

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<u>http://bmjopen.bmj.com</u>).

If you have any questions on BMJ Open's open peer review process please email <u>info.bmjopen@bmj.com</u>

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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

1		
2 3	1	
4 5	2	
6 7	2	
8 9	3	
10 11	4	Diverse experts' perspectives on ethical issues in predicting HIV/AIDS risk in
12 13	5	Sub-Saharan Africa
14 15 16	6	
17 18	7	Ariadne A. Nichol, ^{1*} Eran Bendavid, ¹ Farirai Mutenherwa, ^{2,3} Chirag Patel, ⁴
19 20	8	Mildred K. Cho ¹
21 22 23	9	
23 24 25	10	
26 27	11	¹ Stanford University School of Medicine, Stanford, California, United States
28 29	12	² School of Applied Human Sciences, University of KwaZulu-Natal, Pietermaritzburg,
30 31 32	13	South Africa
33 34	14	³ KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of
35 36	15	Health Sciences, University of KwaZulu-Natal, Durban, South Africa
37 38 30	16	⁴ Harvard Medical School, Boston, Massachusetts, United States
40 41	17	
42 43	18	
44 45 46	19	* Corresponding Author (Ariadne A. Nichol)
47 48	20	Address: 1215 Welch Road, Modular A, Stanford, CA 94305, United States
49 50	21	Email: ariadnen@stanford.edu
51 52	22	
53 54 55 56 57 58	23	Word Count: 3,501
59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 3 of 30

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.	Enseignement Superieur (ABES)	MJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I
--	-------------------------------	--

1		
2 3 4	24	ABSTRACT (271 words)
5 6	25	Objective: To better understand diverse experts' views about the ethical implications of
7 8 0	26	ongoing research funded by the National Institutes of Health that uses machine learning
9 10 11	27	to predict HIV/AIDS risk in Sub-Saharan Africa based on publicly-available
12 13	28	Demographic and Health Surveys data.
14 15	29	Design: Three rounds of semi-structured surveys in an online expert panel.
16 17 18	30	Participants: Experts in informatics, African public health and HIV/AIDS and bioethics
19 20	31	were invited to participate.
21 22	32	Measures: Perceived importance of or agreement about relevance of ethical issues on
23 24 25	33	5-point uni-polar Likert scales. Qualitative data analysis identified emergent themes
26 27 28 29	34	related to ethical issues and development of an ethical framework and
	35	recommendations for open-ended questions.
30 31 32	36	Results: Of the thirty-five invited experts, 22 participated in the online expert panel
33 34	37	(63%). Emergent themes were the inclusion of African researchers in all aspects of
35 36	38	study design, analysis, and dissemination to identify and address local contextual
37 38 30	39	issues, as well as engagement of communities. Experts focused on engagement with
40 41	40	health and science professionals to address risks, benefits, and communication of
42 43	41	findings. Respondents prioritized the mitigation of stigma to research participants but
44 45 46 47 48	42	recognized trade-offs between privacy and the need to disseminate findings to realize
	43	public health benefits. Strategies for responsible communication of results were
49 50	44	suggested, including careful word choice in presentation of results and limited
51 52 53 54 55 56	45	dissemination to need-to-know stakeholders such as public health planners.

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text and

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

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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1		B M C O
2 3 4	66	• The main limitation of this study is that a small number of respondents completed
5 6	67	all three surveys, however, our expert respondents did represent diverse 물
7 8 9	68	perspectives in informatics, bioethics of Africa-based studies, and African public
10 11	69	health and HIV/AIDS.
12 13	70	ected t
14 15 16	71	y cop
17 18	72	vright,
19 20	73	includ
21 22 23	74	ing for
23 24 25	75	Ense uses r
26 27	76	elated
28 29 20	77	to text
30 31 32	78	aded fr and dr
33 34	79	ABEE
35 36 37	80	j) - ming, A
37 38 39	81	l trainii
40 41	82	ng, ang ang
42 43	83	d simila
44 45 46	84	ar tech
47 48	85	nologi i
49 50	86	es. Age
51 52 53	87	
55 54 55	88	
56 57		a) phi q
58 59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

INTRODUCTION 89

5 6	90	
7 8	91	It is now well recognized that the use of big data for health research poses
9 10 11	92	significant ethical challenges. ^{1–3} In particular, such research poses risks to the privacy
12 13	93	of sensitive information as well as the potential for re-identification, stigmatization, and
14 15	94	bias. ^{4–6} Many research cohort datasets with individual or patient-level information are
16 17 18	95	available, such as those from epidemiological studies from biobanks (e.g., UK Biobank),
19 20	96	repositories (such as dbGaP), and surveillance programs (e.g., Demographic and
21 22	97	Health Surveys and US Centers for Disease Control and Prevention).
23 24 25	98	Several research studies aim to predict HIV/AIDS status in Sub-Saharan African
25 26 27	99	(SSA) countries using data from the Demographic and Health Surveys (DHS).7,8 While
28 29	100	there were no specific regulatory barriers to this research, it raised concerns for the
30 31 22	101	researchers about whether existing ethical frameworks were adequate to address its
32 33 34	102	specific constellation of characteristics. Namely, these included the particularly sensitive
35 36	103	nature of HIV/AIDS, especially in SSA countries, the region's history of human rights
37 38	104	abuses and exploitation, and the goal of predicting HIV/AIDS status using easily
39 40 41	105	ascertainable features.
42 43	106	We therefore conducted a series of surveys of an expert panel with diverse
44 45	107	expertise, including bioethics of Africa-based studies, informatics, and African public
46 47 48	108	health and HIV/AIDS to better understand the ethical implications and concerns about
49 50	109	this type of research and to inform an ethical framework and recommendations for
51 52	110	researchers.
53 54	111	
55 56 57		
58 59		5
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 2			
2 3 4	112	METHODS	
5 6	113		
7 8	114	Approach	
9 10 11	115	Our overall approach was modeled after the Delphi method, but was heavily	
12 13	116	modified because our goal was not to achieve consensus but to document the range of	i
14 15	117	perspectives of experts from diverse backgrounds about ethical issues and converge or	n
16 17 18	118	recommendations for addressing them. Therefore, we relied largely on qualitative	
19 20	119	analysis, based on responses to open-ended questions to identify themes not already	
21 22	120	identified in the literature. We also asked closed-ended questions to better understand	
23 24	121	how individuals prioritized specific ethical issues and recommendations. We surveyed	
25 26 27	122	an expert panel in multiple rounds, building on responses to each round to develop the	
28 29	123	questions for the next one. We focused on identifying questions that required	
30 31	124	clarification or that indicated areas of disagreement that could be probed with more	
32 33	125	specificity in the subsequent survey.	
34 35 36	126		
37 38	127	Sample	
39 40	128	We identified 35 experts in informatics (n=10), African public health and	
41 42 43	129	HIV/AIDS (n=9) and bioethics of Africa-based studies (n=16) through searches of the	
44 45	130	biomedical and ethics literature and by snowball sampling. All but one of the public	
46 47	131	health and bioethics experts were from African countries (Ethiopia, Ghana, Kenya,	
48 49 50	132	Nigeria, Rwanda, South Africa, Uganda, Zambia, Zimbabwe), and all of the informatics	
50 51 52	133	experts had their primary academic appointments in the United States, but did work on	
53 54 55	134	health in Africa. Experts were invited by email and were offered US\$200 for	
56 57			
58 59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	6

1 2		
2 3 4	135	participation in all three surveys. Twenty-two agreed by email to participate
5 6 7	136	(22/35=63%). Five actively declined, and eight did not respond to the initial invitation or
7 8 9	137	to follow-up emails.
10 11	138	
12 13	139	Surveys
14 15 16	140	We administered a series of three online, scenario-based semi-structured
17 18	141	surveys, anonymously via Qualtrics, to make participation convenient and encourage
19 20	142	frank responses. Respondents were allowed approximately three weeks to respond,
21 22 23	143	with two reminder emails to all 22 who initially agreed to participate.
23 24 25	144	Survey 1 began with a scenario describing an actual research study funded by
26 27	145	the National Institute of Allergy and Infectious Diseases at the US National Institutes of
28 29 20	146	Health (Box 1). The study utilizes large, publicly-available survey cohort data that
30 31 32	147	includes detailed health data and HIV status of millions of survey participants
33 34	148	throughout the world, socioeconomic data, and Global Positioning System (GPS)
35 36	149	coordinates of randomly displaced neighborhoods by up to 5km to protect privacy. Data
37 38 39	150	collection is conducted with informed consent by interviewers in person, and reviewed
40 41	151	and approved by an ethics review board in individual countries. Data users must
42 43	152	register for data access and undergo review by their local ethics review board (Box 1).
44 45 46		Box 1: Survey 1 scenario
47 48 49 50 51 52 53 54 55 56		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
57 58		

