BMJ Open

Diverse experts' perspectives on ethical issues in predicting HIV/AIDS risk in Sub-Saharan Africa

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-052287
Article Type:	Original research
Date Submitted by the Author:	11-Apr-2021
Complete List of Authors:	Nichol, Ariadne; Stanford University School of Medicine, Center for Biomedical Ethics Bendavid, Eran; Stanford University, Mutenherwa, Farirai ; University of KwaZulu-Natal; University of KwaZulu-Natal, School of Applied Human Sciences Patel, Chirag; Harvard Medical School, Department of Biomedical Informatics Cho, Mildred; Stanford University School of Medicine, Center for Biomedical Ethics
Keywords:	Public health < INFECTIOUS DISEASES, ETHICS (see Medical Ethics), Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, MEDICAL ETHICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Diverse experts' perspectives on ethical issues in predicting HIV/AIDS risk in Sub-Saharan Africa

Ariadne A. Nichol,[1*] Eran Bendavid,[1] Farirai Mutenherwa,[2,3] Chirag Patel,[4] Mildred K. Cho[1]

[1] Stanford University School of Medicine, Stanford, California, United States

[2] School of Applied Human Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa

[3] KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

[4] Harvard Medical School, Boston, Massachusetts, United States

* Corresponding Author (Ariadne A. Nichol)

Address: 1215 Welch Road, Modular A, Stanford, CA 94305, United States

Email: ariadnen@stanford.edu

Word Count: 3,501

ABSTRACT (271 words)

Objective: To better understand diverse experts' views about the ethical implications of ongoing research funded by the National Institutes of Health that uses machine learning to predict HIV/AIDS risk in Sub-Saharan Africa based on publicly-available Demographic and Health Surveys data.

Design: Three rounds of semi-structured surveys in an online expert panel.

Participants: Experts in informatics, African public health and HIV/AIDS and bioethics were invited to participate.

Measures: Perceived importance of or agreement about relevance of ethical issues on 5-point uni-polar Likert scales. Qualitative data analysis identified emergent themes related to ethical issues and development of an ethical framework and recommendations for open-ended questions.

Results: Of the thirty-five invited experts, 22 participated in the online expert panel (63%). Emergent themes were the inclusion of African researchers in all aspects of study design, analysis, and dissemination to identify and address local contextual issues, as well as engagement of communities. Experts focused on engagement with health and science professionals to address risks, benefits, and communication of findings. Respondents prioritized the mitigation of stigma to research participants but recognized trade-offs between privacy and the need to disseminate findings to realize public health benefits. Strategies for responsible communication of results were suggested, including careful word choice in presentation of results and limited dissemination to need-to-know stakeholders such as public health planners.

Conclusion: Experts identified ethical issues specific to the African context and to research on sensitive, publicly-available data, and strategies for addressing these issues. These findings can be used to inform an ethical implementation framework with research stage-specific recommendations on how to utilize publicly-available data for machine-learning-based predictive analytics to predict HIV/AIDS risk in Sub-Saharan Africa.

Strengths and limitations of this study

- A strength of this study is that it represents the perspectives of diverse experts on the unique ethical issues raised by the use of predictive analytics for HIV/AIDS risk on large public health datasets in Sub-Saharan Africa.
- Another strength of the study is our use of open-ended questions and qualitative analysis of anonymously collected data to enhance breadth and validity of responses, and three rounds of iterative surveys to identify and resolve areas of disagreement.
- A third strength of the study is that it elicited specific suggestions from experts to navigate ethical tradeoffs, such as alternative methods of describing and disseminating findings of predictive analytics to minimize risks to privacy and of stigmatization, and suggestions for prioritizing specific groups for community engagement.

- The main limitation of this study is that a small number of respondents completed all three surveys, however, our expert respondents did represent diverse perspectives in informatics, bioethics of Africa-based studies, and African public health and HIV/AIDS.

INTRODUCTION

It is now well recognized that the use of big data for health research poses significant ethical challenges.[1–3] In particular, such research poses risks to the privacy of sensitive information as well as the potential for re-identification, stigmatization, and bias.[4–6] Many research cohort datasets with individual or patient-level information are available, such as those from epidemiological studies from biobanks (e.g., UK Biobank), repositories (such as dbGaP), and surveillance programs (e.g., Demographic and Health Surveys and US Centers for Disease Control and Prevention).

Several research studies aim to predict HIV/AIDS status in Sub-Saharan African (SSA) countries using data from the Demographic and Health Surveys (DHS).[7,8] While there were no specific regulatory barriers to this research, it raised concerns for the researchers about whether existing ethical frameworks were adequate to address its specific constellation of characteristics. Namely, these included the particularly sensitive nature of HIV/AIDS, especially in SSA countries, the region's history of human rights abuses and exploitation, and the goal of predicting HIV/AIDS status using easily ascertainable features.

We therefore conducted a series of surveys of an expert panel with diverse expertise, including bioethics of Africa-based studies, informatics, and African public health and HIV/AIDS to better understand the ethical implications and concerns about this type of research and to inform an ethical framework and recommendations for researchers.

METHODS

Approach

Our overall approach was modeled after the Delphi method, but was heavily modified because our goal was not to achieve consensus but to document the range of perspectives of experts from diverse backgrounds about ethical issues and converge on recommendations for addressing them. Therefore, we relied largely on qualitative analysis, based on responses to open-ended questions to identify themes not already identified in the literature. We also asked closed-ended questions to better understand how individuals prioritized specific ethical issues and recommendations. We surveyed an expert panel in multiple rounds, building on responses to each round to develop the questions for the next one. We focused on identifying questions that required clarification or that indicated areas of disagreement that could be probed with more specificity in the subsequent survey.

Sample

We identified 35 experts in informatics (n=10), African public health and HIV/AIDS (n=9) and bioethics of Africa-based studies (n=16) through searches of the biomedical and ethics literature and by snowball sampling. All but one of the public health and bioethics experts were from African countries (Ethiopia, Ghana, Kenya, Nigeria, Rwanda, South Africa, Uganda, Zambia, Zimbabwe), and all of the informatics experts had their primary academic appointments in the United States, but did work on health in Africa. Experts were invited by email and were offered US$200 for

participation in all three surveys. Twenty-two agreed by email to participate (22/35=63%). Five actively declined, and eight did not respond to the initial invitation or to follow-up emails.

Surveys

We administered a series of three online, scenario-based semi-structured surveys, anonymously via Qualtrics, to make participation convenient and encourage frank responses. Respondents were allowed approximately three weeks to respond, with two reminder emails to all 22 who initially agreed to participate.

Survey 1 began with a scenario describing an actual research study funded by the National Institute of Allergy and Infectious Diseases at the US National Institutes of Health (Box 1). The study utilizes large, publicly-available survey cohort data that includes detailed health data and HIV status of millions of survey participants throughout the world, socioeconomic data, and Global Positioning System (GPS) coordinates of randomly displaced neighborhoods by up to 5km to protect privacy. Data collection is conducted with informed consent by interviewers in person, and reviewed and approved by an ethics review board in individual countries. Data users must register for data access and undergo review by their local ethics review board (Box 1).

Box 1: Survey 1 scenario

A group of American scientists funded by the US government is developing big data tools to identify individuals and groups at elevated risk of acquiring HIV in Sub-Saharan Africa. The purpose of the project is to help ministries of health and international public health organizations target testing and treatment programs to the individuals and groups most at-risk. The scientists are using large, publicly-available datasets that identify the HIV status of millions of individuals, and hundreds of additional personal and household features of these individuals, some of which is collected by surveys. Household wealth, educational history, marital status, and the GPS coordinates of the households' village

or neighborhood, among others, are characterized in detail. The data are readily available on the web for anyone who registers, and the source code for using the data and executing the HIV risk identification procedures are posted for public access. Policy makers in African countries have expressed interest in the findings, but have not specified how they plan to use the new information.

The survey began with three open-ended questions about 1) ethical issues they believed should be addressed by researchers conducting the study; 2) any details about the study that were not provided in the scenario but would be important to understanding the associated ethical issues; and 3) any specific recommendations for researchers conducting this or similar studies. We then asked respondents to rate the importance of seven ethical issues that we identified in the literature as potentially relevant to this scenario, using a 5-point uni-polar Likert scale ranging from 1=*Not important at all* to 5=*Absolutely essential*. Ethical issues included privacy, validity, power disparities, alignment and conflicts of interests, benefit-sharing, stigma, and bias. We specifically presented these seven issues after the open-ended questions in order to avoid anchoring or constraining open-ended responses, in hopes of eliciting a wide range of ethical issues and recommendations.

Surveys 2 and 3 were designed based on the responses to the previous surveys, as described below in Results.

Patient and public involvement

Patients and the public were not involved in research question development, study design, or analysis since the research specifically sought to elucidate experts' opinions on research utilizing big data for predicting HIV/AIDS. The expert panelists did

propose appropriate approaches for community and public engagement and for disseminating sensitive research findings.

RESULTS

Survey 1

Of the 22 experts who agreed to participate in the panel, 16/22 (73%) responded to Survey 1 (overall response rate 16/35 = 46%). Because survey responses were anonymous, we do not know what proportion of respondents were experts in informatics, public health, or bioethics. In responses to closed-ended questions (see Table 1) respondents rated almost all issues as "of average importance", "very important", or "absolutely essential" (6 of 7 issues had a mean rating of at least 4.0 on a scale of 1-5), and did not rate any of the 7 issues as *Not Important At All* or *Of Little Importance*. Nevertheless, two items clearly emerged as being most important. First was the potential to stigmatize groups or populations that are uniquely identified by the research (all rated this issue as *Absolutely Essential*) and, second, the privacy of individuals (14 rated this as *Absolutely Essential*, 2 as *Very Important*). The next two most important issues identified were the validity of findings using big data tools, and potential for bias.

Open-ended responses were exceptionally rich, and reflected issues of re-identification, stigma, discrimination against individuals, families, or geographically defined and/or socially defined groups, consistent with the importance accorded these issues in the responses to the closed-ended questions which were asked later in the

survey. Respondents brought up several general ethical concerns commonly raised in relation to biobanking in SSA, such as data ownership and access, data security and privacy, research priority setting, and benefit-sharing.[9–11]

Several responses raised concerns around individual autonomy and consent obtained to use personal data, whether at the initial collection of DHS data or at the start of research utilizing machine learning predictive analytics to analyze the data. These responses indicated that respondents needed more detail on how informed consent and ethical review processes were conducted for data collection for the DHS and for data use by individual researchers. As a result, we significantly expanded the description of the study for Survey 2 (Box 2).

