global Excellence and Stature strategic initiative [23]. The protocol will receive support from Offices of the Deputy Vice Chancellor: Research and Internationalisation, Research, and Innovation Support; Strategic Initiatives & Administration; Internationalisation; and the Faculty of Health Sciences. The UJ Research Intelligence Unit in the research office will provide bibliometric and research performance analyses to support reporting on research progress, impact visibility, and footprint as well as to inform strategic research decisions. These analyses will provide opportunities for capacity-building among academic trainees and stakeholders to enhance ability to gather and analyze epidemiological and program data as well as other data types and develop skills to monitor and evaluate programs and policies.

The Boloka data repository seeks to optimize HIV data systems to mitigate the epidemic consequences of the unmet HIV prevention and treatment needs for KPs. To achieve the stated objectives, this study project will be developed in the following five key stages (Figure 1):

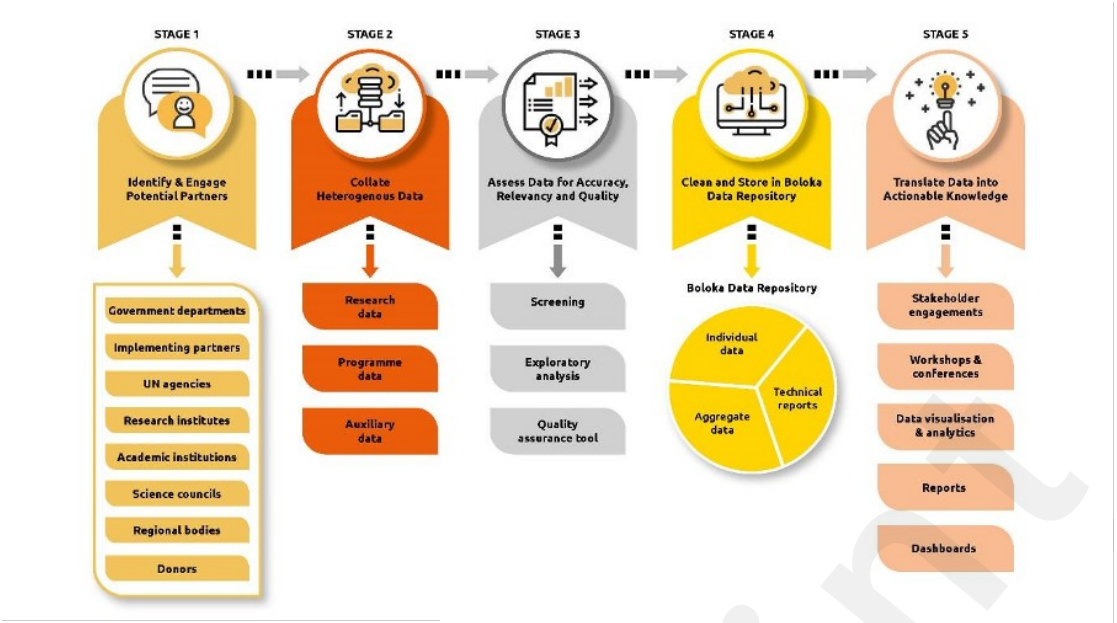


Figure 1: Five stages of the Boloka data repository initiative [24]

Stage 1: Engage key stakeholders to develop meaningful data partnerships

This study will utilize a diverse range of data sources. To facilitate the collation and organization of these diverse data sources, we will engage stakeholders at multiple levels representing myriad institutions. A stakeholder analysis will be conducted each year to understand the landscape of individuals and institutions relevant to KPs in South Africa. This analysis process will consist of a stakeholder mapping activity to identify and describe key stakeholders utilizing a project tracking tool in Excel. A structured interview guide will be developed for engagements with stakeholders. This will build upon the investigators' existing network of HIV practitioners, policymakers, program implementers, and researchers. Additionally, new partners will be identified and engaged at scientific conferences, workshops, technical working meetings, and related engagements. Appendix 1 shows an overview of current and potential partners for the Boloka data repository.

A key stakeholder in this work is SANAC, which will act as a partner in both the sharing of data and the facilitation of connections to additional stakeholders. Ultimately, SANAC will ensure that the findings from this study and the ongoing use of the repository are incorporated into the national HIV response strategy in South Africa.

The Boloka protocol will take a transdisciplinary participatory approach to stakeholder engagement to develop meaningful partnerships and ensure collaboration between researchers and stakeholders

[25]. Throughout the project lifespan, the participatory approach will improve buy-in, project performance, co-ownership, and sustainability [26,27]. The engagements will be guided by key principles of mutual trust, sharing, transparency, and responsibility [26,27], all with the common goal of improving the understanding of HIV dynamics for KP. Stakeholders will be given an opportunity to give input at all stages about all components of the Boloka data repository, including which data are collected, how they are stored, data confidentiality, data privacy, data security, and how others may access them going forward using the approved University of Johannesburg (UJ) data management process. The ongoing engagement of stakeholders will ensure that their input is incorporated in the study protocol and this input guides the development of research questions that will be answered using data from the Boloka data repository. This will ensure that the questions asked have utility for those directly involved in the HIV response.

The UJ Data Processing Agreement (DPA) and data partnership agreement or data request form will be signed between UJ and the respective data partners [28]. This DPA will govern the rights and duties of the parties when processing personal information in accordance with the "protection of personal information Act number 4 of 2014", henceforth referred to as the POPI Act, and any other applicable data protection, security, storage, regulation, and legal requirements [29]. The data partnership agreement outlines the obligations, expectations, and roles of both parties, the type(s) of data, as well as utilization and expected outcomes from the shared data. Through these agreements, it will be ensured that the collection, storage, use, disclosure, transfer, disposal, and other processing of any personal information is in line with the prescribed data protection law.

Stage 2: Acquire and collate heterogeneous data

The Boloka data repository will leverage and collate available and diverse data from various sources across places, times, and populations (Appendix 2). The team reviewed the UNAIDS Global Monitoring tool and the NDOH National Indicator Dataset form used for national HIV reporting to identify primary and secondary indicators in line with national and regional priorities. The team also reviewed existing validated questionnaires/instruments from various data partners to further refine priority indicators. From this, an indicator list or broader categories of HIV-related interest areas was formed (Figure 2).