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. Polic makers in African countries have expressed interest in the findings, but have not specified how they plan to use the new information.	cy
The survey began with three open-ended questions about 1) ethical issues they	Protect
believed should be addressed by researchers conducting the study; 2) any details abo	but by
the study that were not provided in the scenario but would be important to	copyri
understanding the associated ethical issues; and 3) any specific recommendations for	ght, ind
researchers conducting this or similar studies. We then asked respondents to rate the	cluding
importance of seven ethical issues that we identified in the literature as potentially	y for us
relevant to this scenario, using a 5-point uni-polar Likert scale ranging from 1= <i>Not</i>	ses rela
important at all to 5=Absolutely essential. Ethical issues included privacy, validity, pow	ver to
disparities, alignment and conflicts of interests, benefit-sharing, stigma, and bias. We	text ar
specifically presented these seven issues after the open-ended questions in order to	nd data
avoid anchoring or constraining open-ended responses, in hopes of eliciting a wide	ı minin
range of ethical issues and recommendations.	g, Al tr
Surveys 2 and 3 were designed based on the responses to the previous survey	′S, ^{aini} ng
as described below in Results.	, and s
	imilar
Patient and public involvement	techno
Patients and the public were not involved in research question development,	ologies
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 di	d
For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	8

Page 10 of 30

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

3 4	173
5 6	174
7 8	175
9 10 11	176
12 13	177
14 15	178
16 17 18	179
19 20	180
21 22	181
23 24 25	182
26 27	183
28 29	184
30 31	185
32 33 34	186
35 36	187
37 38	188
39 40 41	189
41 42 43	190
44 45	191
46 47	192
48 49 50	193
50 51 52	194
53 54	195
55 56	
57 58	
59 60	

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

176 **RESULTS**

1 2

5 178 Survey 1

7179Of the 22 experts who agreed to participate in the panel, 16/22 (73%) responded180to Survey 1 (overall response rate 16/35 = 46%). Because survey responses were181anonymous, we do not know what proportion of respondents were experts in182informatics, public health, or bioethics. In responses to closed-ended questions (see183Table 1) respondents rated almost all issues as "of average importance", "very184important", or "absolutely essential" (6 of 7 issues had a mean rating of at least 4.0 on a185scale of 1-5), and did not rate any of the 7 issues as *Not Important At All* or *Of Little*186*Importance*. Nevertheless, two items clearly emerged as being most important. First187was the potential to stigmatize groups or populations that are uniquely identified by the188research (all rated this issue as *Absolutely Essential*) and, second, the privacy of189individuals (14 rated this as *Absolutely Essential*, 2 as *Very Important*). The next two190most important issues identified were the validity of findings using big data tools, and191potential for bias.

Open-ended responses were exceptionally rich, and reflected issues of reidentification, 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

1 2			
2 3 4	196	survey. Respondents brought up several general ethical concerns commonly raised in	
5 6	197	relation to biobanking in SSA, such as data ownership and access, data security and	
/ 8 9	198	privacy, research priority setting, and benefit-sharing.9-11	
9 10 11	199	Several responses raised concerns around individual autonomy and consent	
12 13	200	obtained to use personal data, whether at the initial collection of DHS data or at the start	
14 15 16	201	of research utilizing machine learning predictive analytics to analyze the data. These	
17 18	202	responses indicated that respondents needed more detail on how informed consent and	
19 20	203	ethical review processes were conducted for data collection for the DHS and for data	
21 22	204	use by individual researchers. As a result, we significantly expanded the description of	
23 24 25	205	the study for Survey 2 (Box 2).	
26 27	Box 2: Survey 2 scenario		
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 9 51 52 53 45		 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 	
57 58 59 60		10 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

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.

Page 13 of 30

BMJ Open

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%).

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de

> 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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 2		
3 4		required.
5 6 7 8		In Survey that the c sufficient
9 10 11 12 13 14		Q4. Data control. A approvals must be r
15 16 17 18		Do you b prevent n
19 20 21		Q5. Stud factors fo
22 23 24 25 26 27		Do you h what thes governme to the gel
27	236	
29 30	237	O
31 32 22	238	in relevar
33 34 35	239	research
36 37	240	implemer
38 39	241	national I
40 41 42	242	3). There
43 44	243	interests
45 46	244	the oppor
47 48	245	considera
49 50	246	mitigate r
52	247	communi
53 54	247	Commun
55 56		
57		
58 59		

Survey 1, some respondents expressed the need for ethics review. Do you believe t the centralized and local review of the DHS survey and by the US university ficient? If not, what additional review should be instituted, by whom, and why? **Data access.** The DHS dataset is publicly available but subject to some access trol. Any requests for access to data must be approved by DHS staff. General provals do not automatically guarantee access to the HIV data. Separate requests st be made to access both the general survey and HIV survey data. you believe that this type of control of access to the DHS dataset is sufficient to vent misuse? If not, what additional controls would you recommend? **Study findings.** The analysis of the DHS data is anticipated to identify a set of risk tors for acquisition of HIV. you have any recommendations for the data analysts for how best to communicate at these risk factors are, assuming that the study findings will be disseminated to vernmental and non-governmental public health organizations, other scientists, and he general public? Overall, community and stakeholder engagement that includes Africans, ideally elevant countries, was seen as key to minimizing risks at several stages of the earch process, including data access, protocol oversight, and dissemination and elementation of findings. Some recommended engagement at the regional as well as ional level, and respondents named a wide range of stakeholder groups (see Table There was also broad support for community engagement in general to protect rests of local communities, groups and individuals. This engagement would provide opportunity to better understand local concerns, values, norms, and cultural isiderations and guide researchers on how to communicate findings in a way that gate risks to communities and individuals. Other purposes of stakeholder and nmunity engagement were to provide education to public health officials and

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

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."

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

1		
2 3 4	271	
5 6	272	Survey 3
7 8 9	273	We invited all 22 experts who agreed to participate in the panel to take Survey 3,
10 11	274	regardless of whether they had taken Surveys 1 or 2. Ten experts responded (10/22 =
12 13	275	45%). Because the surveys were anonymous, we do not know whether these 10
14 15 16	276	respondents were the same as those who responded to Survey 2.
17 18	277	In Survey 3, we presented the same scenario as in Survey 2, providing additional
19 20	278	examples of analysis that could be conducted using the DHS data that highlighted the
21 22 22	279	types of features that could be identified as risk factors using predictive analytics, and
23 24 25	280	presented alternative ways of describing research findings that would pose tradeoffs
26 27	281	between dissemination and privacy (Box 4).
28		Box 4: Survey 3 examples added to research scenario
29 30		box 4. Our vey 5 examples added to research scenario
30 31 32		Here are examples of data analyses that could be conducted with DHS data:
 33 34 35 36 37 38 39 40 41 		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.
42 43 44 45 46 47 48 49		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%.
50 51 52 53 54 55		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
56 57 58 59		16

1 2

59

3 4		risk factor that led to clinical trials and large-scale public health programs.
5 6 7 8 9 10 11 12 13 14		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.
14 15	282	
16 17	283	We then sought to clarify positions expressed by respondents in Survey 2 by
18 19	284	focusing on policies and actions regarding: 1) who to include in stakeholder
20 21	285	engagement; 2) strategies for dissemination of research findings to mitigate
22 23 24	286	stigmatization and discrimination, and 3) requirements for ethical review, with a closed-
24 25 26	287	ended component and opportunity for open-ended explanation. Question 1 about
27 28	288	stakeholder engagement asked respondents to rate the importance of including each of
29 30	289	a list of potential stakeholders on a Likert scale ranging from Critically important to
31 32 33	290	include to Do not include. Questions 2 and 3 asked respondents to rate their level of
34 35	291	agreement with statements on balancing privacy, stigmatization, and discrimination
36 37	292	concerns with the dissemination of useful findings of risk factors for HIV/AIDS and their
38 39 40	293	level of agreement with statements on the level of ethics review that is sufficient.
40 41 42	294	When asked which specific stakeholders would be critically important to include
43 44	295	in stakeholder engagement, over half of participants believed African data scientists,
45 46	296	African ethicists, representatives from a national Ministry of Health and representatives
47 48 49	297	from African universities were necessary to include. Interestingly, responses were split
50 51	298	(roughly in half) on whether African religious leaders and healers, African health
52 53	299	workers and African patients and families were critically important to include or not
54 55 56	300	important to include in stakeholder engagement. There were divergent opinions as to
57 58		

