JAIDS Journal of Acquired Immune Deficiency Syndromes Publish Ahead of Print
DOI: 10.1097/QAI.0000000000003035

Overcoming Challenges of Online Research: Measures to Ensure Enrollment of Eligible Participants

Chadwick K. Campbell, PhD, MPH[1^,*], Samuel Ndukwe, BS[2,^], Karine Dubé, DrPH[2], John A. Sauceda, PhD, MSc[1], Parya Saberi, PharmD, MAS, MFA, AAHIVP[1]

[1]Center for AIDS Prevention Research, Department of Medicine; University of California San Francisco, San Francisco, CA; USA

[2]UNC Gillings School of Global Public Health; University of North Carolina at Chapel Hill; Chapel Hill, NC; USA

[^]Chadwick K. Campbell and Samuel Ndukwe are joint co-first authors.

Corresponding author: Chadwick K. Campbell PhD, MPH 550 16th Street, 3rd Floor San Francisco, California 94143 415-502-1000 ext. 65348 (vm only) (Chadwick.Campbell@ucsf.edu)

The authors report no conflicts of interest related to this work.

Funding: The authors' work was supported by the National Institute of Mental Health (award number R21MH122280).

Running Head: Ensuring eligibility in online research

ABSTRACT

Background: Internet-based surveys are increasingly used for health research as they offer several advantages including greater geographic reach, increased participant anonymity, and reduced financial/time burden. Though, there is also a need to address inherent challenges, such as the likelihood of fraudulent responses and greater difficulty in determining eligibility.

Methods: We conducted an online nationwide survey of 18–29-year-olds living with HIV in the United States, to assess willingness to participate in HIV cure research. To ensure that respondents met age and HIV serostatus inclusion criteria, we instituted screening procedures to identify ineligible respondents utilizing tools that were built into the survey platform (e.g., ReCAPTCHA, geolocation), and required documentation of age and serostatus before providing access to the incentivized study survey.

Results: Of 1,308 eligibility surveys, 569 were incomplete or ineligible due to reported age or serostatus. Of the remaining 739 potentially eligible respondents, we determined that 413 were from fraudulent, bot, or ineligible respondents. We sent individual study survey links to 326 (25% of all eligibility survey respondents) participants whose eligibility was reviewed and confirmed by our study team.

Conclusion: Our multi-component strategy was effective for identifying ineligible and fraudulent responses to our eligibility survey, allowing us to send the study survey link only to those whose eligibility we were able to confirm. Our findings suggest that proactive fraud prevention can be built into the screening phase of the study to prevent wasted resources related to data cleaning and unretrievable study incentives, and ultimately improve the quality of data.

Key words: Internet Research; Eligibility; HIV Research; Online Recruitment

INTRODUCTION

With the increased consumption of social networks such as Facebook, Twitter, and Instagram, Internet and social networking platforms are being used by researchers for study recruitment and data collection in HIV and other health-related research.[1-5] These platforms are beneficial due to the cost-effectiveness of online advertisements and the volume of potential participants/users who may interact with the advertisement.[2,4,6-8] Further, internet-based research expands the geographic reach of recruitment efforts, reduces participation burdens (e.g., financial, travel, time commitment), and improves ability to reach and provide increased anonymity to participants that have been marginalized (e.g., people of color or LGBTQ+ people) and/or those experiencing stigma (e.g., due to gender identity or HIV serostatus).[1,2,6,9,10]

Despite its benefits, online recruitment is a complex and indirect form of communication involving third parties that collect, aggregate, and store participant data.[2] Other challenges of online research include recruiting ineligible individuals (fraud), “bots” (computer software designed to complete automated tasks), and people entering duplicate responses.[2,5,6,10,11] Therefore, strategies are needed to ensure that data are collected from individuals who are eligible for research participation.