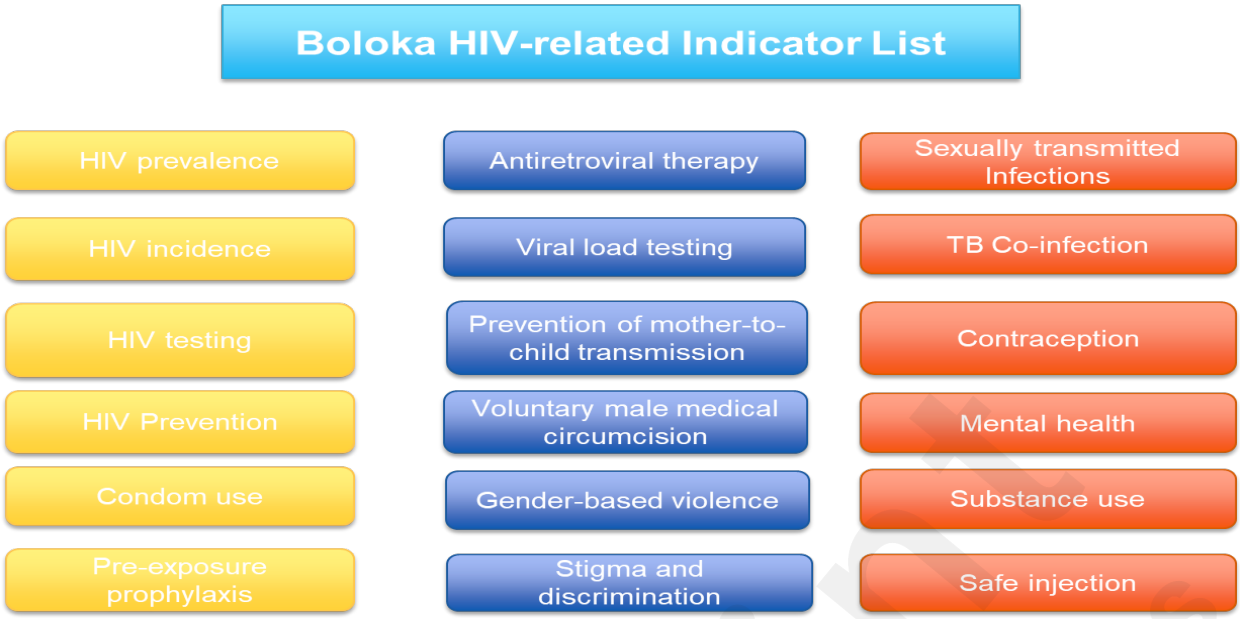


Figure 2: Boloka HIV-related Indicator list

Where possible, data will be disaggregated at individual, sub-district, or district levels to enable advanced analyses that are not possible with higher-order aggregated data. This will help us develop an in-depth understanding of various subsets of the populations within the larger datasets. Where data are aggregated, we will seek data disaggregated by sex, gender, age groups, socio-economic status, geo-location, facility type, and temporal factors, to understand a range of HIV indicators, including heterogeneity of HIV risk and burden, engagement in HIV services (treatment cascade, pre-exposure prophylaxis uptake and continuation), and population size estimates, among others. The process for requesting and accessing data varies based on the data type and institution. As such, the process to acquire and collate the data is discussed below for the different data types.

Research Data

For open-access data, which is typically aggregated or de-identified individual data, we will adhere to procedures set by the respective institutions to obtain access. This may require the submission of designated data access request forms prior to gaining direct access to the dataset. For data which are only available upon request, we will contact the data partner to understand their specific process(es) for accessing data. This typically requires the completion of a data access request form and the sharing of relevant information such as the project proposal, ethics approval, and timelines for analysis. For institutions such as the National Department of Health (NDoH), district health information data is made available for public use upon request and completion of their data user agreement (DUA) forms. The data or indicators requested should be aligned with the approved

National Indicator Data Set, which outlines the data elements and indicators collected by the NDoH. In such cases, the request and DUA forms will be completed to access data. For institutions in possession of data, which are not typically available for public access (e.g., implementing partners), we will seek to develop formal partnership agreements and additionally utilize the UJ DPA to ensure adherence to the POPI Act in terms of data privacy and security prior to any data sharing. Once agreed upon, a formalized partnership agreement approved by the legal team of both parties and UJ DPA will be signed by the two parties. These agreements will be reviewed every three years by the two parties.

Program data

For institutions in possession of data, which are not typically available for public access (e.g., implementing partners), we will seek to develop formal partnership agreements and additionally utilize the UJ DPA to ensure adherence to the POPI Act in terms of data privacy and security prior to any data sharing. Once agreed upon, a formalized partnership agreement approved by the legal team of both parties and UJ DPA will be signed by the two parties. These agreements will be reviewed every three years by the two parties.

Stage 3: Screen data for relevance and quality

Once data are acquired and determined to be eligible according to the inclusion and exclusion criteria (Table 1), these data will go through a process of screening and assessment for accuracy, quality, completeness, and consistency before their inclusion in the Boloka data repository (Figure 1). All data types will be screened and filtered for relevance and inclusion by a member of the research team. Due to limited information and validated tools on how to specifically check for relevance and quality for some of the data types [30], we will align and adopt guidelines on assessing data quality such as the Framework for Data Quality developed by the Federal Committee on Statistical Methodology (FCSM) [31] and the Information Quality Assessment Framework where applicable [32]. Below are further details on the data screening process by the different data types.

Table 1: Inclusion and Exclusion Criteria

Parameter	Inclusion Criteria	Exclusion Criteria
Study Topic	HIV, Key Populations	Non-HIV
Study Area	South Africa	Non-South Africa
Language	English	Non-English

Time frame	Collected in the year 2000 or later	Collected before the year 2000

Research data

For research data, a multi-step quality assessment process will be carried out using the Global HIV Quality Assessment Tool (Appendix 3) [33]. This tool will be used to review and verify the suitability and quality of research data sources in terms of study design and implementation along with criteria for HIV indicators; specifically, prevalence, incidence, engagement in the HIV care continuum, and population size estimates [33]. There will be close supervision and checks to minimize errors by utilizing an honest data broker mechanism to manage and maintain datasets. Two individuals (research assistants) will perform the initial quality assessment. Any discrepancies identified will be addressed by a third assessor (Project Manager/Principal Investigator). Exploratory analyses will be conducted to understand the nature of the data and to identify outliers and missing data that will be cleaned to ensure data quality and accuracy. This will be an iterative process to ensure that we leverage and assemble the best available data. The UJ Statistical Consultation Services and the UJ Faculty of Science, Statistical Departments will provide statistical expertise on the proposed work.

Program data

It is important to highlight that program data present their own unique data quality challenges. Program data are typically aggregated to administrative units and information on individuals is generally not available, hence it is not always feasible to link program exposure directly to an outcome [34]. In terms of screening the program data for inclusion into the Boloka data repository, the WHO's "Data quality review: a toolkit for facility data quality assessment—Module 1: Framework and metrics" will be adapted for use [35]. This framework is made up of four dimensions. For this study, dimensions 1 and 2 will be utilized: completeness and timeliness of data (dimension 1), and internal consistency of reported data (dimension 2). Dimension 1 speaks to the completeness of specific data elements, and the data elements and indicators collected from, for example, the NDoH and implementing partners, will be assessed, and checked for completeness [35]. Dimension 2, i.e. internal consistency of the data relates to the coherence of the data being evaluated. In this regard, four metrics of internal consistency will be used [35]: (i) presence of outliers; (ii) consistency over time; (iii) consistency between indicators; and (iv) consistency of reported data and original records. The presence of outliers will involve checking if a data value in a series of values is extreme with respect to the other values in the series. Consistency over time will be assessed for credibility of reported results for selected program indicators and elements in terms of the history of

reporting of the indicators. Trends will be evaluated to ascertain if reported values are extreme in relation to other values reported during the year or over several years. Consistency between indicators, such as program indicators, which have a predictable relation are to be assessed for whether the expected relation exists between those indicators. In other words, this process assesses if the observed relation between the indicators, as depicted in the reported data, is that which is expected. Consistency of reported data and original record will involve an evaluation of the reporting accuracy for selected indicators, facilitated by a review of source documents in facilities.

In this initial screening, articles, reports, and datasets will be excluded if their scope does not adhere to the inclusion criteria mentioned in Table 1. Data that has a focus on the general population will be included in addition to data with a focus on KPs, as the general population provides important KP contextualization. Furthermore, data that has been empirically collected will be prioritized. Estimates and modeled data will be included but carefully reviewed to ensure they meet quality standards and de-duplicated within the data sets. Data that are deemed to be “low” quality due to the inherently lower quality of certain data types will not be excluded; a field/variable will be added to denote the quality of data in the data repository.

Stage 4: Clean and store data in data repository

Data will be extracted from data sources and put into a staging area, prior to being transformed, and loaded into the data repository: Figure 3. We will use the Amazon Web Services (AWS) platform or similar (Figure 3), which will enable the provision of much-needed timely information at a level and scale that will improve understanding of HIV heterogeneities. All structured data will be pre-processed and normalized into a standardized format using the Boloka Data Harmonisation Tool [36]. The repository will be flexible and updatable for structured data by design. All unstructured data will be stored in NoSQL data store. Pre-processing will entail data cleaning, transformation, and integration to make the data complete for analysis. Data management will involve assembling multiple big, existing HIV data sources and datasets, capturing, pre-processing, cleaning, integrating, and building them into the Boloka data repository.