Page 19 of 30

60

BMJ Open

1 2			
3 4	301	the necessity of representatives from local communities as well. This is perhaps in part	
5 6	302	due to the nature of the data having been collected within multiple countries, where	
/ 8 9	303	'local' might be difficult to define given the context. One respondent articulated the	
10 11	304	concern that it might prove difficult to identify and engage with local communities with	
12 13	305	data coming from over one million people.	
14 15 16	306	Representatives from the African regional WHO office (AFRO) and	
10 17 18	307	representatives from the African Academy of Science and public-health-related NGOs	
19 20	308	were viewed as stakeholders to include if resources were available, but not critical. For	
21 22 22	309	some participants, the stage of the study influenced which stakeholders they felt were	
23 24 25	310	relevant to engage. For example, community members only need to be engaged to	
26 27	311	minimize risks once the analysis is complete and agencies intend to take action based	
28 29	312	on results of the analysis.	
30 31 32	313	When asked about the balance between the benefits of disseminating the	
33 34	314	research findings and risks of identification and stigma, there was support for some	
35 36	315	limitations on reporting (i.e. reporting overall performance of predictive models rather	
37 38 30	316	than individual risk factors) but less agreement about restricting reporting of findings	
40 41	317	that could identify small numbers of individuals if the findings would be less useful for	
42 43	318	public health officials. There was broad agreement on the need for community	
44 45 46	319	representatives to have input on how risk factors are described in publications, but less	
40 47 48	320	so for input by public health officials. There was strong disagreement with the proposed	
49 50	321	statement that researchers cannot do anything to protect against stigmatization based	
51 52	322	on risk factors. Several strategies on the communication of results were suggested,	
53 54 55 56 57	323	including reviewing and validating predictive models, careful word choice in the	
58 59		1	8

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text and

BMJ Open

> 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

1 2		
3 4	347	panel largely agreed that the current practice of ethical review at the point of data
5 6	348	collection and individual projects using large datasets was sufficient even in the SSA
7 8 0	349	context.
9 10 11	350	While experts in our panel pointed to other problematic features of big data and
12 13	351	predictive analytics such as bias, the preponderance of responses to open-ended
14 15	352	questions highlighted ethical concerns that would apply to much of biomedical research
16 17 18	353	generally, but with a focus on contextual factors. These factors included a history of
19 20	354	human rights abuses, lack of trust in government and in non-African researchers,
21 22	355	misuse of research findings, and obligations of US researchers to help build research
23 24	356	capacity in Africa. ¹²
25 26 27 28 29	357	On the other hand, there was some acknowledgement of the potential benefits
	358	from research presented in the scenario, as well as recognition of the inability to
30 31	359	maintain anonymity of research data. As a result, respondents were reluctant to support
32 33 34	360	a complete block of the dissemination of findings. Respondents put forward a number of
35 36 37 38 39 40 41 42 43	361	practical and feasible suggestions aimed at big data research, including privacy-
	362	preserving approaches for reporting the findings of predictive analytic models such as
	363	reporting overall performance of predictive models rather than individual risk factors. In
	364	addition, respondents suggested that benefits from research would be enhanced by
44 45	365	validation of predictive analytic tools.
46 47	366	Our findings are consistent with previous studies that raised concerns over
48 49 50	367	privacy, confidentiality, consent, and data misuse in the African context. ^{13–15} Our results
51 52	368	demonstrate that consent and ensuring individual privacy and confidentiality were of
53 54	369	primary concern to the expert panel, especially given use of predictive analytics.
55 56 57		
57 58 59		20
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 22 of 30

BMJ Open

> 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

Page 23 of 30

59

60

1		
2 3 4	392	context of genomics and biobanking in Africa.9,22,23 Some have also suggested these
5 6	393	stakeholders are in fact "community" in a broader interpretation. ¹⁵
/ 8 0	394	Public health data shared appropriately depends on "the trust and confidence of
9 10 11	395	those from whom such data are derived and relate to". ¹³ Community and stakeholder
12 13	396	engagement activities are key to developing such trust, and should be considered by
14 15 16	397	researchers conducting studies using data from populations where trust has historically
16 17 18	398	been threatened. The expert panel's recommendations around which stakeholders are
19 20	399	essential to include can help researchers using predictive analytics and artificial
21 22	400	intelligence engage with relevant communities.
23 24 25	401	One limitation of this study was the small number of respondents that completed
26 27	402	all three surveys. However, the participants were experts in their respective fields of
28 29	403	informatics, public health, HIV/AIDS, and bioethics in Africa, which increased
30 31 32	404	confidence in the insights reported. In addition, the informatics experts were largely US-
33 34	405	based, although they had experience in working on HIV/AIDS and internationally.
35 36	406	
37 38 20	407	CONCLUSION
40 41	408	Experts identified a number of ethical issues involved in carrying out research
42 43	409	using big data predictive analytics to identify high-risk individuals or groups for
44 45	410	HIV/AIDS in SSA. While many of these issues were not specific to big data or predictive
46 47 48	411	analytics, our expert panel did focus on features specific to the SSA context, especially
49 50	412	the inclusion of African researchers in all aspects of research. The expert panel offered
51 52	413	strategies for navigating the trade-off between protection of privacy of sensitive and big
53 54 55 56 57 58	414	data and dissemination of results, as well as priorities for which communities to involve

1 2		
2 3 4	415	in stakeholder engagement. Overall, the findings from this study can potentially inform
5 6	416	an ethical implementation framework with research stage-specific recommendations on
7 8 9	417	how to utilize machine-learning-based predictive analytics to predict risk of HIV/AIDS
) 10 11	418	and other potentially sensitive conditions (such as COVID-19) in SSA. The
12 13	419	recommendations could also be applicable to studies conducted in the context of
14 15	420	serving historically disadvantaged or exploited groups more broadly.
16 17 18	421	
19 20	422	FOOTNOTES
21 22	423	Contributors
23 24 25	424	EB and MKC contributed to the conception and design of the study. AAN and MKC
26 27	425	completed the data collection and analysis and drafted the initial manuscript. EB, FM,
28 29	426	and CP drafted and finalized the manuscript with equal contributions. All authors
30 31 32	427	contributed to drafts and approved the final manuscript.
32 33 34	428	
35 36	429	Funding
37 38	430	This work was funded by the U.S. National Institutes of Health (1 R01 AI127250-01, Big
39 40 41	431	Data Analysis of HIV Risk and Epidemiology in Sub-Saharan Africa). Eran Bendavid
42 43	432	and Chirag Patel are Principal Investigators and Mildred Cho is a co-investigator of the
44 45	433	grant. The views expressed are solely those of the authors.
46 47 48	434	
40 49 50	435	Competing interests
51 52	436	None declared.
53 54	437	
55 56 57		
57 58		
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 2			
- 3 4	438	Patie	ent consent for publication
5 6	439	Not r	required.
7 8 9	440		
) 10 11	441	Data	availability statement
12 13	442	Data	are available upon reasonable request. Data are de-identified survey responses.
14 15 16	443	Requ	uests can be made to the corresponding author.
10 17 18	444		
19 20	445	REF	ERENCES
21 22	446	1.	Vayena E, Madoff L. Navigating the Ethics of Big Data in Public Health. In:
23 24 25	447		Mastroianni AC, Kahn JP, Kass NE, eds. The Oxford Handbook of Public Health
26 27	448		Ethics. Oxford University Press; 2019:353-367.
28 29	449		doi:10.1093/oxfordhb/9780190245191.013.31
30 31 32	450	2.	Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data
32 33 34	451		in Public Health. Bourne PE, ed. PLOS Comput Biol. 2015;11(2):e1003904.
35 36	452		doi:10.1371/journal.pcbi.1003904
37 38	453	3.	Goodman KW, Meslin EM. Ethics, Information Technology, and Public Health:
39 40 41	454		Duties and Challenges in Computational Epidemiology. In: Magnuson JA, Fu P,
42 43	455		eds. Public Health Informatics and Information Systems. 2nd ed. Springer;
44 45	456		2014:191–209.
46 47 48	457	4.	Rothstein MA. Is Deidentification Sufficient to Protect Health Privacy in Research?
49 50	458		Am J Bioeth. 2010;10(9):3-11. doi:10.1080/15265161.2010.494215
51 52	459	5.	Gasser U. Perspectives on the Future of Digital Privacy. ZSR II. 2015;(134):426-
53 54	460		427.
55 56 57			
58 59			24
60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1