Box 2: Survey 2 scenario

Research team: US-based scientists with expertise in infectious diseases and bioinformatics.
Funding source: US Department of Health and Human Services.

Rationale: HIV is the largest single cause of death among adults in Sub-Saharan Africa, responsible for about a fifth of all adult deaths in 2017. However, despite the dramatic increase in the availability of antiretroviral therapy, over 1.2 million people were newly infected in Sub-Saharan Africa in 2017, an incidence rate more than 10-fold higher than in the United States. A better understanding of the social, behavioral, environmental, and economic contexts that influence HIV risk could improve the effectiveness and efficiency of prevention and treatment programs.

Aims: The overall goal is to analyze large-scale datasets of HIV in Sub-Saharan Africa to identify new risk factors with potential to improve HIV care, and to help ministries of health and international public risk factors with potential to improve HIV care, and to help ministries of health and international public health organizations target testing and treatment programs.

Methods: The primary approach entails aligning HIV test results (positive or negative for HIV-1) with all social, economic, behavioral, and environmental features collected on individuals in the Demographic and Health Surveys (DHS). The DHS has completed home-based HIV testing on over 1,000,000 individuals in sub-Saharan Africa, and the entirety of the DHS information – over 1,000 potential predictors for the average person – is available for each individual, de-identified as described below (see section on Data

privacy, access and ethical review, below). For all biomarker testing, verbal pre- and post-test counseling and printed information are provided to respondents, and test results are kept confidential. HIV-positive respondents are referred to a local health care facility for appropriate care. Analytic approaches include testing for association of HIV status with each of the predictors, as well as building sophisticated prediction models of HIV status using statistical learning approaches such as LASSO and Elastic Net.

Data sources: USAID Demographic and Health Surveys (DHS) from all Sub-Saharan African countries. All survey data are publicly available and are collected through a Household Questionnaire, and Individual Man's or Woman's Questionnaire, and a Biomarker Questionnaire. Household wealth, educational history, marital status, and the GPS coordinates of the households' village or neighborhood, among others, are characterized in detail. Biomarker testing for HIV status has been conducted in all endemic sub-Saharan countries since 2003.

Data privacy, access, and ethical review: Respondent interview and data files are initially identified by enumeration area (EA) and household numbers and then coversheets with these identifiers are destroyed and EA/household numbers are randomly reassigned. Geographic coordinates of each survey are displaced in a random direction and distance up to 2 km (urban) or 5 km (rural) and randomly selected rural clusters displaced up to 10 km.

DHS questionnaires and general data collection procedures are reviewed and approved by an external Institutional Review Board (IRB) and country-specific protocols are reviewed and approved by an IRB from the individual country, which ensures that the survey complies with national laws and norms. Informed consent is conducted by interviewers in person, in a private location to provide privacy about sensitive topics, and includes a discussion of the purpose of the interview or test, privacy about sensitive topics, and includes a discussion of the purpose of the interview or test, expected duration, procedures, potential risks and benefits to the respondent, and contact information for a person who can provide more information. Consent for those undergoing HIV testing for DHS also explains that test results cannot be provided to individuals because names are not attached, but that a free voucher for health services that can provide HIV testing, and a list of local testing facilities is provided for study participants and their partners.

In order to access the DHS data, the US researchers registered for data access on the DHS website. Registration requires a project description and consent for maintaining the data secure and publishing only aggregated findings (i.e., not individual-level data). Once access was granted, the US researchers downloaded the data to secure servers with password protected access. The US researchers' protocol has been reviewed and approved by their university's IRB but is not considered human subjects research because it is considered research on an existing publicly- available, de-identified and non-coded dataset.

206

The specific use of big data predictive analytics generated several ethical issues that respondents wanted to ensure were properly addressed prior to any research, including assessment of the potential for bias, independent review of the validity of the predictive analytics tools, and establishment of a plan for monitoring interventions for harm that could result based on which individuals or groups were identified as being high risk for HIV/AIDS. Several respondents also emphasized the need for researchers to think through how the big data predictive analytics outcomes can be used to inform testing and treatment programs beyond simply identifying high-risk individuals or groups.

Respondents articulated a number of ethical issues that were not mentioned in the closed-ended questions, especially concerns about using DHS data sources to predict HIV/AIDS risks specific to the African context. Contextual factors cited included a history of human rights abuses, lack of trust in government, misuse of research findings, HIV-associated characteristics (e.g., homosexuality) that are crimes in some African countries, lack of expertise in big data analysis, lack of agency of African researchers and ethicists, compliance with or lack of country-specific laws and policies, and the need for engaging African scientists in order to provide contextual knowledge to inform best research and ethics practices. Another theme that emerged was concern about data on Africans being used by non-African researchers (see Table 2).

Survey 2

We invited all 22 experts who agreed to participate in the panel to take Survey 2, regardless of whether they had taken Survey 1. Ten experts responded (10/22 = 45%).

Survey 2 focused on issues that respondents indicated were most important in Survey 1. All 5 of the survey questions were open-ended, and were presented to participants as themes reflecting areas of consensus that had emerged in the previous survey. Survey 2 questions focused on 1) stakeholder engagement; 2) privacy/stigmatization/discrimination; 3) ethics review; 4) data access; and 5) dissemination and communication of study findings (Box 3).

Box 3: Survey 2 questions

Q1. Stakeholder and community engagement. A theme that emerged from responses to Survey 1 was the need for the researchers to engage stakeholders in the planning, design, analysis and dissemination of the research in order to identify and address contextual factors, including local laws and attitudes. The stakeholders included African scientists, ethicists, public health policy makers, and communities.

Given that the DHS data come from a large number of countries and are intended to be nationally representative, how would you suggest that the task of stakeholder engagement be approached, and by whom?

Q2. Privacy, stigmatization and discrimination. Data privacy was clearly identified in Survey 1 as the most important ethical concern about the HIV Big Data research project, primarily because of the potential for stigmatization of and discrimination against people with HIV/AIDS. Even though data obtained by the researchers have been stripped of explicit identifiers, and data have been randomly displaced geographically, re-identification of individuals, families, and groups defined by geographical or phenotypic characteristics could still be a concern because of the large amount of data collected about each individual. The US researchers have assured their IRB that they will not attempt to re- identify individuals or groups from the subset of DHS data that they have obtained, but risk factors that emerge from their analysis could be used to identify and thus stigmatize or discriminate against those with those characteristics.

How would you suggest that the US researchers minimize the chances that their identification of risk factors is misused?

Q3. Ethics review. Data collection for the DHS surveys was conducted with informed consent and with centralized ethics review of the general protocol and local review of country- specific protocols. Because the data are publicly available, the US researchers' IRB does not consider the secondary analysis of the data to be human subjects research. Although the US researchers obtained IRB approval from their university for their study, it was considered "exempt", so further review and informed consent was not

required.

In Survey 1, some respondents expressed the need for ethics review. Do you believe that the centralized and local review of the DHS survey and by the US university sufficient? If not, what additional review should be instituted, by whom, and why?

Q4. Data access. The DHS dataset is publicly available but subject to some access control. Any requests for access to data must be approved by DHS staff. General approvals do not automatically guarantee access to the HIV data. Separate requests must be made to access both the general survey and HIV survey data.

Do you believe that this type of control of access to the DHS dataset is sufficient to prevent misuse? If not, what additional controls would you recommend?

Q5. Study findings. The analysis of the DHS data is anticipated to identify a set of risk factors for acquisition of HIV.

Do you have any recommendations for the data analysts for how best to communicate what these risk factors are, assuming that the study findings will be disseminated to governmental and non-governmental public health organizations, other scientists, and to the general public?

Overall, community and stakeholder engagement that includes Africans, ideally in relevant countries, was seen as key to minimizing risks at several stages of the research process, including data access, protocol oversight, and dissemination and implementation of findings. Some recommended engagement at the regional as well as national level, and respondents named a wide range of stakeholder groups (see Table 3). There was also broad support for community engagement in general to protect interests of local communities, groups and individuals. This engagement would provide the opportunity to better understand local concerns, values, norms, and cultural considerations and guide researchers on how to communicate findings in a way that mitigate risks to communities and individuals. Other purposes of stakeholder and community engagement were to provide education to public health officials and

policymakers, clinicians, and communities, enhance buy-in, identify opportunities for capacity building and translation, and ultimately build trust and collaboration.

While Survey 1 indicated consensus on privacy as a primary concern, in Survey 2, statements about how researchers could address this issue were mixed. Some acknowledged limits on researchers' ability to prevent misuse of findings or to completely protect data privacy; however, others also proposed specific actions to minimize harms. For example, one respondent said, "Of course there is nothing like absolute anonymization of data. I suggest that if sensitive results are obtained, it is imperative that the US research team works with communities in the affected countries on how best to disseminate the findings." Another suggested, "Decide not to report data sub-groups containing very small numbers of individuals."

There was a lack of consensus on the adequacy of centralized versus local ethics review and whether research on publicly-available or de-identified data was considered exempt from ethics review. There was also disagreement about the adequacy of existing data access control and protection against stigma and discrimination from study findings. While one respondent suggested that data access controls were sufficient because data were de-identified, another would require "a clear data analysis and dissemination plan", and another stated that protocol-specific data sharing agreements were necessary, because "Africa has suffered most from exploitation; both for research subjects and researchers."

The findings from Survey 2 were used to design Survey 3 to probe areas of disagreement, and to elicit details that could inform draft recommendations about stakeholder engagement and ethics review.

Survey 3

We invited all 22 experts who agreed to participate in the panel to take Survey 3, regardless of whether they had taken Surveys 1 or 2. Ten experts responded (10/22 = 45%). Because the surveys were anonymous, we do not know whether these 10 respondents were the same as those who responded to Survey 2.

In Survey 3, we presented the same scenario as in Survey 2, providing additional examples of analysis that could be conducted using the DHS data that highlighted the types of features that could be identified as risk factors using predictive analytics, and presented alternative ways of describing research findings that would pose tradeoffs between dissemination and privacy (Box 4).

Box 4: Survey 3 examples added to research scenario

Here are **examples of data analyses** that could be conducted with DHS data:

These analyses would use data collected in 30 African countries, and include: the results of HIV tests from about 1,000,000 men and women between the ages of 15 and 49 (women) or 15 and 59 (men) who had consented to an HIV test; household data (e.g. floor material, water source, electricity); family information (marital status, number of children); health information (hemoglobin measurement, height and weight); family planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among others.

