BMJ Open

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

Ariadne A Nichol ,[1] Eran Bendavid,[2] Farirai Mutenherwa,[3,4] Chirag Patel,[5] Mildred K Cho[1]

To cite: Nichol AA, Bendavid E, Mutenherwa F, *et al*. Diverse experts' perspectives on ethical issues of using machine learning to predict HIV/AIDS risk in sub-Saharan Africa: a modified Delphi study. *BMJ Open* 2021;**11**:e052287. doi:10.1136/bmjopen-2021-052287

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-052287).

Received 11 April 2021
Accepted 19 July 2021

Check for updates

For numbered affiliations see end of article.

Correspondence to
Ariadne A Nichol;
ariadnen@stanford.edu

ABSTRACT

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

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

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

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

Results Of the 35 invited experts, 22 participated in the online expert panel (63%). Emergent themes were the inclusion of African researchers in all aspects of study design, analysis and dissemination to identify and address local contextual issues, as well as engagement of communities. Experts focused on engagement with health and science professionals to address risks, benefits and communication of findings. Respondents prioritised the mitigation of stigma to research participants but recognised trade-offs between privacy and the need to disseminate findings to realise public health benefits. Strategies for responsible communication of results were suggested, including careful word choice in presentation of results and limited dissemination to need-to-know stakeholders such as public health planners.

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

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 trade-offs, such as alternative methods of describing and disseminating findings of predictive analytics to minimise risks to privacy and of stigmatisation, and suggestions for prioritising specific groups for community engagement.
- The main limitation of this study is that a small number of respondents completed all three surveys. However, our expert respondents did represent diverse perspectives in informatics, bioethics of Africa-based studies and African public health and HIV/AIDS.

INTRODUCTION

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

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

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

METHODS

Approach

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

Sample

Our multidisciplinary 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 USA, but did work on health in Africa. All panellists 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 three 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 3 weeks to respond, with two reminder emails to all 22 who initially agreed to participate. The initial survey was designed to capture a wide range of ethical issues, including those that might not have been already identified in the literature using broad open-ended questions, as well as to assess the perceived importance of previously raised concerns. Responses were then analysed to identify areas that were most frequently identified as important but where there was also disagreement about what to do. Subsequent survey questions were developed to identify how experts would prioritise values or make trade-offs between conflicting values to address ethical issues.

Survey 1

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

Box 1 Survey 1 scenario

- ► A group of American scientists funded by the US government is developing big data tools to identify individuals and groups at elevated risk of acquiring HIV in sub-Saharan Africa. The purpose of the project is to help ministries of health and international public health organisations target testing and treatment programmes 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 global positioning system (GPS) coordinates of the households' village or neighbourhood, among others, are characterised 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.

Table 1 Importance of ethical issues (Survey 1)

Item	Mean (1=not important at all, 5=absolutely essential) n=16	SD
Potential to stigmatise 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

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

Survey 2

Responses to Survey 1 indicated that, in order to comment on ethical issues and to make recommendations, respondents needed more detail on how informed consent and ethical review processes were conducted for data collection for the DHS and for data use by individual researchers. As a result, we significantly expanded the description of the study for Survey 2 to include details on what data were collected and how data privacy, access and ethical review of the DHS were handled (online supplemental figure 1). This survey's questions were open-ended, reflecting areas of consensus on importance that had emerged in the previous survey: (1) stakeholder engagement; (2) privacy/stigmatisation/discrimination; (3) ethics review; (4) data access and (5) dissemination and communication of study findings (online supplemental figure 2).

Survey 3

The findings from Survey 2 were used to design Survey 3 to probe areas of disagreement, and to elicit details that could inform draft recommendations about stakeholder engagement and ethics review. In Survey 3, we presented the same scenario as in Survey 2 (online supplemental figure 1), 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 trade-offs between dissemination and privacy (online supplemental figure 3). 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 stigmatisation and discrimination (rating level of agreement with statements on balancing privacy, stigmatisation and discrimination concerns with the dissemination of useful findings of risk factors for HIV/AIDS) and (3) requirements for ethical review (rating level of agreement with statements on the type of ethics review that is sufficient), with a closed-ended component and opportunity for open-ended explanation.