58 59

2			
- 3 4	461	6.	Sweeney L. Simple Demographics Often Identify People Uniquely.; 2000.
5 6	462	7.	Patel CJ, Bhattacharya J, Ioannidis JPA, Bendavid E. Systematic identification of
7 8 0	463		correlates of HIV infection. AIDS. 2018;32(7):933-943.
9 10 11	464		doi:10.1097/QAD.000000000001767
12 13	465	8.	Bendavid E, Claypool K, Chow E, Chung J, Mai D, Patel C. The Demographic,
14 15	466		Social, and Economic Correlates of HIV Infection Status in Sub-Saharan Africa.
16 17 18	467		Preprints. Published online 2020. doi:10.20944/preprints202012.0507.v1
19 20	468	9.	Staunton C, Moodley K. Challenges in biobank governance in Sub-Saharan
21 22	469		Africa. BMC Med Ethics. 2013;14(1):35. doi:10.1186/1472-6939-14-35
23 24 25	470	10.	Upshur RE, Lavery J V, Tindana PO. Taking tissue seriously means taking
26 27	471		communities seriously. BMC Med Ethics. 2007;8(1):11. doi:10.1186/1472-6939-8-
28 29	472		11
30 31 22	473	11.	Virani AH, Longstaff H. Ethical Considerations in Biobanks: How a Public Health
32 33 34	474		Ethics Perspective Sheds New Light on Old Controversies. J Genet Couns.
35 36	475		2015;24(3):428-432. doi:10.1007/s10897-014-9781-9
37 38	476	12.	Barry M. Ethical Considerations of Human Investigation in Developing Countries.
39 40 41	477		N Engl J Med. 1988;319(16):1083-1086. doi:10.1056/NEJM198810203191609
42 43	478	13.	Denny SG, Silaigwana B, Wassenaar D, Bull S, Parker M. Developing Ethical
44 45	479		Practices for Public Health Research Data Sharing in South Africa: The Views
46 47 48	480		and Experiences From a Diverse Sample of Research Stakeholders. J Empir Res
49 50	481		Hum Res Ethics. 2015;10(3):290-301. doi:10.1177/1556264615592386
51 52	482	14.	Parker M, Bull S. Sharing Public Health Research Data. J Empir Res Hum Res
53 54 55	483		Ethics. 2015;10(3):217-224. doi:10.1177/1556264615593494
56 57			

Page 27 of 30

59

60

		BMJ O
al global healt:	h	oen: fir
019.1703504		st pub
nt strategies fo	or	lished
hics.		as 10. Prot
		1136/b ected
ld Health		mjope by cop
Research		n-2021 yright,
article/B6VC6-		-05228 includ
		7 on 2 ling foi
ement and the	÷	8 July Ense r uses
2014;15(1):84	4.	2021. [eignerr relatec
		Downlc Itent Su I to tex
n in Research.		aded f uperieu tand c
004.058933		rom ht Ir (ABE data mi
Concepts: The	e	tp://bn S) . ining, /
earch. Public		njopen Al trair
		.bmj.c 1ing, al
inities':		om/ on nd sim
Philos Trans R		June ' lar tec
6.0305		14, 202 hnoloç
context of		5 at Ag jies.
low resource		gence
2910-020-		Bibliog
		ıraphic
	26	lne de
Imm		

1 2			
2 3 4	484	15.	Adhikari B, Pell C, Cheah PY. Community engagement and ethical global health
5 6	485		research. Glob Bioeth. 2020;31(1):1-12. doi:10.1080/11287462.2019.1703504
/ 8 9	486	16.	Tindana P, de Vries J, Campbell M, et al. Community engagement strategies for
10 11	487		genomic studies in Africa: a review of the literature. BMC Med Ethics.
12 13	488		2015;16(1):24. doi:10.1186/s12910-015-0014-z
14 15 16	489	17.	Council for International Organizations of Medical Sciences, World Health
10 17 18	490		Organization. International Ethical Guidelines for Health-Related Research
19 20	491		Involving Humans.; 2016. http://www.sciencedirect.com/science/article/B6VC6-
21 22	492		45F5X02-9C/2/e44bc37a6e392634b1cf436105978f01
23 24 25	493	18.	King KF, Kolopack P, Merritt MW, Lavery J V. Community engagement and the
26 27	494		human infrastructure of global health research. BMC Med Ethics. 2014;15(1):84.
28 29	495		doi:10.1186/1472-6939-15-84
30 31 32	496	19.	Dickert N, Sugarman J. Ethical Goals of Community Consultation in Research.
33 34	497		Am J Public Health. 2005;95(7):1123-1127. doi:10.2105/AJPH.2004.058933
35 36	498	20.	Marsh VM, Kamuya DK, Parker MJ, Molyneux CS. Working with Concepts: The
37 38 30	499		Role of Community in International Collaborative Biomedical Research. Public
40 41	500		Health Ethics. 2011;4(1):26-39. doi:10.1093/phe/phr007
42 43	501	21.	Wilkinson A, Parker M, Martineau F, Leach M. Engaging 'communities':
44 45	502		anthropological insights from the West African Ebola epidemic. Philos Trans R
40 47 48	503		Soc B Biol Sci. 2017;372(1721):20160305. doi:10.1098/rstb.2016.0305
49 50	504	22.	Nyirenda D, Sariola S, Kingori P, et al. Structural coercion in the context of
51 52	505		community engagement in global health research conducted in a low resource
53 54 55 56 57 58	506		setting in Africa. BMC Med Ethics. 2020;21(1):90. doi:10.1186/s12910-020-

2 3	507		00530-1
4 5 6	508	23.	Tindana P, Campbell M, Marshall P, et al. Developing the science and methods of
7 8	509		community engagement for genomic research and biobanking in Africa. <i>Glob Heal</i>
9 10	510		<i>Epidemiol Genomics</i> . 2017;2:e13. doi:10.1017/gheg.2017.9
11 12	511		
13 14 15	512		
16 17			
18 19 20			
20 21 22			
23 24			
25 26 27			
28 29			
30 31			
32 33 34			
35 36			
37 38 20			
40 41			
42 43			
44 45 46			
40 47 48			
49 50			
51 52			
54 55			
56 57			
58 59			For peer review only - http://bmiopen.bmi.com/site/about/quidelines.xhtml 27
00			. e. peer retter en j integr, wingepenwingteen, site, as out guidennessmann

	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
15			

e in and a or ies o in w.	BMJ Open: first published as 10.1136/bmjopen-20; Protected by copyrigh
•••,	21-05228 ht, inclu
]	87 on 28 July En: ding for use
heir	2021. Down seignement \$ s related to to
29	aded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de l .perieur (ABES) . .t and data mining, Al training, and similar technologies.

<u>2</u>		
3 4 -	516 517	Table 2: Contextual factors – exemplar quotes from Survey 1 respondents
5		Need for engaging African scientists
7 8 9 10 11 12 13 14 15 16		"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."
17 18 19 20		"The exclusion of African researchers from research about Africans, in my view, means that we do not maximize the opportunity to be effective."
20 21 22 23		"The Americans (and their funders) should be in Africa, training Africans in big data methods and tools."
24		Data on Africans being used by non-African researchers
25 26 27 28		"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."
29		
30 31 32 33	518	
34 35 36		
37 38 39 40		
41 42 43		
44 45 46		
47 48 49		
50 51 52		
53 54 55		
50 57 58 59		

519 520	Table 3: Potential relevant stakeholders for engagement identified by expert pane

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
Ċ	African scientists (clinical and public health scientists, biomedical researchers, and data scientists)	Traditional healers
	African healthcare	
	African ethicists	

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

SCHOLARONE[™] Manuscripts



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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies



1		
2		
3	1	
4		
5	2	
7		
8	3	
9		
10	4	Diverse experts' perspectives on ethical issues of utilizing machine learning to
11		
12	5	predict HIV/AIDS risk in Sub-Saharan Africa: A modified Delphi study
13	-	provide the second s
14	6	
15	Ũ	
17	7	Ariadne A. Nichol ^{1*} Fran Bendavid ¹ Farirai Mutenherwa ^{2,3} Chirag Patel ⁴
18	/	
19	8	Mildred K. Cho ¹
20	0	
21	0	
22)	
25 24	10	
25	10	
26	11	1 Stanford University School of Medicine, Stanford, California, United States
27	11	
28	10	² School of Applied Human Sciences, University of KwaZulu Notel, Distormaritzburg
29	12	- School of Applied Human Sciences, Oniversity of Kwazulu-Natal, Fietermanizburg,
30	12	South Africa
31	13	South Amea
32 33	1 /	³ KwaZulu Natal Desearch Innovation and Sequencing Diatform (KDISD). College of
34	14	* Kwazulu-Nalal Research Innovation and Sequencing Platform (KRISP), College of
35	15	Health Sciences, University of KwoZulu Notel, Durban, South Africa
36	13	Health Sciences, University of Kwazulu-Natal, Durban, South Anica
37	16	4 Harvard Madical School, Bacton, Magaachusatta, United States
38	10	Tarvard Medical School, Boston, Massachusetts, United States
39	17	
40 41	1/	
42	10	
43	18	
44	10	* Opmannending Authon (Ariedon A. Nichel)
45	19	" Corresponding Author (Ariadne A. Nichol)
46	•	
4/	20	Address: 1215 Weich Road, Modular A, Stanford, CA 94305, United States
48 70	• •	
49 50	21	Email: ariadnen@stanford.edu
51		
52	22	
53	_	
54	23	Word Count: 3,501
55		
56 57		
57 58		
50 59		1
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml
60