The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small subsets of characteristics. The predictors may or may not have been identified by previous epidemiological research, but may be strongly predictive. For example, bicycle ownership is, in some surveys, a strong predictor of HIV status, and adding it to a risk prediction model can improve prediction accuracy of HIV status from 82% to 85%.

One type of analysis would identify the individual features that are most closely tied to HIV status. This would have the potential to improve targeting of public health programs or help design interventions. For example, if widowhood is identified as a strong predictor of being HIV-positive, this can help design testing and prevention programs that are tailored to widows. This is similar to the identification of male circumcision as a

risk factor that led to clinical trials and large-scale public health programs.

Another type of analysis would create risk scores that are a weighted combination of many individual features. This risk score would emerge from a commonly-used "black box" machine learning approach that chooses the combination of features that best predicts HIV status. The product of this analysis may not disclose any individual risk factors, and indeed some factors might only be predictive in combination with others. The analysis could report how well models predict the chance of being HIV-positive given a combination of features.

We then sought to clarify positions expressed by respondents in Survey 2 by focusing on policies and actions regarding: 1) who to include in stakeholder engagement; 2) strategies for dissemination of research findings to mitigate stigmatization and discrimination, and 3) requirements for ethical review, with a closed-ended component and opportunity for open-ended explanation. Question 1 about stakeholder engagement asked respondents to rate the importance of including each of a list of potential stakeholders on a Likert scale ranging from *Critically important to include* to *Do not include*. Questions 2 and 3 asked respondents to rate their level of agreement with statements on balancing privacy, stigmatization, and discrimination concerns with the dissemination of useful findings of risk factors for HIV/AIDS and their level of agreement with statements on the level of ethics review that is sufficient.

When asked which specific stakeholders would be critically important to include in stakeholder engagement, over half of participants believed African data scientists, African ethicists, representatives from a national Ministry of Health and representatives from African universities were necessary to include. Interestingly, responses were split (roughly in half) on whether African religious leaders and healers, African health workers and African patients and families were critically important to include or not important to include in stakeholder engagement. There were divergent opinions as to

the necessity of representatives from local communities as well. This is perhaps in part due to the nature of the data having been collected within multiple countries, where 'local' might be difficult to define given the context. One respondent articulated the concern that it might prove difficult to identify and engage with local communities with data coming from over one million people.

Representatives from the African regional WHO office (AFRO) and representatives from the African Academy of Science and public-health-related NGOs were viewed as stakeholders to include if resources were available, but not critical. For some participants, the stage of the study influenced which stakeholders they felt were relevant to engage. For example, community members only need to be engaged to minimize risks once the analysis is complete and agencies intend to take action based on results of the analysis.

When asked about the balance between the benefits of disseminating the research findings and risks of identification and stigma, there was support for some limitations on reporting (i.e. reporting overall performance of predictive models rather than individual risk factors) but less agreement about restricting reporting of findings that could identify small numbers of individuals if the findings would be less useful for public health officials. There was broad agreement on the need for community representatives to have input on how risk factors are described in publications, but less so for input by public health officials. There was strong disagreement with the proposed statement that researchers cannot do anything to protect against stigmatization based on risk factors. Several strategies on the communication of results were suggested, including reviewing and validating predictive models, careful word choice in the

packaging of results, and limited dissemination to need-to-know stakeholders such as public health planners.

In clarifying the divergence of responses in Survey 2 on the amount of ethics review required for data collection and data analysis, a majority of respondents agreed that the combination of centralized ethics review of data collection and institutional ethics review of data analysis by the researchers' institution would be sufficient. Most respondents disagreed with the suggestions of requiring additional ethics review by all national research ethics committees of countries involved in data collection or additional ethics review by regional or African organizations.

DISCUSSION

It is increasingly recognized that the use of predictive analytics and artificial intelligence techniques such as machine learning on health data raises new ethical concerns or exacerbate existing issues. Most research to date has unearthed issues arising in high-income contexts, but we demonstrate here that many of these issues are salient for lower-income contexts as well. The rapid convergence and availability of new analytic methods and big data bring out issues arising from the use of such techniques on publicly-available datasets, especially with sensitive data such as HIV/AIDS status. We explored these issues as they pertained to actual US-funded research that uses data from individuals in SSA, a context that could sharpen ethical and social concerns. We demonstrate that issues of data privacy, stigma and discrimination, which are well-documented concerns of big data, were identified as key issues. However, our expert

panel largely agreed that the current practice of ethical review at the point of data collection and individual projects using large datasets was sufficient even in the SSA context.

While experts in our panel pointed to other problematic features of big data and predictive analytics such as bias, the preponderance of responses to open-ended questions highlighted ethical concerns that would apply to much of biomedical research generally, but with a focus on contextual factors. These factors included a history of human rights abuses, lack of trust in government and in non-African researchers, misuse of research findings, and obligations of US researchers to help build research capacity in Africa.[12]

On the other hand, there was some acknowledgement of the potential benefits from research presented in the scenario, as well as recognition of the inability to maintain anonymity of research data. As a result, respondents were reluctant to support a complete block of the dissemination of findings. Respondents put forward a number of practical and feasible suggestions aimed at big data research, including privacy-preserving approaches for reporting the findings of predictive analytic models such as reporting overall performance of predictive models rather than individual risk factors. In addition, respondents suggested that benefits from research would be enhanced by validation of predictive analytic tools.

Our findings are consistent with previous studies that raised concerns over privacy, confidentiality, consent, and data misuse in the African context.[13–15] Our results demonstrate that consent and ensuring individual privacy and confidentiality were of primary concern to the expert panel, especially given use of predictive analytics.

Our findings mirror statements of others such as the H3Africa working group on ethics, who identified that community engagement is needed to support the informed consent process in the context of genomic research in Africa.[16] Others have stressed that community engagement in public health research in Africa is not only instrumental to recruitment and retainment of participants in research studies, but also intrinsically valuable as good ethical practice.[13] Divergent from these findings, we saw no explicit connection drawn between community engagement and informed consent. However, both topics were raised by our expert panel as separate issues that needed to be addressed.

Community engagement is seen as a critical part of health research in general, especially in SSA, given the history of exploitation.[15,17–19] Yet, there is extensive variation in defining what constitutes a "community".[15,16,20,21] Our findings indicate recognition of the need for engagement at national, regional, and local levels with a wide array of proposed participants. Interestingly, our panel of experts found other stakeholders (i.e. African ethicists, university researchers, data scientists, and representatives from ministries of health and universities) beyond local communities to be crucial to engage with in order to minimize risks of stigmatization and discrimination. It is important to consider the nature of predictive analytics and big data research as transnational and inclusive of many individuals' data. Therefore, this focus on broader stakeholders' engagement is explicable and perhaps a somewhat unique feature to research involving big data. Of course, these stakeholders have been identified as proposed participants of engagement for health research before, including within the

context of genomics and biobanking in Africa.[9,22,23] Some have also suggested these stakeholders are in fact "community" in a broader interpretation.[15]

Public health data shared appropriately depends on "the trust and confidence of those from whom such data are derived and relate to".[13] Community and stakeholder engagement activities are key to developing such trust, and should be considered by researchers conducting studies using data from populations where trust has historically been threatened. The expert panel's recommendations around which stakeholders are essential to include can help researchers using predictive analytics and artificial intelligence engage with relevant communities.

One limitation of this study was the small number of respondents that completed all three surveys. However, the participants were experts in their respective fields of informatics, public health, HIV/AIDS, and bioethics in Africa, which increased confidence in the insights reported. In addition, the informatics experts were largely US-based, although they had experience in working on HIV/AIDS and internationally.

CONCLUSION

Experts identified a number of ethical issues involved in carrying out research using big data predictive analytics to identify high-risk individuals or groups for HIV/AIDS in SSA. While many of these issues were not specific to big data or predictive analytics, our expert panel did focus on features specific to the SSA context, especially the inclusion of African researchers in all aspects of research. The expert panel offered strategies for navigating the trade-off between protection of privacy of sensitive and big data and dissemination of results, as well as priorities for which communities to involve

in stakeholder engagement. Overall, the findings from this study can potentially inform an ethical implementation framework with research stage-specific recommendations on how to utilize machine-learning-based predictive analytics to predict risk of HIV/AIDS and other potentially sensitive conditions (such as COVID-19) in SSA. The recommendations could also be applicable to studies conducted in the context of serving historically disadvantaged or exploited groups more broadly.

FOOTNOTES

Contributors

EB and MKC contributed to the conception and design of the study. AAN and MKC completed the data collection and analysis and drafted the initial manuscript. EB, FM, and CP drafted and finalized the manuscript with equal contributions. All authors contributed to drafts and approved the final manuscript.

Funding

This work was funded by the U.S. National Institutes of Health (1 R01 AI127250-01, Big Data Analysis of HIV Risk and Epidemiology in Sub-Saharan Africa). Eran Bendavid and Chirag Patel are Principal Investigators and Mildred Cho is a co-investigator of the grant. The views expressed are solely those of the authors.

Competing interests

None declared.

Patient consent for publication

Not required.

Data availability statement

Data are available upon reasonable request. Data are de-identified survey responses. Requests can be made to the corresponding author.

REFERENCES

1. Vayena E, Madoff L. Navigating the Ethics of Big Data in Public Health. In: Mastroianni AC, Kahn JP, Kass NE, eds. *The Oxford Handbook of Public Health Ethics*. Oxford University Press; 2019:353-367. doi:10.1093/oxfordhb/9780190245191.013.31
2. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data in Public Health. Bourne PE, ed. *PLOS Comput Biol*. 2015;11(2):e1003904. doi:10.1371/journal.pcbi.1003904
3. Goodman KW, Meslin EM. Ethics, Information Technology, and Public Health: Duties and Challenges in Computational Epidemiology. In: Magnuson JA, Fu P, eds. *Public Health Informatics and Information Systems*. 2nd ed. Springer; 2014:191–209.
4. Rothstein MA. Is Deidentification Sufficient to Protect Health Privacy in Research? *Am J Bioeth*. 2010;10(9):3-11. doi:10.1080/15265161.2010.494215
5. Gasser U. Perspectives on the Future of Digital Privacy. *ZSR II*. 2015;(134):426–427.