Qualitative data analysis

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

Patient and public involvement

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

RESULTS

Survey 1

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

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

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

The specific use of big data predictive analytics generated several ethical issues that respondents wanted to ensure were properly addressed prior to any research, including assessment of the potential for bias, independent review of the validity of the predictive analytics tools and establishment of a plan for monitoring interventions for harm that could result based on which individuals or groups were identified as being high risk for HIV/AIDS. Several respondents also emphasised the need for researchers to think through how the big data predictive analytics outcomes can be used to inform testing and treatment programmes 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 box 2 for exemplar quotes) included a history of human rights abuses, lack of trust in government, misuse of research findings, HIV-associated characteristics (eg, 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 box 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 on a scale of 1–5), and did not rate any of the seven issues as *Not Important At All* or *Of Little Importance*. Nevertheless, two items clearly emerged as being most important. First was the potential to stigmatise groups or populations that are uniquely identified by the research (all rated this issue as *Absolutely Essential*) and, second, the privacy of individuals (14 rated this as *Absolutely Essential*, 2 as *Very Important*). The next two most important issues identified were the validity of findings using big data tools and potential for bias.

Survey 2

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

Overall, community and stakeholder engagement that includes Africans, ideally in relevant countries, were seen as key to minimising risks at several stages of the research process, including data access, protocol oversight and dissemination and implementation of findings. Some recommended engagement at the regional as well as national level, and respondents named a wide range of stakeholder groups (see table 2). There was also broad support for community engagement in general to protect interests of local communities, groups and individuals. This engagement would provide the opportunity to better understand local concerns, values, norms and cultural considerations and guide researchers on how to communicate findings in a way that mitigate risks to communities and individuals. Other purposes of stakeholder and community engagement were to provide education to public health officials and policymakers, clinicians and communities, enhance buy-in, identify opportunities for capacity building and translation and ultimately build trust and collaboration.

While Survey 1 indicated consensus on privacy as a primary concern, in Survey 2, statements about how researchers could address this issue were mixed. Some acknowledged limits on researchers' ability to prevent misuse of findings or to completely protect data privacy; however, others also proposed specific actions to minimise 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

Box 2 Contextual factors—exemplar quotes from Survey 1 respondents

Need for engaging African scientists

- ‘For instance, a South African HIV researcher would be knowledgeable about existing stigma relating to this condition and to any attributes of the population groups that could be identified through this research. He or she would likely be in a better position than someone who has never been here to assess whether and when particular kinds of scientific results would be likely to fuel existing stigma or discrimination. He or she would also have ongoing access to these communities and would likely have some insight into how such groups should be referred to in publications emanating from this research.’
- ‘The exclusion of African researchers from research about Africans, in my view, means that we do not maximise the opportunity to be effective.’
- ‘The Americans (and their funders) should be in Africa, training Africans in big data methods and tools.’

Data on Africans being used by non-African researchers

- ‘Countries in sub-Saharan Africa (SSA) are concerned with information being used by researchers abroad, and do not appreciate information being stored in servers outside of their countries, or extracted for analysis in abroad.’

History of human rights abuses

- ‘How the researchers protect the privacy of these individuals would be critical considering the gross human rights abuses and poor legal frameworks in certain jurisdictions across Africa.’

Lack of trust in government and potential for misuse of research findings

- ‘The most important ethical consideration would be to ensure that the privacy of the individuals in the dataset is not compromised, and government officials have no way of tracing back individuals in the dataset up to the household level.’
- ‘Trust—Entrusting Ministries/governments could misuse the information—how can this be safeguarded. Information and political use—interventions may be denied where political support is low in some regions. Development of tools which could be abused by authorities or for political reasons.’

HIV-associated characteristics (eg, homosexuality) that are crimes in some African countries

- ‘Since HIV infection is associated with homosexual behavior which is criminal in many SSA countries, individuals identified in the study may also be in legal jeopardy.’
- ‘How will these researchers ensure that their results will be used for good and not for harmful or discriminatory purposes, especially considering that for example, same-gender sexual relationships are illegal in many African countries, and that people who engage in them are actively persecuted in many?’

Lack of expertise in big data analysis

- ‘Knowledge and understanding of what is big data—for ministries and for the populations.’

Lack of agency of African researchers and ethicists

- ‘There is lack of expertise in ethics review and monitoring research involving big data.’
- ‘The Americans (and their funders) should be in Africa, training Africans in big data methods and tools.’

Compliance with or lack of country-specific laws and policies

- ‘Consider laws in each region/country as these may differ significantly, or simply not exist in a functional format. Important to understand what local laws are available and what is constitutionally acceptable.’
- ‘…Information may have been deposited on an open source without permission or in violation of the in-country laws.’