There are a number of available automated data safety mechanisms (e.g., Completely Automated Public Turing Test to tell Computers and Humans Apart (reCAPTCHA), bot detection, anti-ballot stuffing) in survey software but studies show that human participants and sophisticated bots can bypass them.[5,8,12] Other mechanisms from prior survey studies have included questions requiring a meaningful open-ended text responses. In one study, by reviewing responses to questions like "How has coronavirus (COVID-19) affected your life?", the authors identified a considerable number of bots (13.3%) with exact duplicate text responses.[10] Another study recommended open-ended responses with "trap questions" that could not be true for anyone (e.g., "I was born in the 18th century).[12]

Other strategies involve rigorously examining survey data for duplicate, inconsistent, or suspicious responses. For example, geolocation data can identify respondents outside of the recruitment area, names and email addresses can be reviewed to identify duplicates or unusual formats, and phone numbers can be reviewed to identify inactive numbers or duplicates.[4-6,10] As part of a national online survey of young adults living with HIV in the United States, we describe a series of measures we used to ensure that only truly eligible human respondents were enrolled in our study.

METHODS

Study Design

We conducted a cross-sectional online survey with young adults (18-29 years old) living with HIV (YLWH) in the United States to examine willingness, motivators, and deterrents of participating in HIV cure research.[13,14] Consent was received and all procedures were approved by the University of California, San Francisco Institutional Review Board (IRB).

Between April 2021 and August 2021, we recruited participants using social media, ads on mobile dating apps, and through organizations and clinics that serve YLWH. Recruitment ads were included clear language about eligibility criteria ("Are you 18-29 and living with HIV?"), the study purpose ("We are looking for people like you to participate in a survey about HIV cure research"), and incentive ("You can receive $40 for participating."). Eligible participants were: 18-29 years old, living in the US, living with HIV, able to complete the study survey in English, and willing to give consent. Interested persons completed an online eligibility survey. Prior to this eligibility survey, they were presented with consent information and were instructed, "by completing this survey, you are consenting to allowing us to use the information in this brief screening survey for our research." Age was verified with an image of their ID showing their name and date of birth, and HIV status was verified with a photo of their antiretroviral medication vial, a letter of diagnosis from their provider, or lab report showing their HIV status. Eligible participants were emailed or texted informed consent information and a unique Health Insurance Portability and Accountability Act (HIPAA)-compliant Qualtrics survey link. Upon completion of the survey, participants were paid US$40 via a cash transfer app or e-gift card.

Measures to Ensure Enrollment of Eligible Participants

The online screening survey included several layers of measures to ensure the accurate recruitment of YLWH. Initially, a ReCAPTCHA verification asked users to identify items in a series of photos. Next, they were given the option to upload HIV and age verification documents directly into the survey, or to have the study team follow-up with them to receive the documents via encrypted text message, on a study mobile phone only accessible by the Study Coordinator, or through university-issued, HIPAA compliant and encrypted email. The survey question that requested an upload of documents was a barrier to bots getting to the end of the screening survey because if they responded 'yes' to being able to upload images but were unable to do so, they would not be able to complete this screening survey. While we considered activating the protection against "ballot stuffing," which uses cookies to identify users who have already completed the survey, we opted not to so as not to block legitimate responses from two people who might live together or use the same computer.

Further, a number of individuals who completed the survey uploaded images of people, things, or fraudulent identifications. For example, age verification documents included digitally manipulated identification cards using photos of celebrities or politicians (e.g., President Barrack Obama). Another example included a diagnosis letter to confirm HIV status that did not include the respondent's name or was on the letterhead of a clinic located in Montreal yet signed by a physician practicing in Florida, while the screening survey indicated that the respondent lived in California. The Study Coordinator (SC) conducted a detailed review of each screening survey to identify valid and complete surveys and those that contained incomplete information or fraudulent documents prior to sending the incentivized survey link. When survey responses did not include required verification documents, the SC attempted to contact the respondent to obtain legitimate documents. Few respondents who did not upload verification documents chose to provide them after being contacted, and there were no instances in which respondents who uploaded fraudulent documents replied and provided legitimate documents.