6. Sweeney L. *Simple Demographics Often Identify People Uniquely*.; 2000.

7. Patel CJ, Bhattacharya J, Ioannidis JPA, Bendavid E. Systematic identification of correlates of HIV infection. *AIDS*. 2018;32(7):933-943. doi:10.1097/QAD.0000000000001767

8. Bendavid E, Claypool K, Chow E, Chung J, Mai D, Patel C. The Demographic, Social, and Economic Correlates of HIV Infection Status in Sub-Saharan Africa. *Preprints*. Published online 2020. doi:10.20944/preprints202012.0507.v1

9. Staunton C, Moodley K. Challenges in biobank governance in Sub-Saharan Africa. *BMC Med Ethics*. 2013;14(1):35. doi:10.1186/1472-6939-14-35

10. Upshur RE, Lavery J V, Tindana PO. Taking tissue seriously means taking communities seriously. *BMC Med Ethics*. 2007;8(1):11. doi:10.1186/1472-6939-8-11

11. Virani AH, Longstaff H. Ethical Considerations in Biobanks: How a Public Health Ethics Perspective Sheds New Light on Old Controversies. *J Genet Couns*. 2015;24(3):428-432. doi:10.1007/s10897-014-9781-9

12. Barry M. Ethical Considerations of Human Investigation in Developing Countries. *N Engl J Med*. 1988;319(16):1083-1086. doi:10.1056/NEJM198810203191609

13. Denny SG, Silaigwana B, Wassenaar D, Bull S, Parker M. Developing Ethical Practices for Public Health Research Data Sharing in South Africa: The Views and Experiences From a Diverse Sample of Research Stakeholders. *J Empir Res Hum Res Ethics*. 2015;10(3):290-301. doi:10.1177/1556264615592386

14. Parker M, Bull S. Sharing Public Health Research Data. *J Empir Res Hum Res Ethics*. 2015;10(3):217-224. doi:10.1177/1556264615593494

15. Adhikari B, Pell C, Cheah PY. Community engagement and ethical global health research. *Glob Bioeth*. 2020;31(1):1-12. doi:10.1080/11287462.2019.1703504

16. Tindana P, de Vries J, Campbell M, et al. Community engagement strategies for genomic studies in Africa: a review of the literature. *BMC Med Ethics*. 2015;16(1):24. doi:10.1186/s12910-015-0014-z

17. Council for International Organizations of Medical Sciences, World Health Organization. *International Ethical Guidelines for Health-Related Research Involving Humans*.; 2016. http://www.sciencedirect.com/science/article/B6VC6-45F5X02-9C/2/e44bc37a6e392634b1cf436105978f01

18. King KF, Kolopack P, Merritt MW, Lavery J V. Community engagement and the human infrastructure of global health research. *BMC Med Ethics*. 2014;15(1):84. doi:10.1186/1472-6939-15-84

19. Dickert N, Sugarman J. Ethical Goals of Community Consultation in Research. *Am J Public Health*. 2005;95(7):1123-1127. doi:10.2105/AJPH.2004.058933

20. Marsh VM, Kamuya DK, Parker MJ, Molyneux CS. Working with Concepts: The Role of Community in International Collaborative Biomedical Research. *Public Health Ethics*. 2011;4(1):26-39. doi:10.1093/phe/phr007

21. Wilkinson A, Parker M, Martineau F, Leach M. Engaging 'communities': anthropological insights from the West African Ebola epidemic. *Philos Trans R Soc B Biol Sci*. 2017;372(1721):20160305. doi:10.1098/rstb.2016.0305

22. Nyirenda D, Sariola S, Kingori P, et al. Structural coercion in the context of community engagement in global health research conducted in a low resource setting in Africa. *BMC Med Ethics*. 2020;21(1):90. doi:10.1186/s12910-020-

00530-1

23. Tindana P, Campbell M, Marshall P, et al. Developing the science and methods of community engagement for genomic research and biobanking in Africa. *Glob Heal Epidemiol Genomics*. 2017;2:e13. doi:10.1017/gheg.2017.9

Table 1: Importance of ethical issues (Survey 1)

Item	Mean (1= not important at all, 5= absolutely essential) N=16	SD
Potential to stigmatize identifiable groups or populations	5.0	0.00
Privacy of individuals whose data are contained in the databases	4.9	0.35
Validity of big data analytic tools	4.4	0.50
Potential bias introduced by big data analytic tools	4.3	0.70
Alignment of the interests of scientists, funding agency, and the intended beneficiaries	4.1	0.70
Benefit sharing between scientists and survey respondents	4.0	0.90
Power and economic disparities between scientists and survey respondents	3.8	0.75

Table 2: Contextual factors – exemplar quotes from Survey 1 respondents

Need for engaging African scientists
"For instance, a South African HIV researcher would be knowledgeable about existing stigma relating to this condition and to any attributes of the population groups that could be identified through this research. He or she would likely be in a better position than someone who has never been here to assess whether and when particular kinds of scientific results would be likely to fuel existing stigma or discrimination. He or she would also have ongoing access to these communities and would likely have some insight into how such groups should be referred to in publications emanating from this research."
"The exclusion of African researchers from research about Africans, in my view, means that we do not maximize the opportunity to be effective."
"The Americans (and their funders) should be in Africa, training Africans in big data methods and tools."
Data on Africans being used by non-African researchers
"Countries in SSA are concerned with information being used by researchers abroad, and do not appreciate information being stored in servers outside of their countries, or extracted for analysis in abroad."

Table 3: Potential relevant stakeholders for engagement identified by expert panel

Regional level:	National level:	Local level:
African Academy of Sciences	Ministries of Health	Individuals
World Health Organization Regional Office for Africa (AFRO)	Universities	Communities
	Public health-related NGOs	Community advisory boards
	African public health policymakers	Religious leaders
	African scientists (clinical and public health scientists, biomedical researchers, and data scientists)	Traditional healers
	African healthcare workers	
	African ethicists	

BMJ Open

Diverse experts' perspectives on ethical issues of utilizing machine learning to predict HIV/AIDS risk in Sub-Saharan Africa: A modified Delphi study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-052287.R1
Article Type:	Original research
Date Submitted by the Author:	21-Jun-2021
Complete List of Authors:	Nichol, Ariadne; Stanford University School of Medicine, Center for Biomedical Ethics Bendavid, Eran; Stanford University, Mutenherwa, Farirai ; University of KwaZulu-Natal; University of KwaZulu-Natal, School of Applied Human Sciences Patel, Chirag; Harvard Medical School, Department of Biomedical Informatics Cho, Mildred; Stanford University School of Medicine, Center for Biomedical Ethics
Primary Subject Heading:	Ethics
Secondary Subject Heading:	Health informatics, HIV/AIDS, Public health
Keywords:	Public health < INFECTIOUS DISEASES, ETHICS (see Medical Ethics), Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, MEDICAL ETHICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Diverse experts' perspectives on ethical issues of utilizing machine learning to predict HIV/AIDS risk in Sub-Saharan Africa: A modified Delphi study

Ariadne A. Nichol,[1*] Eran Bendavid,[1] Farirai Mutenherwa,[2,3] Chirag Patel,[4] Mildred K. Cho[1]

[1] Stanford University School of Medicine, Stanford, California, United States

[2] School of Applied Human Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa

[3] KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

[4] Harvard Medical School, Boston, Massachusetts, United States

* Corresponding Author (Ariadne A. Nichol)

Address: 1215 Welch Road, Modular A, Stanford, CA 94305, United States

Email: ariadnen@stanford.edu

Word Count: 3,501

ABSTRACT (271 words)

Objective: To better understand diverse experts' views about the ethical implications of ongoing research funded by the National Institutes of Health that uses machine learning to predict HIV/AIDS risk in Sub-Saharan Africa based on publicly-available Demographic and Health Surveys data.

Design: Three rounds of semi-structured surveys in an online expert panel using a modified Delphi approach.

Participants: Experts in informatics, African public health and HIV/AIDS and bioethics were invited to participate.

Measures: Perceived importance of or agreement about relevance of ethical issues on 5-point uni-polar Likert scales. Qualitative data analysis identified emergent themes related to ethical issues and development of an ethical framework and recommendations for open-ended questions.

Results: Of the thirty-five invited experts, 22 participated in the online expert panel (63%). Emergent themes were the inclusion of African researchers in all aspects of study design, analysis, and dissemination to identify and address local contextual issues, as well as engagement of communities. Experts focused on engagement with health and science professionals to address risks, benefits, and communication of findings. Respondents prioritized the mitigation of stigma to research participants but recognized trade-offs between privacy and the need to disseminate findings to realize public health benefits. Strategies for responsible communication of results were suggested, including careful word choice in presentation of results and limited dissemination to need-to-know stakeholders such as public health planners.

Conclusion: Experts identified ethical issues specific to the African context and to research on sensitive, publicly-available data, and strategies for addressing these issues. These findings can be used to inform an ethical implementation framework with research stage-specific recommendations on how to utilize publicly-available data for machine-learning-based predictive analytics to predict HIV/AIDS risk in Sub-Saharan Africa.

Strengths and limitations of this study

- A strength of this study is that it represents the perspectives of diverse experts on the unique ethical issues raised by the use of predictive analytics for HIV/AIDS risk on large public health datasets in Sub-Saharan Africa.
- Another strength of the study is our use of open-ended questions and qualitative analysis of anonymously collected data to enhance breadth and validity of responses, and three rounds of iterative surveys to identify areas of disagreement.
- A third strength of the study is that it elicited specific suggestions from experts to navigate ethical tradeoffs, such as alternative methods of describing and disseminating findings of predictive analytics to minimize risks to privacy and of stigmatization, and suggestions for prioritizing specific groups for community engagement.

- The main limitation of this study is that a small number of respondents completed all three surveys, however, our expert respondents did represent diverse perspectives in informatics, bioethics of Africa-based studies, and African public health and HIV/AIDS.

INTRODUCTION

It is now well recognized that the use of big data for health research poses significant ethical challenges.(1–3) In particular, such research poses risks to the privacy of sensitive information as well as the potential for re-identification, stigmatization, and bias.(4–6) Many research cohort datasets with individual or patient-level information are available, such as those from epidemiological studies from biobanks (e.g., UK Biobank), repositories (such as dbGaP), and surveillance programs (e.g., Demographic and Health Surveys and US Centers for Disease Control and Prevention).