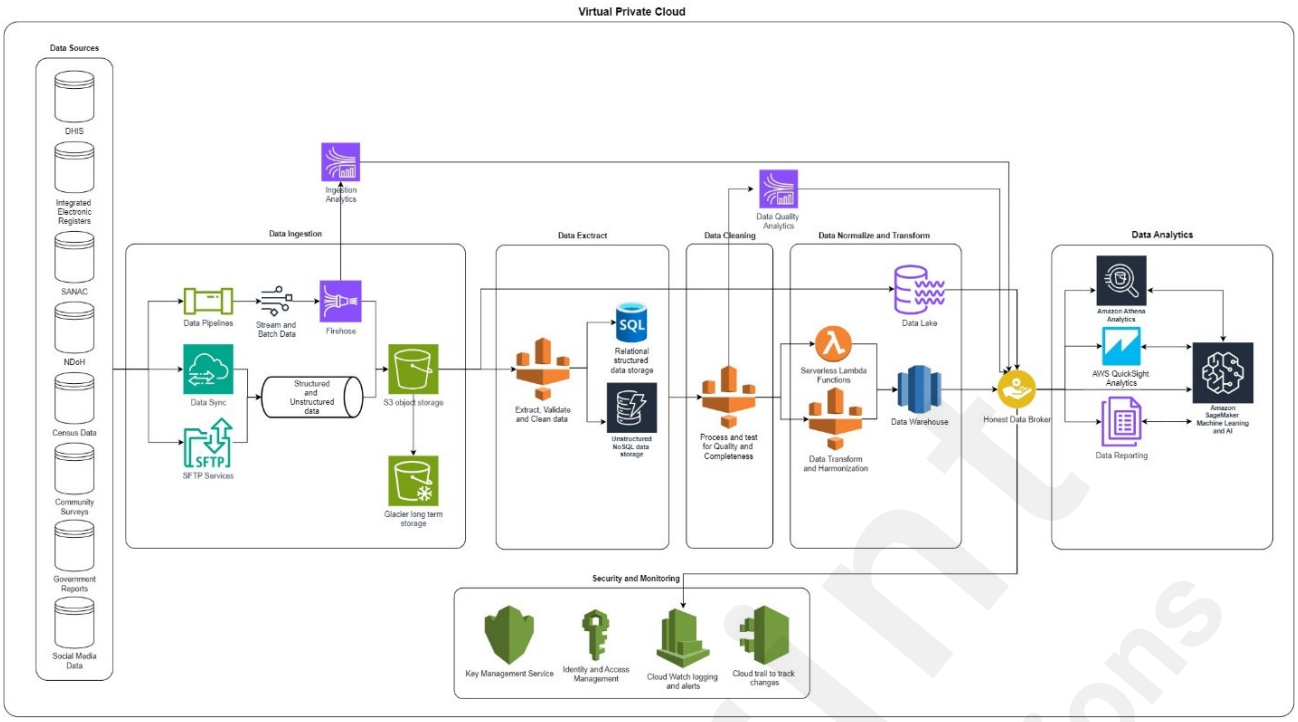


Figure 3: Envisaged Boloka Data Storage Process Flow

For disaggregated data, personal identifiable information will be encrypted and hashed to ensure anonymity and protect privacy. This role is also performed by the appointed honest data broker. Ideally, the Boloka data repository will not receive any identifiable personal information, but a procedure in accordance with the POPI Act will be put in place for handling this issue if such information comes with the data. The honest data broker mechanism will be set up to screen the data for compliance in terms of the latter before it passes to the Boloka data repository. The resulting data repository will be responsive to the size and scope of the collected data. It will be dynamic and flexible to accommodate diverse data that has a high number of dimensions.

The access to the repository will be secured with complex login credentials in conjunction with multi-factor authentication (MFA) to ensure compliance with the POPI Act and avoid any confidentiality breaches and contravention of human rights. The Boloka data repository will contribute to improved storage, retrieval, and access to KP data; linkages of surveillance data collection systems and other data collection efforts; integrated data reporting for monitoring national trends and patterns; recommendations for research, policy, and planning; as well as changes in current and future research. The Boloka data repository will provide a richer empirical basis for policy and program debates in a relatively neglected area of health research in the country. The data, along with its analyses and research outcomes, can form the basis of regularly reported statistics by

stakeholders, such as the Centers for Disease Control and Prevention and WHO. The Boloka data repository will be maintained up to date in order to serve as a real-time sustainable resource to guide HIV planning, resource allocation, policy formulation, and programming.

Stage 5: Translate data into actionable knowledge

Knowledge created from available data will be applied towards addressing real-world challenges. To ensure that findings are incorporated into program and policy decisions in a timely manner, the study will establish a feedback loop with stakeholders and authorized users throughout the project lifespan. Knowledge sharing will be done through dissemination workshops, consultative meetings, reports, dashboards, data visualization, and other methods deemed appropriate. This will support open communication and two-way feedback between researchers and stakeholders (Appendix 4). This participatory process will contribute towards improved utilization of available data and narrowing the gap between science and practice. Various forms of analysis will be planned in partnership with stakeholders to answer research questions that are relevant to their respective organizations and national priorities. Ultimately, the translation of research will seek to strengthen health systems and therefore improve health outcomes throughout South Africa, specifically for KPs and other vulnerable populations. Our goal is to make the Boloka data repository an accessible tool that can be used by local, national, and international stakeholders towards more effective and efficient healthcare management, prevention, and programming.

The UJ Strategic Communication Department designated official will promote and market the initiative using UJ's communication technologies, media, and platforms. They will promote public understanding of study and research outputs throughout. This will include articles in the daily press, magazines, and other popular media; public lectures; and interviews or programs on TV or print media.

Ethical considerations

An ethics approval has been secured from the University of Johannesburg, Faculty of Health Sciences, Research Ethics Committee (REC-1504-2022) (Appendix 5).

Results

Progress to date

As of December 2022, the protocol was approved and funded by the South African Medical Research Council following external reviews. Subsequently, the study received ethics approval from UJ,

Faculty of Health Sciences, Research Ethics Committee. The research assistants, students, and post-doctoral fellows who will partake in the creation of the data repository and subsequent analyses were recruited, onboarded, and received online training and resources in research ethics evaluation (TRREE) on ethics and regulation of health research involving human participants, Management and Analysis of Data in Epidemiology (MADE), data types, and structures. The team will continue to receive ongoing professional development training, including sensitivity training for KPs and training on data privacy, specifically regarding the POPI Act and AWS functionalities. The progress to date is described below according to the five stages.

Stage 1: Identify and engage key stakeholders to develop meaningful data partnerships

The stakeholder analysis process is underway and the Excel project tracking tool listing potential stakeholders, their contact details, and their level of engagement with data (e.g., district, provincial, and national) is in place. The development of the stakeholder interview guides has been in progress.

Research data

The research team engaged with the HSRC on securing the open access data from five (2002, 2005, 2008, 2012, and 2017) South Africa National HIV Prevalence, Incidence, Behavior and Communication Survey (SABSSM). Since the data is open access, no signing of DPA was required to access and secure the data. Additionally, engagement was made with researchers at HSRC on the possibility of using the research data. At this stage, engagement is still ongoing and an example of the UJ DPA has been shared for consideration by the stakeholders. An engagement with NDoH was done in 2022 and a DUA is in place to share HIV-related routine program data with NDoH.

Program data

The research team had a series of meetings with SANAC regarding the Boloka data repository. These meetings have established a strong partnership and system for the coordination of data sharing partnerships. The partnership with SANAC is essential to ensuring that the Boloka data repository is utilized to guide national HIV strategy related to KPs. This strong relationship will also serve to make the Boloka data repository complementary, rather than duplicative, of SANAC's ongoing efforts to develop *The Situation Room*, a central data repository for program data in South Africa that will enable government departments, program implementers, researchers, and other stakeholders to effectively use data sets in real time for decision making. The Situation Room, still in its infancy stages, seeks to enable data checks and balances, dynamic visualization and sharing of the national,

provincial, and district data, to monitor progress towards reaching set targets. The conversations with SANAC will allow the Boloka data repository to be a resource for SANAC with their ongoing efforts.