Bots were also detected by screening for similar combinations of names, email addresses, phone numbers, and Internet Protocol (IP) addresses. Some indicators were first and last names entered in all lower-case letters with no spaces followed by what appeared to be random, additional letters or numbers (e.g., samanthadoekt), combined with email addresses that followed similar, repeated patterns (e.g., samathadoekt9756@gmail.com). Other indicators included responses using the same or similar names but with different email addresses and/or phone numbers. Further, in cases where the SC attempted to contact these respondents, phone numbers were not active. Each completed online eligibility screening survey was reviewed by the SC to confirm eligibility. The SC sent participants their own unique web link to complete the incentivized study survey only after receiving legitimate age and HIV status verification documents and reviewing the screening survey responses. These unique links were the final step of ensuring integrity of study enrollment as they could not be shared with others or accessed by bots.

Time requirements for these detailed reviews varied throughout the recruitment period, peaking immediately after placements of new ads. The SC's effort on this study was 0.5 FTE and, during peak periods, reviewing screening surveys, contacting participants to retrieve verification documents, and sending consent forms and unique study links to eligible participants required the SC's full effort. For example, when we placed a one-week advertisement on a gay dating app, we received 202 screening surveys within 7 days. During slower periods, the SC allocated more time to reminding participants who had already received a study link to complete the survey and following up with those who had not provided age and HIV status verification.

RESULTS

The results of our eligibility verification measures are detailed in **Figure 1**. We received a total of 1308 eligibility surveys, of which four (.3%) were incomplete, 50 (3.8%) indicated they were no longer interested before providing their contact information, and 48 (3.6%) were determined to be duplicates (i.e., surveys completed by the same person more than once). Duplicates occurred if a participant completed the survey twice in a short period of time or had already participated in the study once yet completed the screening survey again at a later date. Finally, 470 (35.9%) respondents were ineligible because they were under 18 or over 29 (n=142), or reported their HIV status as unknown or negative (n=240). Others opted not to provide documentation of their HIV status or age (n=88). Importantly, these are distinct from another group of responses that did not include verification documents as we discuss below. These were determined to be real human respondents (as opposed to bot responses) based on having responded directly to ads, and because they provided working phone numbers and email address. In three cases, the SC was able to reach the respondents via email or phone. These three participants indicated that they were not comfortable providing the requested documentation, while others did not respond to our requests.

Of the remaining 739 potentially eligible responses, 413 (55.8%) were determined to be fraudulent or bot responses. These included 45 responses with uploaded images that did not confirm age (e.g., a selfie; n=5), HIV status (e.g., fake diagnosis letter, photo of non-antiretroviral medication; n=24), or both (n=16), as well as those who, according to geolocation data, were outside of the US (e.g., Zambia, Vietnam, Mexico; n=28). Because we opted not to use the "prevent ballot stuffing" Qualtrics feature, so as not to exclude partners or roommates using the same computer, it was possible for multiple surveys to be completed using the same IP address. Upon review by the SC, we identified 91 surveys that we determined to be "repeated bot responses," a category that included some combination of IP address, name, email, or phone number repeated in multiple surveys over a short period of time (typically within a few minutes), and if none of those responses included complete screening information and verification documents. For example, on May 3, 2021, we received six responses from the same IP address within 33 seconds of each other. These responses all reported detectable HIV viral loads, contained invalid phone numbers, and all responded "no" when asked to upload age and status verification documents. These 91 "repeated bot responses" included only 37 IP addresses, 33 unique names, and 14 names that were repeated between 2–8 times within a few seconds or minutes of each other.

A final 246 surveys did not include age or HIV status verification documents which was due to, presumably, automated responses that were interrupted due to lack of document upload. In addition to not uploading verification documents, these survey responses also had inactive phone numbers, unusual names or email addresses, or a combination of these factors. A total of 326 (24.9% of all received eligibility surveys) unique survey links were sent to eligible participants, and 271 surveys were included in the final sample, giving us a survey completion rate of 83%.