Several research studies aim to predict HIV/AIDS risk in Sub-Saharan African (SSA) countries using data from the Demographic and Health Surveys (DHS).(7,8) While there were no specific regulatory barriers to this research, it raised concerns for the researchers about whether existing ethical frameworks were adequate to address its specific constellation of characteristics (see Fig. 1, Supplemental Information). Namely, these included the particularly sensitive nature of HIV/AIDS, especially in SSA countries, the granularity of the data (including household wealth, educational history, marital status, and the location of households' villages or neighborhoods), the region's history of human rights abuses and exploitation, and the goal of predicting HIV/AIDS risk using easily ascertainable features. While many international regulations, guidelines, and conventions already apply to biomedical research(9–12), we sought to understand whether using new types of predictive analytics on sensitive, publicly available data raised additional issues that warranted special attention by researchers.

We therefore conducted a series of surveys of an expert panel with diverse expertise, including bioethics of Africa-based studies, informatics, and African public health and HIV/AIDS to better understand the ethical implications and concerns about this type of research and to inform an ethical framework and recommendations for researchers.

METHODS

Approach

Our overall approach was modeled after the Delphi method, but was heavily modified because our goal was not to achieve consensus but to document the range of perspectives of experts from diverse backgrounds about ethical issues and converge on recommendations for addressing them. Therefore, we relied largely on qualitative analysis, based on responses to open-ended questions to identify themes not already identified in the literature. We also asked closed-ended questions to better understand how individuals prioritized specific ethical issues and recommendations. We surveyed an expert panel in multiple rounds, building on responses to each round to develop the questions for the next one. We focused on identifying questions that required clarification or that indicated areas of disagreement that could be probed with more specificity in the subsequent survey.

Sample

Our multi-disciplinary research team, with backgrounds in bioethics, biomedical informatics, and public health in developing countries, identified 35 experts in informatics (n=10), African public health and HIV/AIDS (n=9) and bioethics of Africa-based studies (n=16) that were known to team members to have expertise in the context of public health or HIV/AIDS in Africa, through searches of the biomedical and ethics literature (again, focusing on public health, HIV/AIDS, and the African context), and by snowball sampling. All but one of the public health and bioethics experts were from African countries (Ethiopia, Ghana, Kenya, Nigeria, Rwanda, South Africa, Uganda, Zambia, Zimbabwe), and all of the informatics experts had their primary academic appointments in the United States, but did work on health in Africa. All panelists were English-speaking. Experts were invited by email and were offered US$200 for participation in all three surveys. Twenty-two agreed by email to participate (22/35=63%). Five actively declined, and eight did not respond to the initial invitation or to follow-up emails. We invited all 22 experts who agreed to participate in the panel to take Surveys 2 and 3, regardless of whether they had taken prior surveys. Because the surveys were anonymous, we do not know whether the same participants responded to each of the 3 surveys.

Surveys

We administered a series of three online, scenario-based semi-structured surveys, anonymously via Qualtrics, to make participation convenient and encourage frank responses. Respondents were allowed approximately three weeks to respond, with two reminder emails to all 22 who initially agreed to participate. The initial survey

was designed to capture a wide range of ethical issues, including those that might not have been already identified in the literature using broad open-ended questions, as well as to assess the perceived importance of previously-raised concerns. Responses were then analyzed to identify areas that were most frequently identified as important but where there was also disagreement about what to do. Subsequent survey questions were developed to identify how experts would prioritize values or make tradeoffs between conflicting values to address ethical issues.

Survey 1

Two research team members (MC and EB) developed the scenario for Survey 1 that was based on an actual research study funded by the National Institute of Allergy and Infectious Diseases at the US National Institutes of Health and conducted by some of the team members (Box 1). The scenario briefly describes aspects of the DHS survey datasets that are used but does not explicitly name them.

Box 1: Survey 1 scenario

A group of American scientists funded by the US government is developing big data tools to identify individuals and groups at elevated risk of acquiring HIV in Sub-Saharan Africa. The purpose of the project is to help ministries of health and international public health organizations target testing and treatment programs to the individuals and groups most at-risk. The scientists are using large, publicly-available datasets that identify the HIV status of millions of individuals, and hundreds of additional personal and household features of these individuals, some of which is collected by surveys. Household wealth, educational history, marital status, and the GPS coordinates of the households' village or neighborhood, among others, are characterized in detail. The data are readily available on the web for anyone who registers, and the source code for using the data and executing the HIV risk identification procedures are posted for public access. Policy makers in African countries have expressed interest in the findings, but have not specified how they plan to use the new information.

The survey began with three open-ended questions about 1) ethical issues they believed should be addressed by researchers conducting the study; 2) any details about the study that were not provided in the scenario but would be important to understanding the associated ethical issues; and 3) any specific recommendations for researchers conducting this or similar studies. We then asked respondents to rate the importance of seven ethical issues that we identified in the literature as potentially relevant to this scenario, using a 5-point uni-polar Likert scale ranging from 1=*Not important at all* to 5=*Absolutely essential*. Ethical issues included privacy, validity, power disparities, alignment and conflicts of interests, benefit-sharing, stigma, and bias (full item descriptions of the ethical issues can be found in Table 1). We specifically presented these seven issues after the open-ended questions in order to avoid anchoring or constraining open-ended responses, in hopes of eliciting a wide range of ethical issues and recommendations.

Survey 2

Responses to Survey 1 indicated that, in order to comment on ethical issues and to make recommendations, respondents needed more detail on how informed consent and ethical review processes were conducted for data collection for the DHS and for data use by individual researchers. As a result, we significantly expanded the description of the study for Survey 2 to include details on what data were collected and how data privacy, access and ethical review of the DHS survey were handled (Fig. 1, Supplemental Information). This survey's questions were open-ended, reflecting areas of consensus on importance that had emerged in the previous survey: 1) stakeholder engagement; 2) privacy/stigmatization/discrimination; 3) ethics review; 4) data access;

and 5) dissemination and communication of study findings (Fig. 2, Supplemental Information).

Survey 3

The findings from Survey 2 were used to design Survey 3 to probe areas of disagreement, and to elicit details that could inform draft recommendations about stakeholder engagement and ethics review. In Survey 3, we presented the same scenario as in Survey 2 (Fig. 1, Supplemental Information), but provided additional examples of analysis that could be conducted using the DHS data that highlighted the types of features that could be identified as risk factors using predictive analytics, and presented alternative ways of describing research findings that would pose tradeoffs between dissemination and privacy (Fig. 3, Supplemental Information). We then sought to clarify positions expressed by respondents in Survey 2 by focusing on policies and actions regarding: 1) who to include in stakeholder engagement (rating importance of each stakeholder on a Likert scale ranging from *Critically important to include* to *Do not include*); 2) strategies for dissemination of research findings to mitigate stigmatization and discrimination (rating level of agreement with statements on balancing privacy, stigmatization, and discrimination concerns with the dissemination of useful findings of risk factors for HIV/AIDS), and 3) requirements for ethical review (rating level of agreement with statements on the type of ethics review that is sufficient), with a closed-ended component and opportunity for open-ended explanation.

Qualitative data analysis

Responses to open-ended questions were analyzed as qualitative data. Statements were initially coded by one of the research team members (MC) to characterize the types of ethical issues or concerns that were raised, such as stigma, data ownership, or the need for stakeholder engagement. These codes were derived directly from the data. We then identified themes representing the most frequently occurring codes where there was lack of consensus or widely divergent views. SRQR reporting guidelines were used.(13)

Patient and public involvement

Patients and the public were not involved in research question development, study design, or analysis since the research specifically sought to elucidate experts' opinions on research utilizing big data for predicting HIV/AIDS. The expert panelists did propose appropriate approaches for community and public engagement and for disseminating sensitive research findings.

RESULTS

Survey 1

Of the 22 experts who agreed to participate in the panel, 16/22 (73%) responded to Survey 1 (overall response rate 16/35 = 46%). Because survey responses were anonymous, we do not know what proportion of respondents were experts in informatics, public health, or bioethics.

Open-ended responses were exceptionally rich, and reflected issues of re-identification, stigma, discrimination against individuals, families, or geographically defined and/or socially defined groups, especially pointing to the possibility of linking to HIV risk. These responses were consistent with the importance accorded these issues in the responses to the closed-ended questions which were asked later in the survey. As an example of stigma and discrimination, one respondent stated: “Perhaps the most concerning is the possibility of developing models that are based on source codes that could potentially stigmatize people, who will be labeled as 'at risk' individuals. Stigma is one of the most harmful conditions in HIV care today, and effective interventions are very hard to develop.” Respondents brought up several general ethical concerns commonly raised in relation to biobanking in SSA (though not unique to SSA), such as data ownership and access, data security and privacy, research priority setting, and benefit-sharing.(14–16)

Several responses to the question “Are there any details about this study that were not provided here that you feel would be important to understanding the ethical issues related to the study?” elicited questions about whether and how consent was obtained from data donors to use personal data, whether at the initial collection of DHS data or at the start of research utilizing machine learning predictive analytics to analyze the data. Therefore, in the subsequent survey, we made greater distinctions between consent and data use for DHS and for the HIV study.

The specific use of big data predictive analytics generated several ethical issues that respondents wanted to ensure were properly addressed prior to any research, including assessment of the potential for bias, independent review of the validity of the

predictive analytics tools, and establishment of a plan for monitoring interventions for harm that could result based on which individuals or groups were identified as being high risk for HIV/AIDS. Several respondents also emphasized the need for researchers to think through how the big data predictive analytics outcomes can be used to inform testing and treatment programs beyond simply identifying high-risk individuals or groups.

Respondents articulated a number of ethical issues that were not mentioned in the closed-ended questions, especially concerns about using DHS data sources to predict HIV/AIDS risks specific to the African context. Contextual factors cited (see Table 2 for exemplar quotes) included a history of human rights abuses, lack of trust in government, misuse of research findings, HIV-associated characteristics (e.g., homosexuality) that are crimes in some African countries, lack of expertise in big data analysis, lack of agency of African researchers and ethicists, compliance with or lack of country-specific laws and policies, and the need for engaging African scientists in order to provide contextual knowledge to inform best research and ethics practices. Another theme that emerged was concern about data on Africans being used by non-African researchers (see Table 2).

In responses to closed-ended questions (see Table 1) respondents rated almost all issues as “of average importance”, “very important”, or “absolutely essential” (6 of 7 issues had a mean rating of at least 4.0 on a scale of 1-5), and did not rate any of the 7 issues as *Not Important At All* or *Of Little Importance*. Nevertheless, two items clearly emerged as being most important. First was the potential to stigmatize groups or populations that are uniquely identified by the research (all rated this issue as

Absolutely Essential) and, second, the privacy of individuals (14 rated this as *Absolutely Essential*, 2 as *Very Important*). The next two most important issues identified were the validity of findings using big data tools, and potential for bias.