1 2		
2 3 4	24	ABSTRACT (271 words)
5 6	25	Objective: To better understand diverse experts' views about the ethical implications of
7 8	26	ongoing research funded by the National Institutes of Health that uses machine learning
9 10 11	27	to predict HIV/AIDS risk in Sub-Saharan Africa based on publicly-available
12 13	28	Demographic and Health Surveys data.
14 15	29	Design: Three rounds of semi-structured surveys in an online expert panel using a
16 17 18	30	modified Delphi approach.
19 20	31	Participants: Experts in informatics, African public health and HIV/AIDS and bioethics
21 22	32	were invited to participate.
23 24 25	33	Measures: Perceived importance of or agreement about relevance of ethical issues on
26 27	34	5-point uni-polar Likert scales. Qualitative data analysis identified emergent themes
28 29	35	related to ethical issues and development of an ethical framework and
30 31 22	36	recommendations for open-ended questions.
32 33 34	37	Results: Of the thirty-five invited experts, 22 participated in the online expert panel
35 36	38	(63%). Emergent themes were the inclusion of African researchers in all aspects of
37 38	39	study design, analysis, and dissemination to identify and address local contextual
39 40 41	40	issues, as well as engagement of communities. Experts focused on engagement with
42 43	41	health and science professionals to address risks, benefits, and communication of
44 45	42	findings. Respondents prioritized the mitigation of stigma to research participants but
46 47 48	43	recognized trade-offs between privacy and the need to disseminate findings to realize
48 49 50	44	public health benefits. Strategies for responsible communication of results were
51 52	45	suggested, including careful word choice in presentation of results and limited
53 54 55 56 57 58	46	dissemination to need-to-know stakeholders such as public health planners.

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text and

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.

1 2		
3 4	67 •	The main limitation of this study is that a small number of respondents completed
5 6	68	all three surveys, however, our expert respondents did represent diverse
/ 8 9	69	perspectives in informatics, bioethics of Africa-based studies, and African public
10 11	70	health and HIV/AIDS.
12 13	71	
14 15	72	
16 17 18	73	
19 20	74	
21 22	75	
23 24 25	76	
26 27	77	
28 29	78	
30 31 32	79	
33 34	80	
35 36	81	
37 38 39	82	
40 41	83	
42 43	84	
44 45 46	85	
47 48	86	
49 50	87	
51 52	88	
53 54 55	89	
56 57		
58 59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

INTRODUCTION

6	91	
7 8 9	92	It is now well recognized that the use of big data for health research poses
9 10 11	93	significant ethical challenges.(1–3) In particular, such research poses risks to the
12 13	94	privacy of sensitive information as well as the potential for re-identification,
14 15 16	95	stigmatization, and bias.(4–6) Many research cohort datasets with individual or patient-
10 17 18	96	level information are available, such as those from epidemiological studies from
19 20	97	biobanks (e.g., UK Biobank), repositories (such as dbGaP), and surveillance programs
21 22	98	(e.g., Demographic and Health Surveys and US Centers for Disease Control and
23 24 25	99	Prevention).
26 27	100	Several research studies aim to predict HIV/AIDS risk in Sub-Saharan African
28 29	101	(SSA) countries using data from the Demographic and Health Surveys (DHS).(7,8)
30 31 32	102	While there were no specific regulatory barriers to this research, it raised concerns for
33 34	103	the researchers about whether existing ethical frameworks were adequate to address its
35 36	104	specific constellation of characteristics (see Fig. 1, Supplemental Information). Namely,
37 38 39	105	these included the particularly sensitive nature of HIV/AIDS, especially in SSA
40 41	106	countries, the granularity of the data (including household wealth, educational history,
42 43	107	marital status, and the location of households' villages or neighborhoods), the region's
44 45 46	108	history of human rights abuses and exploitation, and the goal of predicting HIV/AIDS
40 47 48	109	risk using easily ascertainable features. While many international regulations,
49 50	110	guidelines, and conventions already apply to biomedical research(9–12), we sought to
51 52	111	understand whether using new types of predictive analytics on sensitive, publicly
55 55	112	available data raised additional issues that warranted special attention by researchers.
56 57		
58 59		5

1 2		
2 3 4	113	We therefore conducted a series of surveys of an expert panel with diverse
5 6 7	114	expertise, including bioethics of Africa-based studies, informatics, and African public
7 8	115	health and HIV/AIDS to better understand the ethical implications and concerns about
9 10 11	116	this type of research and to inform an ethical framework and recommendations for
12 13	117	researchers.
14 15	118	
16 17 18	119	METHODS
19 20	120	
21 22	121	Approach
23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38	122	Our overall approach was modeled after the Delphi method, but was heavily
	123	modified because our goal was not to achieve consensus but to document the range of
	124	perspectives of experts from diverse backgrounds about ethical issues and converge on
	125	recommendations for addressing them. Therefore, we relied largely on qualitative
	126	analysis, based on responses to open-ended questions to identify themes not already
	127	identified in the literature. We also asked closed-ended questions to better understand
	128	how individuals prioritized specific ethical issues and recommendations. We surveyed
39 40 41	129	an expert panel in multiple rounds, building on responses to each round to develop the
42 43	130	questions for the next one. We focused on identifying questions that required
44 45	131	clarification or that indicated areas of disagreement that could be probed with more
46 47 48	132	specificity in the subsequent survey.
40 49 50	133	
51 52	134	Sample
53 54		
55 56		
57		
58 59		6

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

2		
3 4	158	was designed to capture a wide range of ethical issues, including those that might not
5 6	159	have been already identified in the literature using broad open-ended questions, as well
7 8	160	as to assess the perceived importance of previously-raised concerns. Responses were
9 10 11	161	then analyzed to identify areas that were most frequently identified as important but
12 13	162	where there was also disagreement about what to do. Subsequent survey questions
14 15 16	163	were developed to identify how experts would prioritize values or make tradeoffs
10 17 18	164	between conflicting values to address ethical issues.
19 20	165	Survey 1
21 22 22	166	Two research team members (MC and EB) developed the scenario for Survey 1
25 24 25	167	that was based on an actual research study funded by the National Institute of Allergy
26 27	168	and Infectious Diseases at the US National Institutes of Health and conducted by some
28 29 20	169	of the team members (Box 1). The scenario briefly describes aspects of the DHS survey
30 31 32	170	datasets that are used but does not explicitly name them.
33 34	171	
35 36 27		Box 1: Survey 1 scenario
38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55	172	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.
56 57 58 59 60		8 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 10 of 39

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

3	
4	
5	
6	
7	
, Q	
0	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
20 21	
∠ I วา	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
25	
22	
30	
3/	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
<u>4</u> 0	
79 50	
50	
51	
52	
53	
54	
55	
56	
57	
58	
59	

60

1 2

173 The survey began with three open-ended questions about 1) ethical issues they 174 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 175 176 understanding the associated ethical issues; and 3) any specific recommendations for 177 researchers conducting this or similar studies. We then asked respondents to rate the 178 importance of seven ethical issues that we identified in the literature as potentially 179 relevant to this scenario, using a 5-point uni-polar Likert scale ranging from 1=Not 180 important at all to 5=Absolutely essential. Ethical issues included privacy, validity, power 181 disparities, alignment and conflicts of interests, benefit-sharing, stigma, and bias (full 182 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 183 184 anchoring or constraining open-ended responses, in hopes of eliciting a wide range of 185 ethical issues and recommendations. 186 Survey 2 Responses to Survey 1 indicated that, in order to comment on ethical issues and 187 188 to make recommendations, respondents needed more detail on how informed consent 189 and ethical review processes were conducted for data collection for the DHS and for 190 data use by individual researchers. As a result, we significantly expanded the 191 description of the study for Survey 2 to include details on what data were collected and 192 how data privacy, access and ethical review of the DHS survey were handled (Fig. 1,

193 Supplemental Information). This survey's questions were open-ended, reflecting areas

194 of consensus on importance that had emerged in the previous survey: 1) stakeholder

195 engagement; 2) privacy/stigmatization/discrimination; 3) ethics review; 4) data access;

BMJ Open

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

198 <u>Survey 3</u>

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.