DISCUSSION

We describe a new combination approach to reduce the threat of fraudulent participants and bot responses in online surveys. In anticipation of fraudulent responses and bots, we created a screening survey with various security measures in place. In addition to using built-in Qualtrics features (e.g., reCAPTCHA), we implemented a document-upload step and human review processes to evaluate screening data prior to sending the incentivized survey link.

Importantly, two of our most important recruitment integrity strategies were built into our study design by including individuals who were 18–29 year of age and living with HIV. We required potential participants to upload a photo ID showing their name, birthdate, and documentation of their HIV serostatus. Because of this document upload, bots were prevented from completing the survey and ineligible human respondents who uploaded fraudulent documents (e.g., photoshopped ID, faked letters of diagnosis) or did not respond when contacted, were excluded. Finally, study procedures included sending separate and unique Qualtrics survey links to each eligible participant which could not be shared with others or accessed by bots.

A number of verification strategies have been used in online-based studies such as having participants enter their names multiple times in the survey (e.g., for e-gift card receipt, emails, consent to be contacted for future research) which can be cross-referenced by the research team.[6] Similar to other research,[9] we reviewed names to identify possible duplicate responses and were able to identify screening surveys with repeat names, emails, and IP addresses (mostly completed within a short period of time). We also identified unusual names or name formats, which were often paired with an unusual and repeated pattern of email addresses (e.g., names followed by a series of random numbers or letters), which is an indication of bot email generation.[6,10,15-17] Lastly, geolocation data showed that the respondents were located outside of the United States. While built-in fraud detection settings are useful in detecting bots, they are grossly insufficient. Therefore, the review of names, email addresses, and phone numbers by the study team can be an effective check to ensure participant eligibility prior to enrollment in the incentivized study.

There are challenges to our approach worth noting. Requiring participants to upload ID and HIV status may lead to under sampling of some important groups (people who use drugs, incarcerated or undocumented) who may not have identification or may not be comfortable uploading these documents. Inconsistent access to broadband and slow internet speeds also present challenges. Lastly, human review of data is time-intensive and tedious, and there is inherent subjectivity when discerning the validity of self-reported data. Thus, we may have overlooked a fraudulent entry or denied a valid participant when evaluating screening survey data. Despite these potential limitations, our screening strategies were effective in identifying a large number of bot and fraudulent human responses. The use of technologies that may be useful in this process. For example, we may be able to use artificial intelligence in place of human review of images and documents. This can reduce the need for research coordinator time to review each image and it may be more precise in detecting fraud (e.g., detecting photoshopped images).

Innovative and multi-component fraud protection protocols are necessary to avoid threats to data security and improve data quality, but does require greater burdens to persons to pass these protections. Our approach and findings suggest that proactive fraud prevention can be built into the screening phase of the study to identify potentially fraudulent responses, prevent wasted resources related to data cleaning and unretrievable study incentives, and ultimately improve data quality. We encourage researchers to implement and share additional strategies in implementing valid online surveys and for journal editors and the peer review process to further scrutinize papers using online research methodology.

REFERENCES

1. Arigo D, Pagoto S, Carter-Harris L, Lillie SE, Nebeker C. Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery. *Digital health*. 2018;4:2055207618771757.

2. Curtis BL. Social networking and online recruiting for HIV research: ethical challenges. *Journal of Empirical Research on Human Research Ethics*. 2014;9(1):58-70.

3. Dubé K, Campbell DM, Perry KE, et al. Reasons people living with HIV might prefer oral daily antiretroviral therapy, long-acting formulations, or future HIV remission options. *AIDS Res Hum Retroviruses*. 2020;36(12):1054-1058.