Survey 2

Ten of 22 experts responded to Survey 2 (10/22 = 45%), which presented only open-ended questions.

Overall, community and stakeholder engagement that includes Africans, ideally in relevant countries, was seen as key to minimizing risks at several stages of the research process, including data access, protocol oversight, and dissemination and implementation of findings. Some recommended engagement at the regional as well as national level, and respondents named a wide range of stakeholder groups (see Table 3). There was also broad support for community engagement in general to protect interests of local communities, groups and individuals. This engagement would provide the opportunity to better understand local concerns, values, norms, and cultural considerations and guide researchers on how to communicate findings in a way that mitigate risks to communities and individuals. Other purposes of stakeholder and community engagement were to provide education to public health officials and policymakers, clinicians, and communities, enhance buy-in, identify opportunities for capacity building and translation, and ultimately build trust and collaboration.

While Survey 1 indicated consensus on privacy as a primary concern, in Survey 2, statements about how researchers could address this issue were mixed. Some acknowledged limits on researchers' ability to prevent misuse of findings or to

completely protect data privacy; however, others also proposed specific actions to minimize harms. For example, one respondent said, “Of course there is nothing like absolute anonymization of data. I suggest that if sensitive results are obtained, it is imperative that the US research team works with communities in the affected countries on how best to disseminate the findings.” Another suggested, “Decide not to report data sub-groups containing very small numbers of individuals.”

There was a lack of consensus on the adequacy of centralized versus local ethics review and whether research on publicly-available or de-identified data was considered exempt from ethics review. Some respondents felt the centralized and local ethics review of the DHS surveys presented in the scenario would be adequate and the secondary data analysis of de-identified data would be exempt. However, one respondent articulated a differing view: “Ethics review from the regional and national bodies will be necessary… National ethics committee may be able to instill confidence that there is some oversight. Also any community and national level concerns may then be addressed.” Another respondent disagreed that research on de-identified data should be considered exempt and believed this protocol should “be reviewed (expedited review) by an IRB (ideally based in SSA)”. There was also disagreement about the adequacy of existing data access control and protection against stigma and discrimination from study findings. While one respondent suggested that data access controls were sufficient because data were de-identified, another would require “a clear data analysis and dissemination plan”, and another stated that protocol-specific data sharing agreements were necessary, because “Africa has suffered most from exploitation; both for research subjects and researchers.”

Survey 3

Ten experts responded (10/22 = 45%) to Survey 3, which primarily presented closed-ended questions, with space provided for participants to explain their responses. When asked which specific stakeholders would be critically important to include in stakeholder engagement, over half of participants believed African data scientists, African ethicists, representatives from a national Ministry of Health and representatives from African universities were necessary to include. Interestingly, responses were split (roughly in half) on whether African religious leaders and healers, African health workers and African patients and families were critically important to include or not important to include in stakeholder engagement. There were divergent opinions as to the necessity of representatives from local communities as well. One respondent articulated the concern that it might prove difficult to identify and engage with local communities with data coming from over one million people.

Representatives from the African regional WHO office (AFRO) and representatives from the African Academy of Science and public-health-related NGOs were viewed as stakeholders to include if resources were available, but not critical. For some participants, the stage of the study influenced which stakeholders they felt were relevant to engage. For example, community members only need to be engaged to minimize risks once the analysis is complete and agencies intend to take action based on results of the analysis.

When asked about the balance between the benefits of disseminating the research findings and risks of identification and stigma, there was support for some limitations on reporting to protect the identity of individuals or small groups (i.e. reporting

overall performance of predictive models rather than individual risk factors) but less agreement about restricting reporting of findings that could identify small numbers of individuals if the findings would be less useful for public health officials. There was broad agreement on the need for community representatives to have input on how risk factors are described in publications (e.g., if local geographic regions were to be mentioned in publications, community representatives would know whether this could lead to stigmatization against those relevant sub-populations), but there was less consensus as to whether it was necessary to obtain input from public health officials. There was strong disagreement with the proposed statement that researchers cannot do anything to protect against stigmatization based on risk factors. Several strategies on the communication of results were suggested, including reviewing and validating predictive models, careful word choice in the packaging of results, and limited dissemination to need-to-know stakeholders such as public health planners.

In clarifying the divergence of responses in Survey 2 on the amount of ethics review required for data collection and data analysis, a majority of respondents agreed that the combination of centralized ethics review of data collection and institutional ethics review of data analysis by the researchers' institution would be sufficient. Most respondents disagreed with the suggestions of requiring additional ethics review by all national research ethics committees of countries involved in data collection or additional ethics review by regional or African organizations.

DISCUSSION

It is increasingly recognized that the use of predictive analytics and artificial intelligence techniques such as machine learning on health data raises new ethical concerns or exacerbate existing issues. Most research to date has unearthed issues arising in high-income contexts, but we demonstrate here that many of these issues are salient for lower-income contexts as well. The rapid convergence and availability of new analytic methods and big data bring out issues arising from the use of such techniques on publicly-available datasets, especially with sensitive data such as HIV/AIDS status. We explored these issues as they pertained to actual US-funded research that uses data from individuals in SSA, a context that could sharpen ethical and social concerns. We demonstrate that issues of data privacy, stigma and discrimination, which are well-documented concerns of big data, were identified as key issues.(1,17,18) However, our expert panel largely agreed that the current practice of ethical review at the point of data collection and individual projects using large datasets was sufficient even in the SSA context.

While experts in our panel pointed to other problematic features of big data and predictive analytics such as bias, the preponderance of responses to open-ended questions highlighted ethical concerns that would apply to much of biomedical research generally, but with a focus on contextual factors. These factors included a history of human rights abuses, lack of trust in government and in non-African researchers, misuse of research findings, and obligations of US researchers to help build research capacity in Africa.(19)

On the other hand, there was some acknowledgement of the potential benefits from research presented in the scenario, as well as recognition of the inability to

maintain anonymity of research data. As a result, respondents were reluctant to support a complete block of the dissemination of findings. Respondents put forward a number of practical and feasible suggestions aimed at big data research, including privacy-preserving approaches for reporting the findings of predictive analytic models such as reporting overall performance of predictive models rather than individual risk factors. In addition, respondents suggested that benefits from research would be enhanced by validation of predictive analytic tools.

Our findings are consistent with previous studies that raised concerns over privacy, confidentiality, consent, and data misuse in the African context.(20–22) Our results demonstrate that consent and ensuring individual privacy and confidentiality were of primary concern to the expert panel, especially given use of predictive analytics.

Our findings mirror statements of others such as the H3Africa working group on ethics, who identified that community engagement is needed to support the informed consent process in the context of genomic research in Africa.(23) Others have stressed that community engagement in public health research in Africa is not only instrumental to recruitment and retainment of participants in research studies, but also intrinsically valuable as good ethical practice.(20) Divergent from these findings, we saw no explicit connection drawn between community engagement and informed consent. However, both topics were raised by our expert panel as separate issues that needed to be addressed.

Community engagement is seen as a critical part of health research in general, especially in SSA, given the history of exploitation.(12,22,24,25) Yet, there is extensive variation in defining what constitutes a “community”.(22,23,26,27) Our findings indicate

recognition of the need for engagement at national, regional, and local levels with a wide array of proposed participants. Interestingly, our panel of experts found other stakeholders (i.e. African ethicists, university researchers, data scientists, and representatives from ministries of health and universities) beyond local communities to be crucial to engage with in order to minimize risks of stigmatization and discrimination. It is important to consider the nature of predictive analytics and big data research as transnational and inclusive of many individuals' data. Therefore, this focus on broader stakeholders' engagement is explicable and perhaps a somewhat unique feature to research involving big data. Of course, these stakeholders have been identified as proposed participants of engagement for health research before, including within the context of genomics and biobanking in Africa.(14),(28),(29) Some have also suggested these stakeholders are in fact "community" in a broader interpretation.(22)

Public health data shared appropriately depends on "the trust and confidence of those from whom such data are derived and relate to".(20) Community and stakeholder engagement activities are key to developing such trust, and should be considered by researchers conducting studies using data from populations where trust has historically been threatened. The expert panel's recommendations around which stakeholders are essential to include can help researchers using predictive analytics and artificial intelligence engage with relevant communities.

One limitation of this study was the small number of respondents that completed all three surveys. However, the participants were experts in their respective fields of informatics, public health, HIV/AIDS, and bioethics in Africa, which increased

confidence in the insights reported. In addition, the informatics experts were largely US-based, although they had experience in working on HIV/AIDS and internationally.

CONCLUSION

Experts identified a number of ethical issues involved in carrying out research using big data predictive analytics to identify high-risk individuals or groups for HIV/AIDS in SSA. While many of these issues were not specific to big data or predictive analytics, our expert panel did focus on features specific to the SSA context, especially the inclusion of African researchers in all aspects of research. The expert panel offered strategies for navigating the trade-off between protection of privacy of sensitive and big data and dissemination of results, as well as priorities for which communities to involve in stakeholder engagement. Overall, the findings from this study can potentially inform an ethical implementation framework with research stage-specific recommendations on how to utilize machine-learning-based predictive analytics to predict risk of HIV/AIDS and other potentially sensitive conditions (such as COVID-19) in SSA. The recommendations could also be applicable to studies conducted in the context of serving historically disadvantaged or exploited groups more broadly.

FOOTNOTES

Contributors

EB and MKC contributed to the conception and design of the study. AAN and MKC completed the data collection and analysis and drafted the initial manuscript. EB, FM,

and CP drafted and finalized the manuscript with equal contributions. All authors contributed to drafts and approved the final manuscript.

Funding

This work was funded by the U.S. National Institutes of Health (1 R01 AI127250-01, Big Data Analysis of HIV Risk and Epidemiology in Sub-Saharan Africa). Eran Bendavid and Chirag Patel are Principal Investigators and Mildred Cho is a co-investigator of the grant. The views expressed are solely those of the authors.

Competing interests

None declared.

Patient consent for publication

Not required.

Ethics approval statement

Ethics approval for exemption was provided by the Stanford Human Subjects Panel (protocol number IRB-50813) and the reason for exemption was exemption #2 according to the common rule: Research involving the use of educational tests, survey procedures, interview procedures or observation of public behavior and information obtained is recorded by the investigator in such manner that the identity of the human subjects cannot readily be ascertained.