SANAC introduced the research team to its implementation partners to acquire potential data partners, namely Beyond Zero (BZ), Networking HIV and AIDS Community of Southern Africa (NACOSA), Sex Worker Education Advocacy Taskforce (SWEAT), Sisonke, and AIDS Foundation South Africa (AFSA). BZ is a non-profit organization (NPO) implementing partner mainly supported by the Global Fund. It provides HIV, TB, and STI services to MSM, transgender people, adolescent girls, and young women (AGYW), and other high-risk communities in seven provinces [37]. Like BZ, NACOSA is a prominent NPO that works with 2,500 community-based organizations (CBOs) in seven provinces to deliver services to FSW, PWUD, and AGYW [38]. SWEAT focuses on advocating for the rights of sex workers and promoting the health for people who choose to sell sex in South Africa [39]. Sisonke is a national alliance of sex workers in South Africa [40]. AFSA focuses on supporting adolescents, young persons, and sex workers by addressing the social and structural drivers of HIV [41]. The Perinatal HIV Research Unit (PHRU) is an affiliate of University of the Witwatersrand and focuses their research on prevention of mother-to-child HIV transmission and on multiple aspects of HIV prevention, treatment, and care, including medical and social research [42].

A series of engagements have been carried out with the respective organizations; the engagements are at different stages. The research team has identified and begun meeting with numerous data stakeholders and developed meaningful research-practice partnerships to facilitate collaboration and data sharing; several collaborative agreements are underway. BZ has been engaged since September 2022, and a final copy of the UJ/BZ data sharing agreement is currently under review by the UJ Legal Team before it is signed by both parties. With NACOSA, a series of meetings occurred early 2023 to identify the data available and what indicators would be requested for the repository. As of July 2023, a UJ Data Partnership Agreement was shared with the organization and is currently being reviewed by the legal entity at NACOSA. As of July 2023, a data partnership agreement is in progress with PHRU. For this partnership, the PHRU data sharing agreement is being developed by PHRU legal department and will be shared. With AFSA, the agreement is in the early stages of communication. Regarding Higher Health, engagement on the data being collected has been done and the next is to discuss DPA. The remaining implementing partners SWEAT and Sisonke have yet to be engaged. In addition to engagements with SANAC and its implementing partners, the research

team has consulted with other community leaders who work/worked with community-based organizations like treatment action campaign.

Stage 2: Acquire and collate *heterogeneous* data

There has been significant progress in securing program data, published research data, and technical reports, particularly from national stakeholders. The security of the acquired data is prioritized during this interim period; thus, the data have been stored in a secure staging area and will later be moved to appropriate data storage software.

Research Data

Open access data from five of the six (2002, 2005, 2008, 2012 and 2017) population-based multi-stage cluster cross-sectional of up to 85,000 randomly selected households in South Africa has been secured from the Human Sciences Research Council (HSRC). The HSRC has four levels of data access, namely: open, registered, restricted, and project team. The SABSSM data is within the restricted access level. The data was accessed by completing an online data request form detailing the name of the project, brief description of the intended use, and the expected date of project completion. Access is subject to approval from the owners, funders, or depositors of the data. An email notification from the HSRC with login credentials to the data repository was received and was used to confirm that access has been authorized. Users can export the datasets post request to the honest data broker in their format of choice including Stata, SPSS, or Excel, along with dataset supporting documents such as questionnaires and code books in PDF format. The surveys provide data on HIV incidence, prevalence, antiretroviral therapy (ART), viral load suppression, drug resistance, risk behaviors and HIV care, among others. The Key Population Implementation Science (KPIS) data on HIV testing and engagement in care was secured at Emory University for the degree purpose of postgraduate students. Further, data on HIV testing proficiency was generated by a postgraduate student from the health care facilities in Eastern Cape Province. Separate permission needs to be requested from the respective institutions to include the data set into the Boloka data repository.

Data from the Demographic Health Surveys (DHI) Program has been secured. The data was obtained from DHIS web portal. This required registration, and submission of data access request form along with justification of the request. These details included the project title, a summary of the project, the populations targeted, the researchers involved, and the organization. A request for HIV routine data was made through an email to an NDoH representative responsible for managing requests and access

to data. A data request form and signed DUA were completed and submitted to NDoH. Information on the project description, purpose, and use of the data, list of indicators, as well as expected outcome, were provided on the data request. The data being requested needs to align with the list of indicators and elements listed in the official National Indicator Data Set. The data includes the following HIV indicators: HIV testing and HIV prevention; antiretroviral therapy (ART) (i.e., ART initiation, ART rate, viral load & CD4 cell count testing, ART type, ART adherence); sexually transmitted infections; maternal and neonatal (antenatal HIV test, ART initiation, ART adherence, live birth, infant HIV test); management of in-patients and management of primary health care facility.

There are plans to use data from the 1173 NDoH High Transmission Area sites across South African provinces based in communities that function like clinics and provide services to KPs. The NDoH's High Transmission Area program reaches key and vulnerable populations with HIV, tuberculosis, and sexually transmitted infections prevention and management services in key hotspots across the country. Consultative discussions have already been held with the Director of HIV/AIDS and STI and with the Deputy Director of HIV/AIDS Prevention and STI.

Program Data

This protocol will also leverage a partnership with the largest HIV service provider for key populations in South Africa, TB HIV Care, an NPO, building on a decade-long period of dynamic and exciting work; preliminary discussions have already taken place with the Chief Executive Officer. TB HIV Care focuses on preventing, finding, and treating HIV and TB across 7 provinces and 22 districts in South Africa [43]. Its KP program provides services to sex workers and PWUD. Further, data sharing agreements are being drafted and checked for BZ, NACOSA, and AFSA, as detailed above. Appendix 1 provides a comprehensive overview of potential data partners.

Stage 3: Assess data for accuracy, relevance, and quality

The data received in stage 2 went through screening and assessment for accuracy, completeness, and consistency before its placement in the staging area. The data received was screened and filtered for relevance and inclusion by a member of the research team using the FCSM and the Information Quality Assessment Framework [32] where applicable.

Research Data

SABSSM data has undergone a quality check to assess the completeness of the data received.

Utilizing the Needs Assessment Form, the DHIS data received has been checked to assess the relevance of the data to the project aims. The assessment checked that the data included the relevant indicators and if the data was complete. This process has been done by two researchers on the team.

Stage 4: Clean and store data in the Boloka data repository

The data have been placed in the staging area prior to being stored in the data repository. PACER will secure access to AWS. The access will grant UJ permission to utilize the AWS platform, along with access to the support tools and resources. Currently, PACER is in the process of exploring the proof of concept in compliance with POPI Act security standards and adhering to the best practices recommended by AWS business associate agreement (BAA). The Resources will be shared to provide access to AWS training materials. Through this platform, KP data will be harmonized into a centralized storage area that is managed and protected. All data is to be cleaned and converted into a standardized format to create a structured, flexible, and updatable data repository.

Stage 5: Translate data into actionable knowledge

Authorized users and stakeholders will have the capability to generate customized reports and export the data to applications such as Stata software for further dissemination. Authorized users and stakeholders will submit a data request prior to receiving log-in credentials to access the data repository. Initial secondary data analyses using analytic methods attuned to the structure of available data, including cross-sectional and longitudinal analyses, are being conducted to improve our understanding of HIV among KPs for a targeted response.

With access to the Boloka data repository, there will be opportunities for complex and important analyses. A suite of epidemiologic methods for multi-level analyses will be employed, applying approaches attuned to the structure of available data, including cross-sectional and longitudinal analyses. The proposed analyses will be a transdisciplinary collaborative effort to enable joint application of theories and methods to share conceptual frameworks, innovations, and best practices for solving public health problems. In this regard, the proposed analyses will be jointly finalized with program partners to identify priority research questions that will measure program progress, guide programmatic decision-making, and ultimately improve the response to HIV.