49 216

Qualitative data analysis

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

3 4	2
5 6	2
7 8	2
9 10 11	2
12 13	2
14 15	2
16 17	2
18 19 20	2
21 22	2
23 24	2
25 26 27	2
28 29	2
30 31	2
32 33	2
34 35 36	2
37 38	2
39 40	2
41 42 43	2
44 45	2
46 47	2
48 49 50	2
50 51 52	2
53 54	
55 56	
57 58 50	
60	

218	Responses to open-ended questions were analyzed as qualitative data.
219	Statements were initially coded by one of the research team members (MC) to
220	characterize the types of ethical issues or concerns that were raised, such as stigma,
221	data ownership, or the need for stakeholder engagement. These codes were derived
222	directly from the data. We then identified themes representing the most frequently
223	occurring codes where there was lack of consensus or widely divergent views. SRQR
224	reporting guidelines were used.(13)
225	
226	Patient and public involvement
227	Patients and the public were not involved in research question development,
228	study design, or analysis since the research specifically sought to elucidate experts'
229	opinions on research utilizing big data for predicting HIV/AIDS. The expert panelists did
230	propose appropriate approaches for community and public engagement and for
231	disseminating sensitive research findings.
232	
233	RESULTS
234	
235	Survey 1
236	Of the 22 experts who agreed to participate in the panel, 16/22 (73%) responded
237	to Survey 1 (overall response rate 16/35 = 46%). Because survey responses were
238	anonymous, we do not know what proportion of respondents were experts in
239	informatics, public health, or bioethics.

Page 13 of 39

BMJ Open

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.

260 The specific use of big data predictive analytics generated several ethical issues 261 that respondents wanted to ensure were properly addressed prior to any research, 262 including assessment of the potential for bias, independent review of the validity of the

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Page 14 of 39

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 15 of 39

BMJ Open

	do Ch
Absolutely Essential) and, second, the privacy of individuals (14 rated this as Absolute	≱ly ^e n: fir
Essential, 2 as Very Important). The next two most important issues identified were th	e st pub
validity of findings using big data tools, and potential for bias.	lished
	as 10. Prot
Survey 2	1136/b ected I
Ten of 22 experts responded to Survey 2 (10/22 = 45%), which presented only	mjoper oy cop
open-ended questions.	ז-2021 yright,
Overall, community and stakeholder engagement that includes Africans, ideally	-05228 includ
in relevant countries, was seen as key to minimizing risks at several stages of the	7 on 28 ling for
research process, including data access, protocol oversight, and dissemination and	3 July Ense uses
implementation of findings. Some recommended engagement at the regional as well a	נוסבין. ב Pignem יelated
national level, and respondents named a wide range of stakeholder groups (see Table	ent Su to tex
3). There was also broad support for community engagement in general to protect	aded fr perieu t and d
interests of local communities, groups and individuals. This engagement would provid	r (ABE lata mi
the opportunity to better understand local concerns, values, norms, and cultural	tp://bm S) . ning, /
considerations and guide researchers on how to communicate findings in a way that	ijopen. Al train
mitigate risks to communities and individuals. Other purposes of stakeholder and	bmj.cc ing, an
community engagement were to provide education to public health officials and	om/ on Id simi
policymakers, clinicians, and communities, enhance buy-in, identify opportunities for	June 1 lar tec
capacity building and translation, and ultimately build trust and collaboration.	4, 202 hnolog
While Survey 1 indicated consensus on privacy as a primary concern, in Surve	5 at Ag jies.
2, statements about how researchers could address this issue were mixed. Some	yence I
acknowledged limits on researchers' ability to prevent misuse of findings or to	Bibliog
	raphiq
For neer review only - http://bmionon.hmi.com/cito/about/quidalines.yhtml	14 de
For peer review only integrating periodification site about guidelines. And the	

Ξ

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

> 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."

59

60

d nses.	BMJ Open: first published as 10.1136/br Protected b
ives	njope у сор
split	n-2021-052287 on 2 yright, including fo
to	8 July 2021. Downloaded from h Enseignement Superieur (ABE r uses related to text and data m
Os	ttp://b ES) . iining,
For	njope Al trai
ere	n.bmj.c ning, a
)	nd sin
sed	n June 14, nilar techn
	2025 at A ologies.
;	gence
orting	Bibliographic
16	que de l

3 4	332	Survey 3
5 6	333	Ten experts responded (10/22 = 45%) to Survey 3, which primarily presented
7 8 0	334	closed-ended questions, with space provided for participants to explain their response
9 10 11	335	When asked which specific stakeholders would be critically important to include in
12 13	336	stakeholder engagement, over half of participants believed African data scientists,
14 15	337	African ethicists, representatives from a national Ministry of Health and representative
16 17 18	338	from African universities were necessary to include. Interestingly, responses were spli
19 20	339	(roughly in half) on whether African religious leaders and healers, African health
21 22	340	workers and African patients and families were critically important to include or not
23 24 25	341	important to include in stakeholder engagement. There were divergent opinions as to
26 27	342	the necessity of representatives from local communities as well. One respondent
28 29	343	articulated the concern that it might prove difficult to identify and engage with local
30 31 32	344	communities with data coming from over one million people.
33 34	345	Representatives from the African regional WHO office (AFRO) and
35 36	346	representatives from the African Academy of Science and public-health-related NGOs
37 38 30	347	were viewed as stakeholders to include if resources were available, but not critical. Fo
40 41	348	some participants, the stage of the study influenced which stakeholders they felt were
42 43	349	relevant to engage. For example, community members only need to be engaged to
44 45 46	350	minimize risks once the analysis is complete and agencies intend to take action based
40 47 48	351	on results of the analysis.
49 50	352	When asked about the balance between the benefits of disseminating the
51 52	353	research findings and risks of identification and stigma, there was support for some
55 54 55 56 57 58	354	limitations on reporting to protect the identity of individuals or small groups (i.e. reporti

Page 18 of 39

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text and

BMJ Open

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

Page 19 of 39

BMJ Open

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 welldocumented 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

395 generally, but with a focus on contextual factors. These factors included a history of

guestions highlighted ethical concerns that would apply to much of biomedical research

396 human rights abuses, lack of trust in government and in non-African researchers,

397 misuse of research findings, and obligations of US researchers to help build research398 capacity in Africa.(19)

399 On the other hand, there was some acknowledgement of the potential benefits400 from research presented in the scenario, as well as recognition of the inability to

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

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 privacypreserving 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.

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

Page 21 of 39

BMJ Open

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

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

446 confidence in the insights reported. In addition, the informatics experts were largely US447 based, although they had experience in working on HIV/AIDS and internationally.

449 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,

1 2						
3 4	468	and CP drafted and finalized the manuscript with equal contributions. All authors				
5 6	469	contributed to drafts and approved the final manuscript.				
/ 8 0	470					
10 11 12 13 14 15	471	Funding				
	472	This work was funded by the U.S. National Institutes of Health (1 R01 AI127250-01, Big				
	473	Data Analysis of HIV Risk and Epidemiology in Sub-Saharan Africa). Eran Bendavid				
16 17 18	474	and Chirag Patel are Principal Investigators and Mildred Cho is a co-investigator of the				
19 20	475	grant. The views expressed are solely those of the authors.				
21 22	476					
23 24 25	477	Competing interests				
26 27	478	None declared.				
28 29	479					
30 31 32	480	Patient consent for publication				
32 33 34	481	Not required.				
35 36	482					
37 38	483	Ethics approval statement				
39 40 41	484	Ethics approval for exemption was provided by the Stanford Human Subjects Panel				
42 43	485	(protocol number IRB-50813) and the reason for exemption was exemption #2				
44 45	486	according to the common rule: Research involving the use of educational tests, survey				
46 47 48	487	procedures, interview procedures or observation of public behavior and information				
49 50	488	obtained is recorded by the investigator in such manner that the identity of the human				
51 52	489	subjects cannot readily be ascertained.				
53 54	490					
56 57						
58 59		22				
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml				