Data availability statement

Data are available upon reasonable request. Data are de-identified survey responses. Requests can be made to the corresponding author.

REFERENCES

1. Vayena E, Madoff L. Navigating the Ethics of Big Data in Public Health. In: Mastroianni AC, Kahn JP, Kass NE, editors. The Oxford Handbook of Public Health Ethics [Internet]. Oxford University Press; 2019. p. 353–67. Available from: http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190245191.001.0001/oxfordhb-9780190245191-e-31
2. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data in Public Health. Bourne PE, editor. PLOS Comput Biol [Internet]. 2015 Feb 9;11(2):e1003904. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003904
3. Goodman KW, Meslin EM. Ethics, Information Technology, and Public Health: Duties and Challenges in Computational Epidemiology. In: Magnuson JA, Fu P, editors. Public Health Informatics and Information Systems. 2nd ed. New York: Springer; 2014. p. 191–209.
4. Rothstein MA. Is Deidentification Sufficient to Protect Health Privacy in Research? Am J Bioeth [Internet]. 2010 Sep 9;10(9):3–11. Available from: http://www.tandfonline.com/doi/abs/10.1080/15265161.2010.494215
5. Gasser U. Perspectives on the Future of Digital Privacy. ZSR II. 2015;(134):426–427.

6. Sweeney L. Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper 3. Pittsburgh; 2000.

7. Patel CJ, Bhattacharya J, Ioannidis JPA, Bendavid E. Systematic identification of correlates of HIV infection. AIDS [Internet]. 2018 Apr 24;32(7):933–43. Available from: https://journals.lww.com/00002030-201804240-00013

8. Bendavid E, Claypool K, Chow E, Chung J, Mai D, Patel C. The Demographic, Social, and Economic Correlates of HIV Infection Status in Sub-Saharan Africa. Preprints. 2020;

9. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA [Internet]. 2013 Nov 27;310(20):2191–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24141714

10. European Union. CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION [Internet]. Official Journal of the European Union. 2012 [cited 2021 Jun 11]. Available from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:12012P/TXT

11. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Washington, D.C.; 1978.

12. Council for International Organizations of Medical Sciences, World Health Organization. International Ethical Guidelines for Health-related Research Involving Humans [Internet]. Council for International Organizations of Medical Sciences (CIOMS). 2016. 1–119 p. Available from:

http://www.sciencedirect.com/science/article/B6VC6-45F5X02-9C/2/e44bc37a6e392634b1cf436105978f01

13. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research. Acad Med [Internet]. 2014 Sep;89(9):1245–51. Available from: http://journals.lww.com/00001888-201409000-00021

14. Staunton C, Moodley K. Challenges in biobank governance in Sub-Saharan Africa. BMC Med Ethics [Internet]. 2013 Dec 11;14(1):35. Available from: https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-14-35

15. Upshur RE, Lavery J V, Tindana PO. Taking tissue seriously means taking communities seriously. BMC Med Ethics [Internet]. 2007 Dec 26;8(1):11. Available from: https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-8-11

16. Virani AH, Longstaff H. Ethical Considerations in Biobanks: How a Public Health Ethics Perspective Sheds New Light on Old Controversies. J Genet Couns [Internet]. 2015 Jun 29;24(3):428–32. Available from: http://doi.wiley.com/10.1007/s10897-014-9781-9

17. Enserink M, Chin G. The end of privacy. Science (80-) [Internet]. 2015 Jan 30;347(6221):490–1. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.347.6221.490

18. Beck EJ, Gill W, De Lay PR. Protecting the confidentiality and security of personal health information in low- and middle-income countries in the era of SDGs and Big Data. Glob Health Action [Internet]. 2016 Dec 1;9(1):32089. Available from: https://www.tandfonline.com/doi/full/10.3402/gha.v9.32089

19. Barry M. Ethical Considerations of Human Investigation in Developing Countries. N Engl J Med [Internet]. 1988 Oct 20;319(16):1083–6. Available from: http://www.nejm.org/doi/abs/10.1056/NEJM198810203191609

20. Denny SG, Silaigwana B, Wassenaar D, Bull S, Parker M. Developing Ethical Practices for Public Health Research Data Sharing in South Africa: The Views and Experiences From a Diverse Sample of Research Stakeholders. J Empir Res Hum Res Ethics [Internet]. 2015 Jul 21;10(3):290–301. Available from: http://journals.sagepub.com/doi/10.1177/1556264615592386

21. Parker M, Bull S. Sharing Public Health Research Data. J Empir Res Hum Res Ethics [Internet]. 2015 Jul 21;10(3):217–24. Available from: http://journals.sagepub.com/doi/10.1177/1556264615593494

22. Adhikari B, Pell C, Cheah PY. Community engagement and ethical global health research. Glob Bioeth [Internet]. 2020 Jan 1;31(1):1–12. Available from: https://www.tandfonline.com/doi/full/10.1080/11287462.2019.1703504

23. Tindana P, de Vries J, Campbell M, Littler K, Seeley J, Marshall P, et al. Community engagement strategies for genomic studies in Africa: a review of the literature. BMC Med Ethics [Internet]. 2015 Dec 12;16(1):24. Available from: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-015-0014-z

24. King KF, Kolopack P, Merritt MW, Lavery J V. Community engagement and the human infrastructure of global health research. BMC Med Ethics [Internet]. 2014 Dec 13;15(1):84. Available from: https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-15-84

25. Dickert N, Sugarman J. Ethical Goals of Community Consultation in Research.

Am J Public Health [Internet]. 2005 Jul;95(7):1123–7. Available from: http://ajph.aphapublications.org/doi/10.2105/AJPH.2004.058933

26. Marsh VM, Kamuya DK, Parker MJ, Molyneux CS. Working with Concepts: The Role of Community in International Collaborative Biomedical Research. Public Health Ethics [Internet]. 2011 Apr 1;4(1):26–39. Available from: https://academic.oup.com/phe/article-lookup/doi/10.1093/phe/phr007

27. Wilkinson A, Parker M, Martineau F, Leach M. Engaging 'communities': anthropological insights from the West African Ebola epidemic. Philos Trans R Soc B Biol Sci [Internet]. 2017 May 26;372(1721):20160305. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0305

28. Nyirenda D, Sariola S, Kingori P, Squire B, Bandawe C, Parker M, et al. Structural coercion in the context of community engagement in global health research conducted in a low resource setting in Africa. BMC Med Ethics [Internet]. 2020 Dec 21;21(1):90. Available from: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-020-00530-1

29. Tindana P, Campbell M, Marshall P, Littler K, Vincent R, Seeley J, et al. Developing the science and methods of community engagement for genomic research and biobanking in Africa. Glob Heal Epidemiol Genomics [Internet]. 2017 Sep 4;2:e13. Available from: https://www.cambridge.org/core/product/identifier/S2054420017000094/type/journal_article

Table 1: Importance of ethical issues (Survey 1)

Item	Mean (1= not important at all, 5= absolutely essential) N=16	SD
Potential to stigmatize identifiable groups or populations	5.0	0.00
Privacy of individuals whose data are contained in the databases	4.9	0.35
Validity of big data analytic tools	4.4	0.50
Potential bias introduced by big data analytic tools	4.3	0.70
Alignment of the interests of scientists, funding agency, and the intended beneficiaries	4.1	0.70
Benefit sharing between scientists and survey respondents	4.0	0.90
Power and economic disparities between scientists and survey respondents	3.8	0.75

Table 2: Contextual factors – exemplar quotes from Survey 1 respondents

Need for engaging African scientists
"For instance, a South African HIV researcher would be knowledgeable about existing stigma relating to this condition and to any attributes of the population groups that could be identified through this research. He or she would likely be in a better position than someone who has never been here to assess whether and when particular kinds of scientific results would be likely to fuel existing stigma or discrimination. He or she would also have ongoing access to these communities and would likely have some insight into how such groups should be referred to in publications emanating from this research."
"The exclusion of African researchers from research about Africans, in my view, means that we do not maximize the opportunity to be effective."
"The Americans (and their funders) should be in Africa, training Africans in big data methods and tools."
Data on Africans being used by non-African researchers
"Countries in SSA are concerned with information being used by researchers abroad, and do not appreciate information being stored in servers outside of their countries, or extracted for analysis in abroad."
History of human rights abuses
"How the researchers protect the privacy of these individuals would be critical considering the gross human rights abuses and poor legal frameworks in certain jurisdictions across Africa."
Lack of trust in government and potential for misuse of research findings
"The most important ethical consideration would be to ensure that the privacy of the individuals in the dataset is not compromised, and government officials have no way of tracing back individuals in the dataset up to the household level."
"Trust - Entrusting Ministries/governments could misuse the information - how can this be safeguarded. Information and political use - interventions may be denied where political support is low in some regions. Development of tools which could be abused by authorities or for political reasons."
HIV-associated characteristics (e.g., homosexuality) that are crimes in some African countries
"Since HIV infection is associated with homosexual behavior which is criminal in many SSA countries, individuals identified in the study may also be in legal jeopardy."
"How will these researchers ensure that their results will be used for good and not for harmful or discriminatory purposes, especially considering that e.g. same-

gender sexual relationships are illegal in many African countries, and that people who engage in them are actively persecuted in many?"
Lack of expertise in big data analysis
"Knowledge and understanding of what is big data - for ministries and for the populations."
Lack of agency of African researchers and ethicists
"There is lack of expertise in ethics review and monitoring research involving big data."
""The Americans (and their funders) should be in Africa, training Africans in big data methods and tools."
Compliance with or lack of country-specific laws and policies
"Consider laws in each region/country as these may differ significantly, or simply not exist in a functional format. Important to understand what local laws are available and what is constitutionally acceptable."
"…Information may have been deposited on an open source without permission, or in violation of the in-country laws."

611

Table 3: Potential relevant stakeholders for engagement identified by expert panel

Regional level:	National level:	Local level:
African Academy of Sciences	Ministries of Health	Individuals
World Health Organization Regional Office for Africa (AFRO)	Universities	Communities
	Public health-related NGOs	Community advisory boards
	African public health policymakers	Religious leaders
	African scientists (clinical and public health scientists, biomedical researchers, and data scientists)	Traditional healers
	African healthcare workers	
	African ethicists	

SUPPLEMENTAL INFORMATION

Figure 1: Survey 2 scenario

Research team: US-based scientists with expertise in infectious diseases and bioinformatics.
Funding source: US Department of Health and Human Services.