Data synthesis and initial analyses are being planned with a transdisciplinary team of infectious disease epidemiologists, public health scientists, statisticians, data scientists, data analysts, implementing partners, and data partners who will bring complementary expertise on global key populations insights, key population modeling, epidemiology, human rights, and the development of

a sustainable data platform.

The transdisciplinary team proposed in this protocol consists of a statistician from the UJ Statistics Department who will offer advanced statistical analysis expertise in identifying HIV predictors, determining associations of data and precision, as well as understanding HIV acquisition and transmission from various data sources iteratively. We have already worked with some of the data partners to determine the research agenda that targets policy and practice outcomes, i.e., determining common research questions, models, and methods to improve South Africa's HIV response. The PhD students who will utilize the Boloka data repository for their dissertations have begun writing their research proposals and seeking ethical approval.

Discussion

There are factors that are critical for the success of this project to achieve the envisaged medium- and long-term impacts as shown in Figure 4.

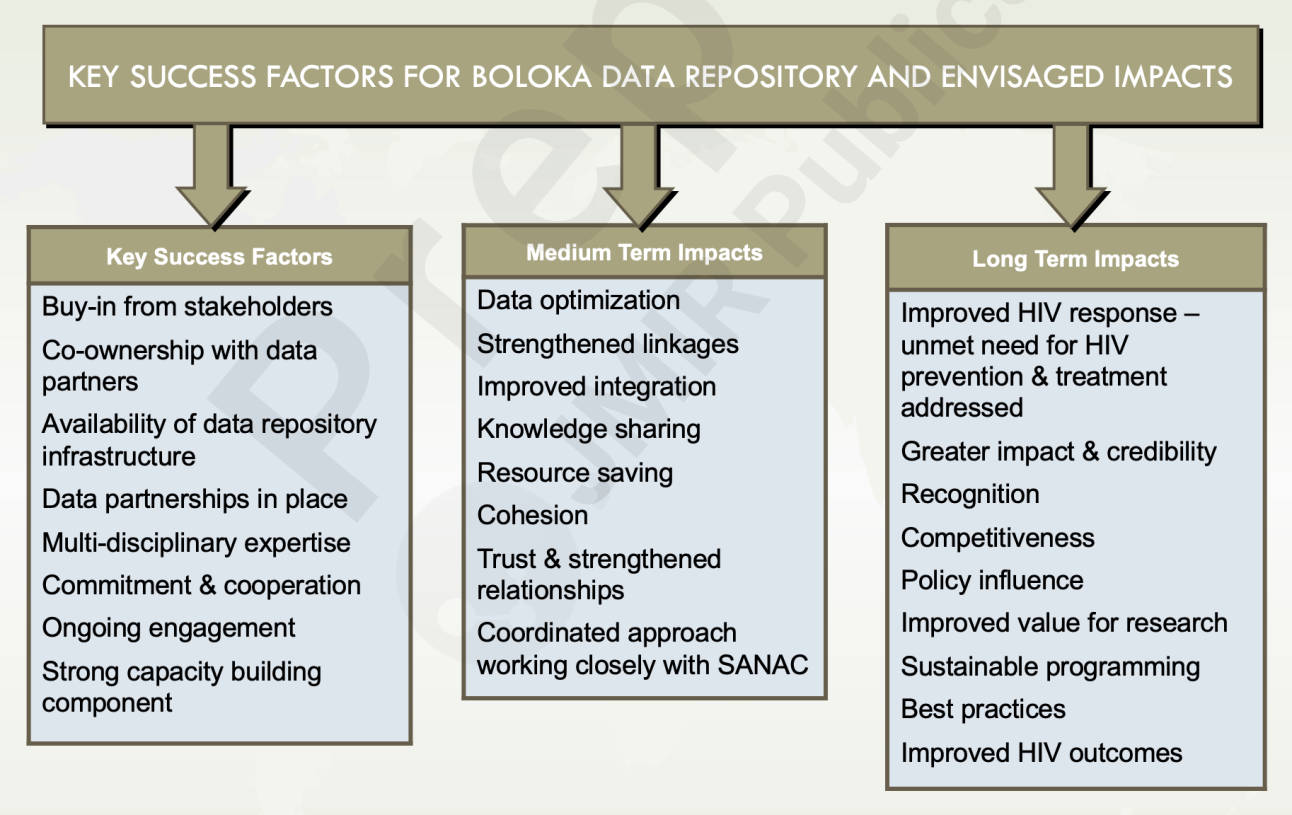


Figure 4: Key success factors for Boloka data repository and envisaged impacts

It is anticipated that there will be data challenges which are common across projects of this nature due to the POPI Act. There is already a mechanism that UJ has established to address these. Potential POPI Act related challenges have been highlighted in the UJ Data Processing Agreement which is

standard to all data driven projects. Alternative strategies to address each challenge/risk have been laid out in the below Table 2.

Table 2: Anticipated challenges and envisaged solutions

Anticipated Challenges	Envisaged solutions
Security and integrity in confidentiality of the Personal Information	Notification of each party of any data security bridge or incident. Restricted access to Personal Information using log-in username and password for each person requiring accessing the data.
Unauthorized disclosure	The parties will undergo POPI Act training
Compliance	Data sharing agreements will be signed with the respective data partners

The Boloka data repository will be a lasting resource for the country to guide regional policies and strategies to address HIV, particularly among KPs. Therefore, in addition to traditional research dissemination led by the research team and the proposed feedback loop (Appendix 4), the data repository will be widely accessible to all data partners and stakeholders. This co-ownership with stakeholders will enable alternative possibilities for analyses and translation of data to real-world settings. The knowledge, historical context, and applied understanding that stakeholders possess provide opportunities for transformative change. The protocol will provide the opportunity to document lessons learned into a knowledge base, which may be applied to other countries in SSA. Appendix 5 gives a summary of planned and executed milestones for this project.

Conclusions

We posit that a truly “complete” data infrastructure that systematically and rigorously integrates empiric, contextual, observational, and programmatic data for KPs will not only improve our understanding of local epidemics but will also improve HIV interventions and policies. Further, it will inform future research directions, and become an incredible institutional mechanism for epidemiological and public health training. Achieving epidemic control in South Africa necessitates moving beyond HIV data silos, harnessing under-utilized heterogeneous data, and conducting unconventional analyses. Creating a comprehensive and accessible data infrastructure inclusive of heterogeneous data will improve our understanding of HIV among KPs and accelerate the production of high-quality evidence. Empowered with this evidence, scientists, program leaders, community stakeholders, and policymakers will be able to tailor and optimize HIV service delivery to the most

vulnerable populations, thus making the South African NSP goal possible. The amassed evidence will provide opportunities for comprehensive and innovative analyses which seek to address priority research questions to improve our understanding of HIV among KPs, assist in setting program targets, guide programmatic decision-making, and ultimately improve the response to HIV in South Africa and other SSA countries.

Acknowledgements

The work reported herein was made possible through funding by the South African Medical Research Council (SAMRC) Project Code #57035 (SAMRC File ref no: HDID8528/KR/202) through its Division of Research Capacity Development under the Mid-Career Scientist Programme through funding received from the South African National Treasury. The content hereof is the sole responsibility of the authors and do not necessarily represent the official views of the SAMRC". This research is also funded, in part, by the U.S. National Institute of Allergy and Infectious Diseases under award R01AI170249. The content hereof is the sole responsibility of the authors and do not necessarily represent the official views of the funders

On the author contributions, RNPM, SDB conceptualised the protocol, initial draft of the manuscript and provided overall leadership for executing the review to completion. RNPM, EP, AR, KW, KR, AV, AB, CS, BS, MS, MJ, LO, KSJ, HIS, XL, OB, HN, PD, FW, SY, LM, MR, SDB contributed to the introduction, methodology and progress updates of the study. RNPM, EP, AV, AB, CS, BS, MS, MJ, LO, KSJ, HIS were responsible for engaging with key stakeholders and initiating data sharing process. RPNM and EP were responsible for engaging with legal team and finalising data sharing agreements. LM and MR were responsible for the facilitation and linkage to key stakeholders. HN and PD provided insights on the Boloka Data Storage Process Flow. H.I.S and KSJ drafted the Boloka Indicator list; AV and AB drafted the Boloka Harmonisation Tool – and the rest of the team reviewed the tools for further inputs. RNPM was responsible for management and coordination of review activities. All authors discussed the results and contributed to the final manuscript.