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

1 2			
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40	491	Data	a availability statement
	492	Data	a are available upon reasonable request. Data are de-identified survey responses.
	493	Req	uests can be made to the corresponding author.
	494		
	495	REF	ERENCES
	496	1.	Vayena E, Madoff L. Navigating the Ethics of Big Data in Public Health. In:
	497		Mastroianni AC, Kahn JP, Kass NE, editors. The Oxford Handbook of Public
	498		Health Ethics [Internet]. Oxford University Press; 2019. p. 353–67. Available from:
	499		http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780190245191.001.0001/oxf
	500		ordhb-9780190245191-e-31
	501	2.	Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data
	502		in Public Health. Bourne PE, editor. PLOS Comput Biol [Internet]. 2015 Feb
	503		9;11(2):e1003904. Available from:
	504		https://dx.plos.org/10.1371/journal.pcbi.1003904
	505	3.	Goodman KW, Meslin EM. Ethics, Information Technology, and Public Health:
	506		Duties and Challenges in Computational Epidemiology. In: Magnuson JA, Fu P,
	507		editors. Public Health Informatics and Information Systems. 2nd ed. New York:
42 43	508		Springer; 2014. p. 191–209.
44 45	509	4.	Rothstein MA. Is Deidentification Sufficient to Protect Health Privacy in Research?
46 47 48	510		Am J Bioeth [Internet]. 2010 Sep 9;10(9):3–11. Available from:
49 50	511		http://www.tandfonline.com/doi/abs/10.1080/15265161.2010.494215
51 52	512	5.	Gasser U. Perspectives on the Future of Digital Privacy. ZSR II. 2015;(134):426-
53 54	513		427.
55 56 57			
58			

59

Page 25 of 39

59

60

1 2			
3 4	514	6.	Sweeney L. Simple Demographics Often Identify People Uniquely. Data Privacy
5 6 7 8 9 10 11 12 13 14 15	515		Working Paper 3. Pittsburgh; 2000.
	516	7.	Patel CJ, Bhattacharya J, Ioannidis JPA, Bendavid E. Systematic identification of
	517		correlates of HIV infection. AIDS [Internet]. 2018 Apr 24;32(7):933-43. Available
	518		from: https://journals.lww.com/00002030-201804240-00013
	519	8.	Bendavid E, Claypool K, Chow E, Chung J, Mai D, Patel C. The Demographic,
16 17 18	520		Social, and Economic Correlates of HIV Infection Status in Sub-Saharan Africa.
19 20	521		Preprints. 2020;
21 22	522	9.	World Medical Association. World Medical Association Declaration of Helsinki:
 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 	523		ethical principles for medical research involving human subjects. JAMA [Internet].
	524		2013 Nov 27;310(20):2191–4. Available from:
	525		http://www.ncbi.nlm.nih.gov/pubmed/24141714
	526	10.	European Union. CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN
	527		UNION [Internet]. Official Journal of the European Union. 2012 [cited 2021 Jun
	528		11]. Available from: https://eur-lex.europa.eu/legal-
	529		content/EN/ALL/?uri=CELEX:12012P/TXT
	530	11.	National Commission for the Protection of Human Subjects of Biomedical and
42 43	531		Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for
44 45	532		the Protection of Human Subjects of Research. Washington, D.C.; 1978.
46 47 48	533	12.	Council for International Organizations of Medical Sciences, World Health
48 49 50	534		Organization. International Ethical Guidelines for Health-related Research
51 52	535		Involving Humans [Internet]. Council for International Organizations of Medical
53 54	536		Sciences (CIOMS). 2016. 1–119 p. Available from:
55 56 57			
58			

Page 26 of 39

1 2				
3 4	537		http://www.sciencedirect.com/science/article/B6VC6-45F5X02-	
5 6 7 8 9 10 11 12 13 14 15	538		9C/2/e44bc37a6e392634b1cf436105978f01	
	539	13.	O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting	
	540		Qualitative Research. Acad Med [Internet]. 2014 Sep;89(9):1245–51. Available	
	541		from: http://journals.lww.com/00001888-201409000-00021	
	542	14.	Staunton C, Moodley K. Challenges in biobank governance in Sub-Saharan	
16 17 18	543		Africa. BMC Med Ethics [Internet]. 2013 Dec 11;14(1):35. Available from:	
19 20	544		https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-14-35	
21 22	545	15.	Upshur RE, Lavery J V, Tindana PO. Taking tissue seriously means taking	
23 24	546		communities seriously. BMC Med Ethics [Internet]. 2007 Dec 26;8(1):11.	
25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40	547		Available from: https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-	
	548		6939-8-11	
	549	16.	Virani AH, Longstaff H. Ethical Considerations in Biobanks: How a Public Health	
	550		Ethics Perspective Sheds New Light on Old Controversies. J Genet Couns	
	551		[Internet]. 2015 Jun 29;24(3):428–32. Available from:	
	552		http://doi.wiley.com/10.1007/s10897-014-9781-9	
	553	17.	Enserink M, Chin G. The end of privacy. Science (80-) [Internet]. 2015 Jan	
41 42 43	554		30;347(6221):490–1. Available from:	
44 45	555		https://www.sciencemag.org/lookup/doi/10.1126/science.347.6221.490	
46 47	556	18.	Beck EJ, Gill W, De Lay PR. Protecting the confidentiality and security of personal	I
48 49 50	557		health information in low- and middle-income countries in the era of SDGs and	
50 51 52	558		Big Data. Glob Health Action [Internet]. 2016 Dec 1;9(1):32089. Available from:	
53 54	559		https://www.tandfonline.com/doi/full/10.3402/gha.v9.32089	
55 56				
57 58				
59 60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml 25	5

Page 27 of 39

1

59

60

countries.	BMJ Open: first published
ithical	d as 10. Pro
Views	1136/b ected
Empir Res	mjope by cop
ım Res	n-2021-052287 on 28 July 2021. Enseigner yright, including for uses relate
al health	Downloaded from http://br nent Superieur (ABES) . d to text and data mining,
ew of the	njopen Al trair
om:	.bmj.co ving, ar
)014-z	om/ on 1d simi
and the	June 1 lar tecl
et]. 2014	4, 2025 at A hnologies.
5-84	vgence
search.	Bibliograp
26	hique de

2			
3 4	560	19.	Barry M. Ethical Considerations of Human Investigation in Developing Countries.
5 6	561		N Engl J Med [Internet]. 1988 Oct 20;319(16):1083–6. Available from:
7 8	562		http://www.nejm.org/doi/abs/10.1056/NEJM198810203191609
9 10 11	563 20.		Denny SG, Silaigwana B, Wassenaar D, Bull S, Parker M. Developing Ethical
12 13	564		Practices for Public Health Research Data Sharing in South Africa: The Views
14 15	565		and Experiences From a Diverse Sample of Research Stakeholders. J Empir Res
16 17	566		Hum Res Ethics [Internet]. 2015 Jul 21;10(3):290–301. Available from:
10 19 20	567		http://journals.sagepub.com/doi/10.1177/1556264615592386
21 22	568	21.	Parker M, Bull S. Sharing Public Health Research Data. J Empir Res Hum Res
23 24 25	569		Ethics [Internet]. 2015 Jul 21;10(3):217–24. Available from:
25 26 27	570		http://journals.sagepub.com/doi/10.1177/1556264615593494
27 28 29 30 31 32 33 34 35 36	571	22.	Adhikari B, Pell C, Cheah PY. Community engagement and ethical global health
	572		research. Glob Bioeth [Internet]. 2020 Jan 1;31(1):1–12. Available from:
	573		https://www.tandfonline.com/doi/full/10.1080/11287462.2019.1703504
	574	23.	Tindana P, de Vries J, Campbell M, Littler K, Seeley J, Marshall P, et al.
37 38	575		Community engagement strategies for genomic studies in Africa: a review of the
39 40 41	576		literature. BMC Med Ethics [Internet]. 2015 Dec 12;16(1):24. Available from:
42 43	577		https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-015-0014-z
44 45	578	24.	King KF, Kolopack P, Merritt MW, Lavery J V. Community engagement and the
46 47 48	579		human infrastructure of global health research. BMC Med Ethics [Internet]. 2014
48 49 50 51 52	580		Dec 13;15(1):84. Available from:
	581		https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-15-84
53 54 55 56 57 58	582	25.	Dickert N, Sugarman J. Ethical Goals of Community Consultation in Research.