Rationale: HIV is the largest single cause of death among adults in Sub-Saharan Africa, responsible for about a fifth of all adult deaths in 2017. However, despite the dramatic increase in the availability of antiretroviral therapy, over 1.2 million people were newly infected in Sub-Saharan Africa in 2017, an incidence rate more than 10-fold higher than in the United States. A better understanding of the social, behavioral, environmental, and economic contexts that influence HIV risk could improve the effectiveness and efficiency of prevention and treatment programs.

Aims: The overall goal is to analyze large-scale datasets of HIV in Sub-Saharan Africa to identify new risk factors with potential to improve HIV care, and to help ministries of health and international public risk factors with potential to improve HIV care, and to help ministries of health and international public health organizations target testing and treatment programs.

Methods: The primary approach entails aligning HIV test results (positive or negative for HIV-1) with all social, economic, behavioral, and environmental features collected on individuals in the Demographic and Health Surveys (DHS). The DHS has completed home-based HIV testing on over 1,000,000 individuals in sub-Saharan Africa, and the entirety of the DHS information – over 1,000 potential predictors for the average person – is available for each individual, de-identified as described below (see section on Data privacy, access and ethical review, below). For all biomarker testing, verbal pre- and post-test counseling and printed information are provided to respondents, and test results are kept confidential. HIV-positive respondents are referred to a local health care facility for appropriate care. Analytic approaches include testing for association of HIV status with each of the predictors, as well as building sophisticated prediction models of HIV status using statistical learning approaches such as LASSO and Elastic Net.

Data sources: USAID Demographic and Health Surveys (DHS) from all Sub-Saharan African countries. All survey data are publicly available and are collected through a Household Questionnaire, and Individual Man's or Woman's Questionnaire, and a Biomarker Questionnaire. Household wealth, educational history, marital status, and the GPS coordinates of the households' village or neighborhood, among others, are characterized in detail. Biomarker testing for HIV status has been conducted in all endemic sub-Saharan countries since 2003.

Data privacy, access, and ethical review: Respondent interview and data files are initially identified by enumeration area (EA) and household numbers and then coversheets with these identifiers are destroyed and EA/household numbers are

randomly reassigned. Geographic coordinates of each survey are displaced in a random direction and distance up to 2 km (urban) or 5 km (rural) and randomly selected rural clusters displaced up to 10 km.

DHS questionnaires and general data collection procedures are reviewed and approved by an external Institutional Review Board (IRB) and country-specific protocols are reviewed and approved by an IRB from the individual country, which ensures that the survey complies with national laws and norms. Informed consent is conducted by interviewers in person, in a private location to provide privacy about sensitive topics, and includes a discussion of the purpose of the interview or test, privacy about sensitive topics, and includes a discussion of the purpose of the interview or test, expected duration, procedures, potential risks and benefits to the respondent, and contact information for a person who can provide more information. Consent for those undergoing HIV testing for DHS also explains that test results cannot be provided to individuals because names are not attached, but that a free voucher for health services that can provide HIV testing, and a list of local testing facilities is provided for study participants and their partners.

In order to access the DHS data, the US researchers registered for data access on the DHS website. Registration requires a project description and consent for maintaining the data secure and publishing only aggregated findings (i.e., not individual-level data). Once access was granted, the US researchers downloaded the data to secure servers with password protected access. The US researchers' protocol has been reviewed and approved by their university's IRB but is not considered human subjects research because it is considered research on an existing publicly- available, de-identified and non-coded dataset.

Figure 2: Survey 2 questions

Q1. Stakeholder and community engagement. A theme that emerged from responses to Survey 1 was the need for the researchers to engage stakeholders in the planning, design, analysis and dissemination of the research in order to identify and address contextual factors, including local laws and attitudes. The stakeholders included African scientists, ethicists, public health policy makers, and communities.

Given that the DHS data come from a large number of countries and are intended to be nationally representative, how would you suggest that the task of stakeholder engagement be approached, and by whom?

Q2. Privacy, stigmatization and discrimination. Data privacy was clearly identified in Survey 1 as the most important ethical concern about the HIV Big Data research project, primarily because of the potential for stigmatization of and discrimination against people with HIV/AIDS. Even though data obtained by the researchers have been stripped of explicit identifiers, and data have been randomly displaced geographically, re-identification of individuals, families, and groups defined by geographical or phenotypic characteristics could still be a concern because of the large amount of data collected about each individual. The US researchers have assured their IRB that they will not attempt to re- identify individuals or groups from the subset of DHS data that they have obtained, but risk factors that emerge from their analysis could be used to identify and thus stigmatize or discriminate against those with those characteristics.

How would you suggest that the US researchers minimize the chances that their identification of risk factors is misused?

Q3. Ethics review. Data collection for the DHS surveys was conducted with informed consent and with centralized ethics review of the general protocol and local review of country- specific protocols. Because the data are publicly available, the US researchers' IRB does not consider the secondary analysis of the data to be human subjects research. Although the US researchers obtained IRB approval from their university for their study, it was considered "exempt", so further review and informed consent was not required.

In Survey 1, some respondents expressed the need for ethics review. Do you believe that the centralized and local review of the DHS survey and by the US university sufficient? If not, what additional review should be instituted, by whom, and why?

Q4. Data access. The DHS dataset is publicly available but subject to some access control. Any requests for access to data must be approved by DHS staff. General approvals do not automatically guarantee access to the HIV data. Separate requests must be made to access both the general survey and HIV survey data.

Do you believe that this type of control of access to the DHS dataset is sufficient to

prevent misuse? If not, what additional controls would you recommend?

Q5. Study findings. The analysis of the DHS data is anticipated to identify a set of risk factors for acquisition of HIV.

Do you have any recommendations for the data analysts for how best to communicate what these risk factors are, assuming that the study findings will be disseminated to governmental and non-governmental public health organizations, other scientists, and to the general public?

Figure 3: Survey 3 examples added to research scenario

Here are **examples of data analyses** that could be conducted with DHS data:

These analyses would use data collected in 30 African countries, and include: the results of HIV tests from about 1,000,000 men and women between the ages of 15 and 49 (women) or 15 and 59 (men) who had consented to an HIV test; household data (e.g. floor material, water source, electricity); family information (marital status, number of children); health information (hemoglobin measurement, height and weight); family planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among others.

The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small subsets of characteristics. The predictors may or may not have been identified by previous epidemiological research, but may be strongly predictive. For example, bicycle ownership is, in some surveys, a strong predictor of HIV status, and adding it to a risk prediction model can improve prediction accuracy of HIV status from 82% to 85%.

One type of analysis would identify the individual features that are most closely tied to HIV status. This would have the potential to improve targeting of public health programs or help design interventions. For example, if widowhood is identified as a strong predictor of being HIV-positive, this can help design testing and prevention programs that are tailored to widows. This is similar to the identification of male circumcision as a risk factor that led to clinical trials and large-scale public health programs.

Another type of analysis would create risk scores that are a weighted combination of many individual features. This risk score would emerge from a commonly-used “black box” machine learning approach that chooses the combination of features that best predicts HIV status. The product of this analysis may not disclose any individual risk factors, and indeed some factors might only be predictive in combination with others. The analysis could report how well models predict the chance of being HIV-positive given a combination of features.

Standards for Reporting Qualitative Research (SRQR)*

http://www.equator-network.org/reporting-guidelines/srqr/

	Page no(s).
Title and abstract	
Title - Concise description of the nature and topic of the study Identifying the study as qualitative or indicating the approach (e.g., ethnography, grounded theory) or data collection methods (e.g., interview, focus group) is recommended	1
Abstract - Summary of key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results, and conclusions	2-3
Introduction	
Problem formulation - Description and significance of the problem/phenomenon studied; review of relevant theory and empirical work; problem statement	5-6
Purpose or research question - Purpose of the study and specific objectives or questions	5-6
Methods	
Qualitative approach and research paradigm - Qualitative approach (e.g., ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g., postpositivist, constructivist/ interpretivist) is also recommended; rationale**	6
Researcher characteristics and reflexivity - Researchers' characteristics that may influence the research, including personal attributes, qualifications/experience, relationship with participants, assumptions, and/or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results, and/or transferability	6-7
Context - Setting/site and salient contextual factors; rationale**	6-7
Sampling strategy - How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g., sampling saturation); rationale**	7-8
Ethical issues pertaining to human subjects - Documentation of approval by an appropriate ethics review board and participant consent, or explanation for lack thereof; other confidentiality and data security issues	11
Data collection methods - Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources/methods, and modification of procedures in response to evolving study findings; rationale**	7-11

Data collection instruments and technologies - Description of instruments (e.g., interview guides, questionnaires) and devices (e.g., audio recorders) used for data collection; if/how the instrument(s) changed over the course of the study	7-11
Units of study - Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results)	6-7
Data processing - Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymization/de-identification of excerpts	7-11
Data analysis - Process by which inferences, themes, etc., were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale**	10-11
Techniques to enhance trustworthiness - Techniques to enhance trustworthiness and credibility of data analysis (e.g., member checking, audit trail, triangulation); rationale**	10-11

Results/findings

Synthesis and interpretation - Main findings (e.g., interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory	11-17
Links to empirical data - Evidence (e.g., quotes, field notes, text excerpts, photographs) to substantiate analytic findings	12,15,28-30

Discussion

Integration with prior work, implications, transferability, and contribution(s) to the field - Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application/generalizability; identification of unique contribution(s) to scholarship in a discipline or field	17-20
Limitations - Trustworthiness and limitations of findings	20-21

Other

Conflicts of interest - Potential sources of influence or perceived influence on study conduct and conclusions; how these were managed	22
Funding - Sources of funding and other support; role of funders in data collection, interpretation, and reporting	22

*The authors created the SRQR by searching the literature to identify guidelines, reporting standards, and critical appraisal criteria for qualitative research; reviewing the reference lists of retrieved sources; and contacting experts to gain feedback. The SRQR aims to improve the transparency of all aspects of qualitative research by providing clear standards for reporting qualitative research.

**The rationale should briefly discuss the justification for choosing that theory, approach, method, or technique rather than other options available, the assumptions and limitations implicit in those choices, and how those choices influence study conclusions and transferability. As appropriate, the rationale for several items might be discussed together.

Reference:

O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. **Standards for reporting qualitative research: a synthesis of recommendations.** *Academic Medicine*, Vol. 89, No. 9 / Sept 2014 DOI: 10.1097/ACM.0000000000000388