Conflicts of Interest

None declared.

Abbreviations

AFSA: AIDS Foundation South Africa

AGYW: Adolescent Girls, and Young Women

ART: Antiretroviral therapy

BZ: Beyond Zero

DHIS: District Health Information System

DUA: Data User Agreement

FCSM: Federal Committee on Statistical Methodology

FSW: Female Sex Workers

HIV: Human Immunodeficiency Virus

Human Sciences Research Council

KP: Key Populations

MSM: Men who Have sex with Men

NACOSA: Networking HIV and AIDS Community of Southern Africa

NDoH: National Department of Health

PHRU: Perinatal HIV Research Unit

PLHIV: People Living with HIV

PWUD: People Who Use Drugs

SANAC: South African National AIDS Council

SSA: Sub-Saharan Africa

SWEAT: Sex Worker Education Advocacy Taskforce

UJ DPA: University of Johannesburg Data Processing Agreement

References

1. The Joint United Nations Programme on HIV/AIDS (UNAIDS). Fact sheet 2022. UNAIDS; 2022. Available from: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf (accessed 12 September 2023)
2. The Joint United Nations Programme on HIV/AIDS (UNAIDS). Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic. UNAIDS; 2021. Available from: https://www.unaids.org/en/resources/documents/2021/UNAIDS_FactSheet (accessed 12 September 2023)
3. South African National AIDS Council (SANAC). The National Strategic Plan for HIV, TN and STIs

2023-2028. 2023. Available from: https://sanac.org.za/national-strategic-plan-2023-2028/ (accessed 12 June 2023).

4. Thembisa Model 4.4. Report on Provincial HIV Estimates. 2021. https://www.thembisa.org/content/downloadPage/Provinces2021 (accessed 30 September 2023)
5. Zuma K, Simbayi L, Zungu N, Moyo S, Marinda E, Jooste S, North A, Nadol P, Aynalem G, Igumbor E, Dietrich C. The HIV epidemic in South Africa: key findings from 2017 national population-based survey. International Journal of Environmental Research and Public Health. 2022;19(13):8125.
6. World Health Organisation (WHO). Consolidated guidelines on HIV prevention, diagnosis, treatment and care for key populations. 2016. Available at: http://apps.who.int/iris/bitstream/handle/10665/246200/9789241511124-eng.pdf (accessed 10 August 2023)
7. Mishra S, Silhol R, Knight J, Phaswana-Mafuya R, Diouf D, Wang L, Schwartz S, Boily MC, Baral S. Estimating the epidemic consequences of HIV prevention gaps among key populations. Journal of the International AIDS Society. 2021; 24: e25739. https://doi.org/10.1002/jia2.25739.
8. Lyons CE, Schwartz SR, Murray SM, Shannon K, Diouf D, Mothopeng T, Kouanda S, Simplice A, Kouame A, Mnisi Z, Tamoufe U. The role of sex work laws and stigmas in increasing HIV risks among sex workers. Nature communications. 2020;11(1):773.
9. Sullivan MC, Rosen AO, Allen A, Benbella D, Camacho G, Cortopassi AC, Driver R, Ssenyonjo J, Eaton LA, Kalichman SC. Falling short of the first 90: HIV stigma and HIV testing research in the 90–90–90 era. AIDS and Behavior. 2020; 24:357-62.
10. Stangl AL, Pliakas T, Izazola-Licea JA, Ayala G, Beattie TS, Ferguson L, Orza L, Mathur S, Pulerwitz J, Iovita A, Bendaud V. Removing the societal and legal impediments to the HIV response: An evidence-based framework for 2025 and beyond. PloS one. 2022;17(2): e0264249. https://doi.org/10.1371/journal.pone.0264249.
11. Gesesew HA, Tesfay Gebremedhin A, Demissie TD, Kerie MW, Sudhakar M, Mwanri L. Significant association between perceived HIV related stigma and late presentation for HIV/AIDS care in low and middle-income countries: a systematic review and meta-analysis. PLoS One. 2017;12(3):e0173928. 10.1371/journal.pone.0173928.
12. Baral S, Beyrer C, Muessig K, Poteat T, Wirtz AL, Decker MR, Sherman SG, Kerrigan D. Burden of HIV among female sex workers in low-income and middle-income countries: a systematic review and meta-analysis. The Lancet infectious diseases. 2012; 12(7):538-49. 10.1016/S1473-3099(12)70066-X.

13. Savva H. Differentiated Service Delivery for Key Populations, Monitoring and Evaluating Key Populations Programs: South Africa.[PowerPoint Slides] CDC South Africa. 2021. Available from: https://cquin.icap.columbia.edu/wp-content/uploads/2021/08/Savva_Session-6a_FINAL.pdf (accessed 10 March 2023)
14. Lyons C, Bendaud V, Bourey C, Erkkola T, Ravichandran I, Syarif O, Stangl A, Chang J, Ferguson L, Nyblade L, Amon J. Global assessment of existing HIV and key population stigma indicators: A data mapping exercise to inform country-level stigma measurement. PLoS medicine. 2022;19(2):e1003914. https://doi.org/10.1371/journal.pmed.1003914.
15. Long LC, Rosen S, Nichols B, Larson BA, Ndlovu N, Meyer-Rath G. Getting resources to those who need them: The evidence we need to budget for underserved populations in sub-Saharan Africa. Journal of the International AIDS Society. 2021;24: e25707.
16. Stone J, Mukandavire C, Boily MC, Fraser H, Mishra S, Schwartz S, Rao A, Looker KJ, Quaife M, Terris-Prestholt F, Marr A. Estimating the contribution of key populations towards HIV transmission in South Africa. African Journal of Reproduction and Gynaecological Endoscopy. 2021; 24(1): e25650. https://doi.org/10.1002/jia2.25650.
17. Brown T, Peerapatanapokin W. Evolving HIV epidemics: the urgent need to refocus on populations with risk. Curr Opin HIV AIDS. 2019 Sep;14(5):337-353. doi: 10.1097/COH.0000000000000571.
18. Liu C, Lu X. Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community. BMC medical informatics and decision making. 2018; 18:1-0.
19. Garnett GP. Reductions in HIV incidence are likely to increase the importance of key population programmes for HIV control in sub-Saharan Africa. Journal of the International AIDS Society. 2021; 24: e25727. doi:10.1002/jia2.25727.
20. Green D, Tordoff DM, Kharono B, Akullian A, Bershteyn A, Morrison M, Garnett G, Duerr A, Drain PK. Evidence of sociodemographic heterogeneity across the HIV treatment cascade and progress towards 90-90-90 in sub-Saharan Africa–a systematic review and meta-analysis. Journal of the International AIDS Society. 2020 Mar;23(3):e25470.
21. Schwartz SR, Rao A, Rucinski KB, Lyons C, Viswasam N, Comins CA, Olawore O, Baral S. HIV-related implementation research for key populations: designing for individuals, evaluating across populations, and integrating context. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2019 Dec 1;82:S206-16. doi:10.1097/QAI.0000000000002191.
22. South African National AIDS Council: Pisa PT, Chikandiwa A, Luwaca B, Motsieloa L. Rampilo M, Khumalo Z, Odama A. Republic of South Africa: 2021 Global AIDS Monitoring Report.

Johannesburg: AROWANA; 2022. Unpublished.