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 centralised versus local ethics review and whether research on publicly available or de-identified data was considered exempt from ethics review. Some respondents felt the centralised and local ethics review of the DHS 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 institutional review board (IRB) (ideally based in SSA)’. There was also disagreement about the adequacy of existing data access control and protection against stigma and discrimination from study findings. While one respondent suggested that data access controls were sufficient because data were de-identified, another would require ‘a clear data analysis and dissemination plan’, and another stated that protocol-specific data sharing agreements were necessary, because ‘Africa has suffered most from exploitation; both for research subjects and researchers’

Survey 3

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

Table 2 Potential relevant stakeholders for engagement identified by expert panel

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

identify and engage with local communities with data coming from over one million people.

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

When asked about the balance between the benefits of disseminating the research findings and risks of identification and stigma, there was support for some limitations on reporting to protect the identity of individuals or small groups (ie, reporting overall performance of predictive models rather than individual risk factors) but less agreement about restricting reporting of findings that could identify small numbers of individuals if the findings would be less useful for public health officials. There was broad agreement on the need for community representatives to have input on how risk factors are described in publications (eg, if local geographic regions were to be mentioned in publications, community representatives would know whether this could lead to stigmatisation against those relevant subpopulations), 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 stigmatisation 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 centralised 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 organisations.

DISCUSSION

It is increasingly recognised 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 analytical 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 questions highlighted ethical concerns that would apply to much of biomedical research generally, but with a focus on contextual factors. These factors included a history of human rights abuses, lack of trust in government and in non-African researchers, misuse of research findings and obligations of US researchers to help build research capacity in Africa.[19]

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

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

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

Community engagement is seen as a critical part of health research in general, especially in SSA, given the history of exploitation.[12 22 24 25] Yet, there is extensive variation in defining what constitutes a 'community'.[22 23 26 27] Our findings indicate recognition of the need for engagement at national, regional and local levels with a wide array of proposed participants. Interestingly, our panel of experts found other stakeholders (ie, 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 minimise risks of stigmatisation and discrimination. It is important to consider the nature of predictive analytics and big data research as transnational and inclusive of many individuals' data. Therefore, this focus on broader stakeholders' engagement is explicable and perhaps a somewhat unique feature to research involving big data. Of course, these stakeholders have been identified as proposed participants of engagement for health research before, including within the context of genomics and biobanking in Africa.[14 28 29] Some have also suggested these stakeholders are in fact 'community' in a broader interpretation.[22]

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

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

CONCLUSION

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

Author affiliations
[1]Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, California, USA
[2]Department of Primary Care and Population Health, Stanford University School of Medicine, Stanford, California, USA
[3]College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa
[4]School of Applied Human Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[5]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

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

Funding This work was funded by the US National Institutes of Health (1 R01 AI127250-01, Big Data Analysis of HIV Risk and Epidemiology in sub-Saharan Africa). EB and CP are principal investigators and MKC is a coinvestigator of the grant. The views expressed are solely those of the authors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethics approval for exemption was provided by the Stanford Human Subjects Panel (protocol number IRB-50813) and the reason for exemption was exemption #2 according to the US Common Rule: Research involving the use of educational tests, survey procedures, interview procedures or observation of public

behaviour and information obtained is recorded by the investigator in such manner that the identity of the human subjects cannot readily be ascertained.

Provenance and peer review Not commissioned; externally peer reviewed.

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

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD
Ariadne A Nichol http://orcid.org/0000-0001-9174-9671

