SUPPLEMENTAL INFORMATION

Figure 1: Survey 2 scenario

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

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

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

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

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

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

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

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

Figure 2: Survey 2 questions

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

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

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

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

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

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

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

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

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

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

Figure 3: Survey 3 examples added to research scenario

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

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

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

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

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