4. Sterzing PR, Gartner RE, McGeough BL. Conducting anonymous, incentivized, online surveys with sexual and gender minority adolescents: Lessons learned from a national polyvictimization study. *J Interpers Violence*. 2018;33(5):740-761.

5. Teitcher JE, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to “fraudsters” in internet research: ethics and tradeoffs. *J Law Med Ethics*. 2015;43(1):116-133.

6. Ballard AM, Cardwell T, Young AM. Fraud detection protocol for web-based research among men who have sex with men: development and descriptive evaluation. *JMIR public health and surveillance*. 2019;5(1):e12344.

7. Bragard E, Fisher CB, Curtis BL. “They know what they are getting into:” Researchers confront the benefits and challenges of online recruitment for HIV research. *Ethics & behavior*. 2020;30(7):481-495.

8. Storozuk A, Ashley M, Delage V, Maloney EA. Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*. 2020;16(5):472-481.

9. Grey JA, Konstan J, Iantaffi A, Wilkerson JM, Galos D, Rosser BS. An updated protocol to detect invalid entries in an online survey of men who have sex with men (MSM): how do valid and invalid submissions compare? *AIDS Behav*. 2015;19(10):1928-1937.

10. Griffin M, Martino RJ, LoSchiavo C, et al. Ensuring survey research data integrity in the era of internet bots. *Quality & quantity*. 2021:1-12.

11. Grov C, Westmoreland D, Rendina HJ, Nash D. Seeing is believing? Unique capabilities of internet-only studies as a tool for implementation research on HIV prevention for men who

have sex with men: A review of studies and methodological considerations. *Journal of acquired immune deficiency syndromes (1999)*. 2019;82(Suppl 3):S253.

12. Yarrish C, Groshon L, Mitchell J, et al. Finding the signal in the noise: Minimizing responses from bots and inattentive humans in online research. *The Behavior Therapist*. 2019;42(7):235-242.

13. Saberi P, Campbell CK, Sauceda JA, Ndukwe S, Dubé K. Perceptions of Risks and Benefits of Participating in HIV Cure-related Research Among Diverse Youth and Young Adults Living with HIV in the United States. *AIDS Res Hum Retroviruses*. In Press;

14. Saberi P, Eskaf S, Campbell CK, Neilands TB, Sauceda JA, Dubé K. Exploration of a Mobile Technology Vulnerability Scale's Association with HIV Clinical Outcomes among Young Adults Living with HIV in the United States. *mHealth*. In Press;

15. Dewitt J, Capistrant B, Kohli N, et al. Addressing participant validity in a small internet health survey (The Restore Study): protocol and recommendations for survey response validation. *JMIR research protocols*. 2018;7(4):e96.

16. Hall EW, Sanchez TH, Stein AD, et al. Use of videos improves informed consent comprehension in web-based surveys among internet-using men who have sex with men: a randomized controlled trial. *J Med Internet Res*. 2017;19(3):e6710.

17. Macapagal K, Nery-Hurwit M, Matson M, Crosby S, Greene GJ. Perspectives on and preferences for on-demand and long-acting PrEP among sexual and gender minority adolescents assigned male at birth. *Sexuality Research and Social Policy*. 2021;18(1):39-53.