23. University of Johannesburg. Annual Report. 2022. Available at: https://www.uj.ac.za/wp-content/uploads/2023/06/annual-report-2022.pdf (accessed 10 September 2023).
24. Refilwe Nancy Phaswana-Mafuya, Edith Phalane, Katharine S. Journeay, Haley I. Sisel, Claris Siyamayambo, Betty Sebati, Francois Wolmarans, Katherine Rucinski, Amrita Rao, Kalai Willis, Xiaoming Li, Bankole Olatosi, Stefan D. Baral. Harnessing big heterogeneous data to evaluate the potential impact of HIV responses among key populations in generalized epidemic settings in Sub Saharan Africa: the Boloka Data Repository. Proceedings of the 4th National Big Data Health Science Conference. BMC Proc 17 (Suppl 19), 32 (2023). https://doi.org/10.1186/s12919-023-00281-y.
25. Mehari KR, Jeffrey A, Chastang CM, Schnitker SA. Transdisciplinary participatory action research: How philosophers, psychologists, and practitioners can work (well) together to promote adolescent character development within context. The Journal of Positive Psychology. 2023: 1-2.
26. Menken S, Keestra M. An introduction to interdisciplinary research: Theory and practice. An Introduction to Interdisciplinary Research. 2016:1-28.
27. Cornish F, Breton N, Moreno-Tabarez U, Delgado J, Rua M, de-Graft Aikins A, Hodgetts D. Participatory action research. Nature Reviews Methods Primers. 2023 Apr 27;3(1):34. https://doi.org/10.1038/s43586-023-00214-1.
28. Phaswana-Mafuya RN, Phalane E, Rao A, Willis K, Rucinski K, Voet KA, Abdulrahman A, Siyamayambo C, Sebati B, Seloka M, Jaiteh M Olifant L, Journeay K, Sisel H, Wolmarans F, Li X, Olatosi B, Motsieloa L, Rampilo M, Baral SD. Leveraging big heterogenous HIV-related data in the era of Protection of Personal Data Act in Sub Saharan Africa: Lessons learned from the Boloka project. BMC Proceedings. 2024; 18(8): P20.
29. Republic of South Africa: The Presidency. Personal Information Protection Act. 2013. Available from: https://www.gov.za/sites/default/files/gcis_document/201409/3706726-11act4of2013protectionofpersonalinforcorrect.pdf (accessed on 09 November 2023)
30. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data science journal. 2015 May 22;14:2-.
31. Federal Committee on Statistical Methodology (FCSM.) A Framework for Data Quality. 2020. Available from: https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf (accessed 09 February 2023)
32. Bizer C, Cyganiak R. Quality-driven information filtering using the WIQA policy framework.

Journal of Web Semantics. 2009;7(1):1-0.

33. Rao A, Schwartz S, Viswasam N, Rucinski K, Van Wickle K, Sabin K, Wheeler T, Zhao J, Baral S. Evaluating the quality of HIV epidemiologic evidence for populations in the absence of a reliable sampling frame: a modified quality assessment tool. Annals of epidemiology. 2022; 65:78-83.
34. Grzeskowiak LE, Gilbert AL, Morrison JL. Methodological challenges in using routinely collected health data to investigate long-term effects of medication use during pregnancy. Therapeutic advances in drug safety. 2013; 4(1):27-37.
35. World Health Organization (WHO). Data quality review: Module 1 Framework and metrics. Available from: https://iris.who.int/bitstream/handle/10665/259224/9789241512725-eng.pdf?%20sequence=1 (accessed 10 May 2023).
36. Phaswana-Mafuya RN, Phalane E. Harmonization of multiple HIV-related data sources in Sub-Saharan Africa: Lessons Learned from the Boloka Project. BMC Proceedings. 2024; 18(8): O14.
37. Beyond Zero. About us. Available from: https://beyondzero.org.za/about/ (accessed 07 November 2023)
38. The Networking HIV and AIDS Community of Southern Africa (NACOSA). Annual Report 2021. 2021. Available from: https://www.nacosa.org.za/2021/12/15/annual-report-2021/ (accessed 07 November 2023)
39. The Sex Workers Education and Advocacy Taskforce (SWEAT). Annual Report 2022. 2022. Available from: https://aidsfonds.org/resource/annual-report-2022 (accessed 07 November 2023)
40. Sisonke. About Us. Available from: https://www.sisonke.org.za/about-us/. Sisonke (accessed 07 November 2023)
41. AIDS Foundation South Africa (AFSA). 2021/2022 Annual Report. 2022. Available from: https://www.aids.org.za/wp-content/uploads/2023/09/AFSA-Annual-Report-2021-2022.pdf (accessed 07 November 2023)
42. Perinatal HIV Research Unit (PHRU). About us. 2023. Available from: https://phru.witshealth.co.za (accessed 07 November 2023).
43. TB HIV Care. Annual Report 2022. 2022. Available from: https://tbhivcare.org/annual-reports/ (accessed 07 November 2023)

Appendix 1: An overview of current and potential partners for the Boloka data repository

Partner Category	Examples of Partners
Country coordinating mechanism	South African National AIDS Council
International and Regional UN Agencies	Centers for Disease Control and Prevention (CDC), UNICEF, United Nations Development Programme, UNAIDS, WHO, United Nations Population Fund
Government Departments	National Departments of Health, Social Development, Correctional Services, Basic education
Non-governmental organizations (NGOs), non-profit organizations (NPOs)	Beyond Zero (BZ), Networking HIV & AIDS Community of Southern Africa (NACOSA), Sex Workers Education and Advocacy Taskforce (SWEAT), Aurum, TB/HIV Care, Right to Care, Sisonke
Science Councils	Human Sciences Research Council (HSRC), South African Medical Research Council, Council for Scientific and Industrial Research
Research Institutes in Universities	Desmond Tutu HIV Foundation, Centre for the AIDS Programme of Research in South Africa
Local and international donors	Global Fund, President's Emergency Plan for AIDS Relief (PEPFAR), UNAIDS, European Union, Department for international Development, Canadian International Development Agency
National Universities	All South African universities

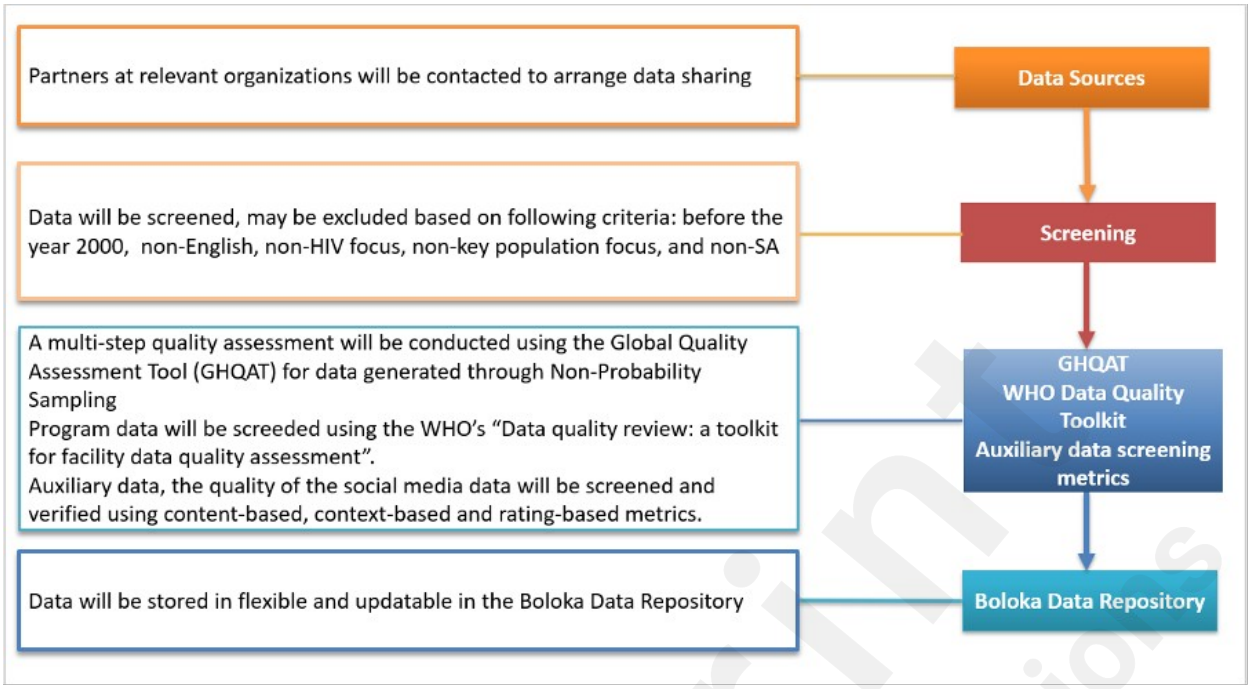
Regional economic communities	Southern Africa Development Community