Page 28 of 39

BMJ Open

1 2			
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 4 35 36 37 38 9 40 41 42 43	583		Am J Public Health [Internet]. 2005 Jul;95(7):1123–7. Available from:
	584		http://ajph.aphapublications.org/doi/10.2105/AJPH.2004.058933
	585	26.	Marsh VM, Kamuya DK, Parker MJ, Molyneux CS. Working with Concepts: The
	586		Role of Community in International Collaborative Biomedical Research. Public
	587		Health Ethics [Internet]. 2011 Apr 1;4(1):26–39. Available from:
	588		https://academic.oup.com/phe/article-lookup/doi/10.1093/phe/phr007
	589	27.	Wilkinson A, Parker M, Martineau F, Leach M. Engaging 'communities':
	590		anthropological insights from the West African Ebola epidemic. Philos Trans R
	591		Soc B Biol Sci [Internet]. 2017 May 26;372(1721):20160305. Available from:
	592		https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0305
	593	28.	Nyirenda D, Sariola S, Kingori P, Squire B, Bandawe C, Parker M, et al. Structural
	594		coercion in the context of community engagement in global health research
	595		conducted in a low resource setting in Africa. BMC Med Ethics [Internet]. 2020
	596		Dec 21;21(1):90. Available from:
	597		https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-020-00530-1
	598	29.	Tindana P, Campbell M, Marshall P, Littler K, Vincent R, Seeley J, et al.
	599		Developing the science and methods of community engagement for genomic
	600		research and biobanking in Africa. Glob Heal Epidemiol Genomics [Internet]. 2017
44 45	601		Sep 4;2:e13. Available from:
46 47 49	602		https://www.cambridge.org/core/product/identifier/S2054420017000094/type/journ
48 49 50	603		al_article
51 52	604		
53 54	605		
55 56 57			
57 58 59			27
60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

	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
608			

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

1 2		
3 4 5		gender sexual relationships are illegal in many African countries, and that people who engage in them are actively persecuted in many?"
6 7		Lack of expertise in big data analysis
, 8 9 10		"Knowledge and understanding of what is big data - for ministries and for the populations."
11		Lack of agency of African researchers and ethicists
12 13 14 15		"There is lack of expertise in ethics review and monitoring research involving big data."
16 17 18		""The Americans (and their funders) should be in Africa, training Africans in big data methods and tools."
19 20		Compliance with or lack of country-specific laws and policies
20 21 22 23 24		"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."
25 26 27 28		"Information may have been deposited on an open source without permission, or in violation of the in-country laws."
29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55	011	
56 57 58 59 60		30 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

612	Table 3: Potential relevant stakeholders for engagement identified by expert panel
613	

015			
	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
	0	African scientists (clinical and public health scientists, biomedical researchers, and data scientists)	Traditional healers
		African healthcare workers	
		African ethicists	
614			

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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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

BMJ Open: first published as 10.1136/bmjopen-2021-052287 on 28 July 2021. Downloaded from Enseignement Superieur (AE

Protected by copyright, including for uses related to text and

data mining, Al training, and similar technologies

http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de

(ABES

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?

to beet terien only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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 149 (women) or 15 and 59 (men) who had consented to an HIV test; household dat (e.g. floor material, water source, electricity); family information (marital status, nu of children); health information (hemoglobin measurement, height and weight); far planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among of The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small su 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 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 tie HIV status. This would have the potential to improve targeting of public health progor help design interventions. For example, if widowhood is identified as a strong predictor of being HIV-positive, this can help design testing and prevention prograt that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "blows" machine learning approach that chooses the combination of features that bese predicts HIV status. The product of this analysis may not disclose any individual rist factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance	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 1£ 49 (women) or 15 and 59 (men) who had consented to an HIV test; household dat (e.g. floor material, water source, electricity); family information (marital status, nur of children); health information (hemoglobin measurement, height and weight); fam planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among ot The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 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 tie HIV status. This would have the potential to improve targeting of public health program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "blo box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive, this can be predictive in combination of features.	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 49 (women) or 15 and 59 (men) who had consented to an HIV test; household dat (e.g. floor material, water source, electricity); family information (marital status, nur of children); health information (hemoglobin measurement, height and weight); fan planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among ot The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 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 tie HIV status. This would have the potential to improve targeting of public health progor 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bbox" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual rise factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the ch	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 49 (women) or 15 and 59 (men) who had consented to an HIV test; household dat (e.g. floor material, water source, electricity); family information (marital status, nur of children); health information (hemoglobin measurement, height and weight); fan planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among ot The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 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 tie HIV status. This would have the potential to improve targeting of public health progor or help design interventions. For example, if widowhood is identified as a strong predictor of being HIV-positive, this can help design desting and prevention program. Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "bloox" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	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 49 (women) or 15 and 59 (men) who had consented to an HIV test; household dat (e.g. floor material, water source, electricity); family information (marital status, nur of children); health information (hemoglobin measurement, height and weight); fam planning information (use of contraception, sexual behavior patterns); and health behavior information (vaccination status, use of antenatal care services) among ot The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 a 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 tier HIV status. This would have the potential to improve targeting of public health progor 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bit box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict t	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 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, nur of children); health information (hemoglobin measurement, height and weight); fam planning information (vaccination status, use of antenatal care services) among otto the analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sut 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 r 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 HIV status. This would have the potential to improve targeting of public health prog 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "blab so" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	Here are exam	nples of data analys	ses that could b	e conducted with DHS	data:
The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small su 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 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 tie HIV status. This would have the potential to improve targeting of public health pro- 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 progra that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positiv given a combination of features.	The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 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 HIV status. This would have the potential to improve targeting of public health progor 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "blabox" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 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 tie HIV status. This would have the potential to improve targeting of public health prog 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 progran that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positiv given a combination of features.	The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 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 tier HIV status. This would have the potential to improve targeting of public health prog 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 progran that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positiv given a combination of features.	The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sul 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 i 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 tier HIV status. This would have the potential to improve targeting of public health prog 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 prograr that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	The analysis tests, statistically, which of these personal characteristics are most strongly associated with HIV status, and the precision of predictions from small sub 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 r 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 HIV status. This would have the potential to improve targeting of public health prog 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 prograr that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with other The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	These analyse results of HIV f 49 (women) or (e.g. floor mate of children); he planning inform behavior inform	es would use data co tests from about 1,00 15 and 59 (men) wh erial, water source, e ealth information (her nation (use of contra nation (vaccination s	llected in 30 Afr 00,000 men and no had consente lectricity); family moglobin measu ception, sexual tatus, use of an	ican countries, and inc women between the a ed to an HIV test; house / information (marital s urement, height and we behavior patterns); and tenatal care services) a	lude: the iges of 15 ehold data tatus, nun ight); fam d health among otl
One type of analysis would identify the individual features that are most closely tie HIV status. This would have the potential to improve targeting of public health pro- 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 progra that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	One type of analysis would identify the individual features that are most closely tied HIV status. This would have the potential to improve targeting of public health progor 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	One type of analysis would identify the individual features that are most closely tied HIV status. This would have the potential to improve targeting of public health proportion 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positiv given a combination of features.	One type of analysis would identify the individual features that are most closely tied HIV status. This would have the potential to improve targeting of public health progor 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	One type of analysis would identify the individual features that are most closely tied HIV status. This would have the potential to improve targeting of public health progor 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	One type of analysis would identify the individual features that are most closely tied HIV status. This would have the potential to improve targeting of public health prog 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 program that are tailored to widows. This is similar to the identification of male circumcision 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 many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	The analysis te strongly assoc of characteristi epidemiologica ownership is, i prediction mod	ests, statistically, whi iated with HIV status ics. The predictors m al research, but may n some surveys, a s lel can improve pred	ich of these pers s, and the precis nay or may not h be strongly prec trong predictor of iction accuracy	sonal characteristics ar ion of predictions from have been identified by dictive. For example, bi of HIV status, and addin of HIV status from 82%	e most small sub previous cycle ng it to a r to 85%.
Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positiv given a combination of features.	Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "bl box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "blbox" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with othe The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	Another type of analysis would create risk scores that are a weighted combination many individual features. This risk score would emerge from a commonly-used "bla box" machine learning approach that chooses the combination of features that bes predicts HIV status. The product of this analysis may not disclose any individual ris factors, and indeed some factors might only be predictive in combination with other The analysis could report how well models predict the chance of being HIV-positive given a combination of features.	One type of an HIV status. Th or help design predictor of be that are tailore risk factor that	alysis would identify is would have the po interventions. For ex ing HIV-positive, this d to widows. This is led to clinical trials a	the individual fe tential to improve cample, if widows can help desig similar to the ide and large-scale p	eatures that are most c ve targeting of public he vhood is identified as a n testing and preventio entification of male circ public health programs	losely tied ealth prog strong n prograr umcision
						Another type o many individua box" machine I predicts HIV st factors, and ine The analysis c given a combin	f analysis would creat al features. This risk learning approach th atus. The product of deed some factors mould report how well nation of features.	ate risk scores t score would em at chooses the this analysis m night only be pre models predict	hat are a weighted con erge from a commonly combination of features ay not disclose any ind edictive in combination the chance of being HI	nbination -used "bla s that bes ividual ris with othe V-positive
						•			4	

Standards for Reporting Qualitative Research (SRQR)*

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

Page no(s).

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 partaining to human subjects. Desumentation of approval by an	
appropriate othics review beard and participant concent, or evplanation for lack	
appropriate ethics review board and participant consent, or explanation for lack	11
thereof; other confidentiality and data security issues	
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
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
the	er	

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.

BMJ Open

**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:

L A. Cock DA. S. instances and a set of the set of t 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.00000000000388