Figure 1: Results of Screening and Data Integrity Procedures

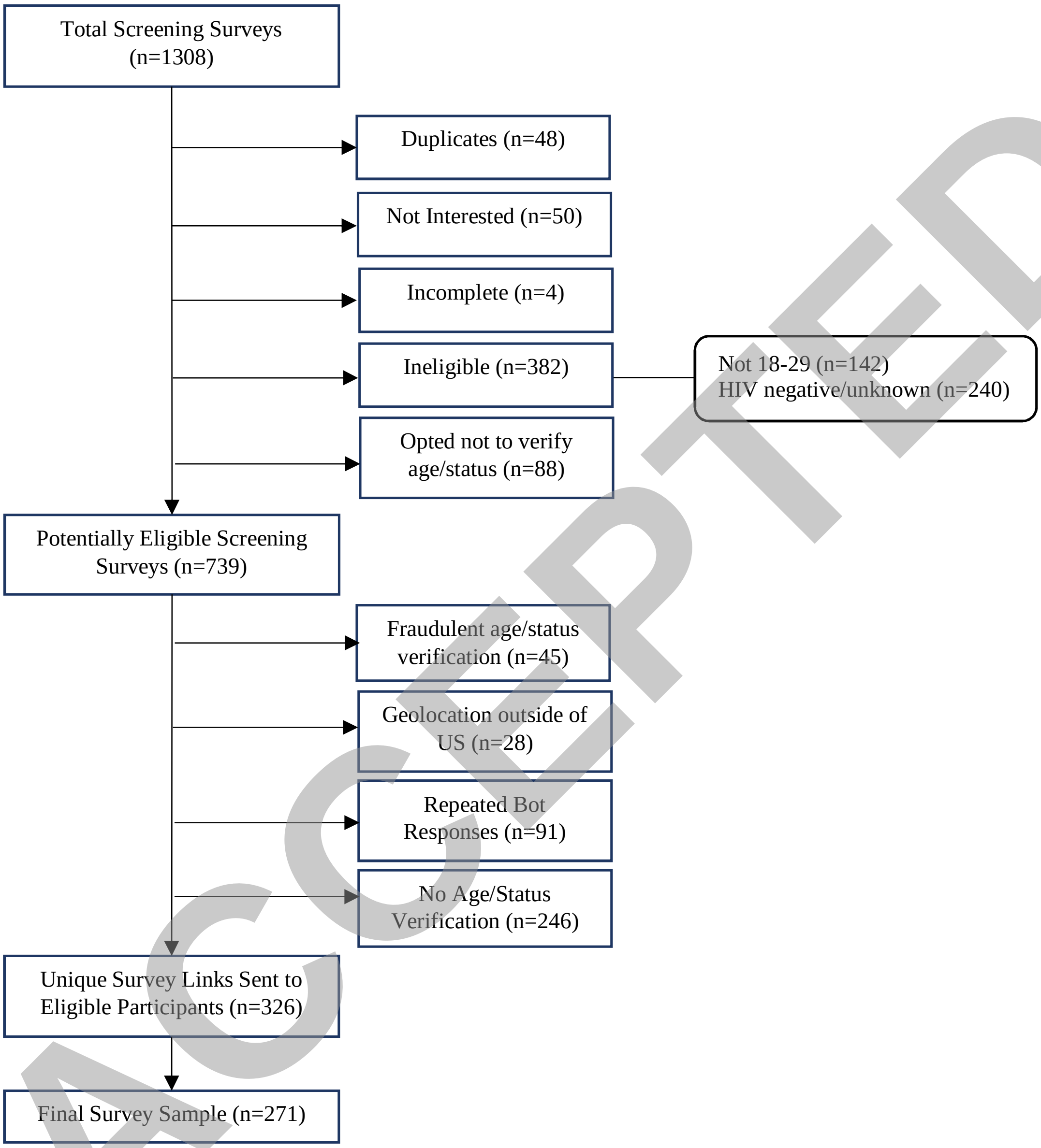
Total Screening Surveys (n=1308)
Duplicates (n=48)
Not Interested (n=50)
Incomplete (n=4)
Ineligible (n=382)
Not 18-29 (n=142)
HIV negative/unknown (n=240)
Opted not to verify age/status (n=88)
Potentially Eligible Screening Surveys (n=739)
Fraudulent age/status verification (n=45)
Geolocation outside of US (n=28)
Repeated Bot Responses (n=91)
No Age/Status Verification (n=246)
Unique Survey Links Sent to Eligible Participants (n=326)
Final Survey Sample (n=271)

Figure 1: Results of Screening and Data Integrity Procedures