Appendix 2: The Boloka data repository will leverage and collate available and diverse data from various sources across places, times, and populations

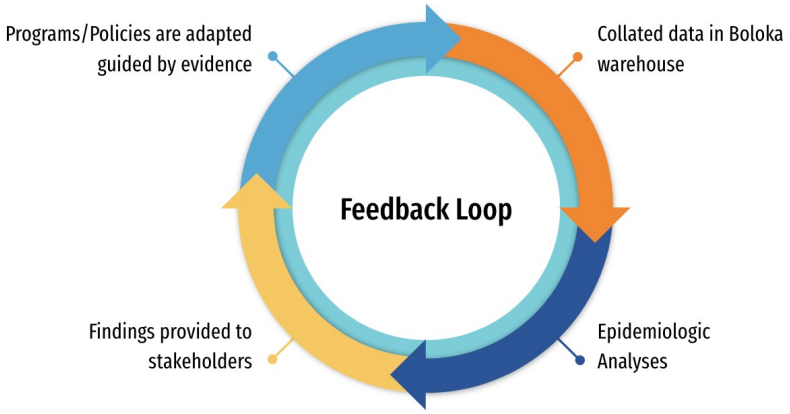
Data Type	Data Source	Purpose of Data Collection	Example(s)
Survey Data	Science Councils, e.g., Human Sciences Research Council, South African Medical Research Council, Government Departments, e.g., National Department of Health, Social Development, Basic Education, etc.	Data to monitor HIV indicators	Country-Specific Surveys, e.g., data from bio behavioral surveillance surveys among FSW and MSM, HIV prevalence survey amongst transgender women (TGW) in South Africa, South African National HIV, Prevalence, Incidence, Behaviour and Communication surveys (SABSSM surveys), South African Men's Health Monitoring Study (SAMHMS), HSRC Botshelo Ba Trans, South African National Health and Nutrition Examination Survey (SANHANES), HIV Surveys, Sexual and reproductive health surveys, periodic South African Stigma Index surveys, the Governance, Public Safety and Justice survey (GPSJS), Demographic Health surveys, online HIV surveys, household surveys, Demographic Health Surveys
Research Data	Academic institutions, research institutions, collaborators	Research conducted to answer priority HIV questions for the country - observational, implementation science, and experimental studies	Data from collaborative studies with Johns Hopkins University, Emory University, Centres for Disease Control and Prevention, Health Science Research Council, Desmond Tutu HIV Foundation, and University of California San Francisco
Program Data from implementing partners	Implementing Partners and main funders (NGOs, community- based organizations (CBOs), Global Fund, PEPFAR)	Routinely collected data at facility, district level and countrywide captured by partners for government	AIDS Foundation of South Africa (AFSA), Networking HIV and AIDS Community of Southern Africa (NACOSA, TB/ HIV Care; Beyond Zero, Aurum, South African Network of People who Use

		reporting population size estimates (PSEs) for FSW and MSM	Drugs (SANPUD)
Program data from national government departments	Government departments, e.g., Department of Health, Social Development and Basic Education	Routinely collected patient-level data and reporting purposes	Electronic health information management system, District Health Management system (DHIS), Tier.net, High Transmission Area data
Data from modeling studies	University of Cape Town	Estimates on key HIV-related indicators for guiding programming and policies	Data from Thembisa and Naomi models
Program data from community-led programs	Community-led monitoring system, developed by organizations representing people living with HIV	Routinely collected data collected for community level monitoring	Ritshidze, High Transmission Area Program
Reports	SANAC Report, Country and departmental annual reports, and organizational reports	Collate information from different sources to gauge the response	Data from the Annual Global AIDS Monitoring Report, NSP Mid-term and Annual Reports, Quarterly Factsheets on NSP, UNAIDS reports, NAC reports, WHO Reports, UNICEF reports

Appendix 3: Global HIV Quality Assessment Tool



Appendix 4: Proposed feedback loop



Appendix 5: Ethics approval letter

FACULTY OF HEALTH SCIENCES
RESEARCH ETHICS COMMITTEE

NHREC Registration: REC 241112-035

ETHICAL CLEARANCE LETTER
(RECX 2.0)

Student/Researcher Name	Refilwe Phaswana-Mafuya	Student Number	720062595
Supervisor Name	Phaswana-Mafuya, Metse		
Department	Environmental Health		
Research Title	HARNESSING BIG HETEROGENEOUS DATA TO EVALUATE THE POTENTIAL IMPACT OF HIV RESPONSES AMONG KEY POPULATIONS IN GENERALIZED EPIDEMIC SETTINGS IN SUB SAHARAN AFRICA		
Date	06 May 2022	Clearance Number	**REC-1504-2022**

Approval of the research proposal with details given above is granted, subject to any conditions under 1 below, and is valid until 2023/05/05.

1. **Conditions:**
Gatekeeper permission, as required.
**Please note that failure to comply with the conditions above (if any) prior to implementation of the research will invalidate this ethical clearance.*

2. **Renewal:**
It is required that this ethical clearance is renewed annually, within two weeks of the date indicated above. Renewal must be done using the Ethical Clearance Renewal Form (REC 10.0), to be completed and submitted to the Faculty Administration office. *See Section 12 of the REC Standard Operating Procedures.*

3. **Amendments:**
Any envisaged amendments to the research proposal that has been granted ethical clearance must be submitted to the REC using the Research Proposal Amendment Application Form (REC 8.0) prior to the research being amended. Amendments to research may only be carried out once a new ethical clearance letter is issued. *See Section 13 of the REC Standard Operating Procedures.*

4. **Adverse Events, Deviations or Non-compliance:**
Adverse events, research proposal deviations or non-compliance must be reported within the stipulated time-frames using the Adverse Event Reporting Form (REC 9.0). *See Section 14 of the REC Standard Operating Procedures.*

The REC wishes you all the best for your studies.

Yours sincerely.

Prof. Christopher Stein
Chairperson: REC
Tel: 011 559 6564
Email: cstein@uj.ac.za

RECX 2.0 – Faculty of Health Sciences **Secretariat:** Ms Raihaanah Pieterse

Appendix 6: Summary of planned and executed milestones for this project

Activity	2022	2023	2024	2025
Study approvals				
Ethics approval, gate keeper approvals (SANAC), data sharing agreements				
Study trainings				
Study training with all the project personnel i.e., ethics, study procedures, POPI Act training				
Acquire and collate heterogeneous data				
Assess and develop meaningful data partnerships and collaborations				
Obtaining & extracting data from multiple data sources				
Assess data for accuracy, relevance, and quality				
Pre-screening for relevance and inclusion, Capturing, pre-processing data, sorting and storing data				
Synthesize/collate/merging data from various sources, critical review according to the checklist, quality control				
Clean and store data in the data repository				
Clean and filter the data				
Engage with Amazon or similar				
Set up the Boloka Data repository				
Translate data into actionable knowledge				
Data integration and data flow control, maintenance, updates,				
License fees, adoption, implementation, analysis, and visualization of data				
The Boloka Project Protocol Paper – Submission for publication				