REFERENCES

1 Vayena E, Madoff L. Navigating the Ethics of Big Data in Public Health. In: Mastroianni AC, Kahn JP, Kass NE, eds. *The Oxford Handbook of public health ethics*. Oxford University Press, 2019: 353–67. http://oxfordhandbooks.com/view/
2 Vayena E, Salathé M, Madoff LC, *et al*. Ethical challenges of big data in public health. *PLoS Comput Biol* 2015;11:e1003904.
3 Goodman KW, Meslin EM. Ethics, Information Technology, and Public Health: Duties and Challenges in Computational Epidemiology. In: Magnuson JA, Fu P, eds. *Public health informatics and information systems*. 2nd ed. New York: Springer, 2014: p. 191–209.
4 Rothstein MA. Is deidentification sufficient to protect health privacy in research? *Am J Bioeth* 2010;10:3–11. doi:10.1080/15265161.2010.494215
5 Gasser U. Perspectives on the future of digital privacy. *Zsr Ii* 2015;134:426–7.
6 Sweeney L. Simple demographics often identify people uniquely. data Privacy515 working paper 3. Pittsburgh; 2000.
7 Patel CJ, Bhattacharya J, Ioannidis JPA, *et al*. Systematic identification of correlates of HIV infection. *AIDS* 2018;32:933–43 https://journals.lww.com/00002030-201804240-00013
8 Bendavid E, Claypool K, Chow E. The demographic, social, and economic correlates of HIV infection status in sub-Saharan Africa. *Preprints* 2020.
9 World Medical Association. World Medical association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013;310:2191–4.
10 European Union. Charter of fundamental rights of the European Union. *Official Journal of the European Union* 2012 https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:12012P/TXT
11 National Commission for the protection of human subjects of biomedical and behavioral research. The Belmont report: ethical principles and guidelines for the protection of human subjects of research. Washington, DC; 1978.
12 Council for International Organizations of Medical Sciences, World Health Organization. International ethical guidelines for health-related research involving humans; 2016: 1–119. http://www.sciencedirect.com/science/article/B6VC6-45F5X02-9C/2/e44bc37a6e392634b1cf436105978f01
13 O'Brien BC, Harris IB, Beckman TJ. Standards for reporting qualitative research. *Academic Medicine* 2014;89:1250–1.
14 Staunton C, Moodley K. Challenges in Biobank governance in sub-Saharan Africa. *BMC Med Ethics* 2013;14:35. doi:10.1186/1472-6939-14-35
15 Upshur REG, Lavery JV, Tindana PO. Taking tissue seriously means taking communities seriously. *BMC Med Ethics* 2007;8:11.
16 Virani AH, Longstaff H. Ethical considerations in biobanks: how a public health ethics perspective sheds new light on old controversies. *J Genet Couns* 2015;24:428–32. doi:10.1007/s10897-014-9781-9
17 Enserink M, Chin G. The end of privacy. *Science* 2015;347:490–1. doi:10.1126/science.347.6221.490
18 Beck EJ, Gill W, De Lay PR. Protecting the confidentiality and security of personal health information in low- and middle-income countries in the era of SDGs and big data. *Glob Health Action* 2016;9:32089. doi:10.3402/gha.v9.32089
19 Barry M. Ethical considerations of human investigation in developing countries: the AIDS dilemma. *N Engl J Med* 1988;319:1083–6. doi:10.1056/NEJM198810203191609
20 Denny SG, Silaigwana B, Wassenaar D, *et al*. Developing ethical practices for public health research data sharing in South Africa: the views and experiences from a diverse sample of research stakeholders. *J Empir Res Hum Res Ethics* 2015;10:290-301–301. doi:10.1177/1556264615592386
21 Parker M, Bull S. Sharing public health research data. *Journal of Empirical Research on Human Research Ethics* 2015;10:217–24. doi:10.1177/1556264615593494
22 Adhikari B, Pell C, Cheah PY. Community engagement and ethical global health research. *Glob Bioeth* 2020;31:1–12. doi:10.1080/11287462.2019.1703504
23 Tindana P, de Vries J, Campbell M, *et al*. Community engagement strategies for genomic studies in Africa: a review of the literature. *BMC Med Ethics* 2015;16:24. doi:10.1186/s12910-015-0014-z
24 King KF, Kolopack P, Merritt MW, *et al*. Community engagement and the human infrastructure of global health research. *BMC Med Ethics* 2014;15:84.
25 Dickert N, Sugarman J. Ethical goals of community consultation in research. *Am J Public Health* 2005;95:1123–7. doi:10.2105/AJPH.2004.058933
26 Marsh VM, Kamuya DK, Parker MJ, *et al*. Working with concepts: the role of community in international collaborative biomedical research. *Public Health Ethics* 2011;4:26–39. doi:10.1093/phe/phr007
27 Wilkinson A, Parker M, Martineau F, *et al*. Engaging 'communities': anthropological insights from the West African Ebola epidemic. *Phil. Trans. R. Soc. B* 2017;372:20160305. doi:10.1098/rstb.2016.0305
28 Nyirenda D, Sariola S, Kingori P, *et al*. Structural coercion in the context of community engagement in global health research conducted in a low resource setting in Africa. *BMC Med Ethics* 2020;21:90.
29 Tindana P, Campbell M, Marshall P, *et al*. Developing the science and methods of community engagement for genomic research and biobanking in Africa. *Glob Health Epidemiol Genom* 2017;2:e13.