Poster #284
Epidemiology and Public Health

CONFERENCE CAHR 2022

Canadian Association for HIV Research

OPTIONS LAB
Collaboratory in HIV/STI treatment and prevention science

Screening for Fraudulent Responses in a Web-Based Survey on Sexual Orientation Disclosure in Healthcare

Robinson Truong[1,2], David Brennan[3], Ann N. Burchell[1,4], Ross Upshur[2,4], & Darrell H.S. Tan[1,2,5,6]

1. MAP Centre for Urban Health Solutions, St. Michael's Hospital, Unity Health Toronto 2. Institute of Medical Science, University of Toronto 3. Factor-Inwentash Faculty of Social Work, University of Toronto 4. Dalla Lana School of Public Health, University of Toronto 5. Division of Infectious Diseases, St. Michael's Hospital, Unity Health Toronto 6. Department of Medicine, St. Michael's Hospital, Unity Health Toronto

Contact: Robinson.truong@unityhealth.to; Robinson.truong@mail.utoronto.ca

Background

- Fraudulent data have been a major barrier to response validity in web-based HIV research.
- Previously used strategies include limiting recruitment, using software, using data analytics, embedding anti-fraud questions, and using specific criteria to remove fraudulent responses.
- There has been much debate on the optimal practice and strategy of limiting or removing fraudulent online responses.

Objectives

- To describe a multi-pronged approach to removing fraudulent responses from a cross-sectional online survey about sexual orientation disclosure in primary care.

Methods

Eligibility Criteria

- Self-identify as a gay, bisexual or other man who has sex with / is sexually attracted to men.
- Is an Ontario Resident.
- Be able to communicate in English and provide informed consent for participation.

Recruitment

- Recruited participants via:
  - a) Information cards at Hassle Free Clinic, a major Toronto sexual health clinic
  - b) Posters in Hassle Free Clinic and LGBTQ2S+ Community-based organizations
  - c) Online advertisements via social media accounts of LGBTQ2S+ Community-based organizations
  - d) Geosocial networking sites such as Scruff, Jack'd, and Grindr
- Data collected using the Hosted In Canada survey platform

act

BLACK C•A•P

ACAS
Improving the Sexual Health of East & Southeast Asians

ASAAP
ALLIANCE FOR SOUTH ASIAN AIDS PREVENTION

GMSH Gay Men's Sexual Health Alliance

MAX
OTTAWA'S HEALTH CONNECTION FOR GUYS INTO GUYS

Methods cont.

Criteria

- 10 potential fraud-detection criteria were considered based on previous literature.

Participant paradata (behaviour) criteria		Participant response criteria
1) Duplicated IP addresses <u>Short survey completion time</u> 2) <5 minutes 3) <10 minutes 4) <15 minutes	<u>Survey start and end times</u> 5) Duplicated start times 6) Duplicated end times 7) Duplicated start or end times 8) Duplicated start and end times	9) Anti-fraud questions pertaining to: a) Embedded directives (e.g. QBotA: "Select the fifth option…") b) Community knowledge (e.g., QBotB: "When is Pride month in Ontario?...") c) Honeypot questions that only bots can see (e.g., QBotC: "If you are a bot, select any answers…") 10) Straightlining (i.e., repeated responses) on open-ended questions

Analysis

- Assuming IP address duplication is the most reliable criterion, we calculated tetrachoric correlation coefficients (r_{tet}) between this and other criteria.
- Assuming internal consistency of psychometric scales would improve with removal of fraudulent responses, we calculated Cronbach's alpha (α) for the 10-item Nebraska Outness Scale (NOS), after removing responses meeting each criterion. The NOS was used because half of the items are reverse scored.
- One criterion from both criteria 2-4 and 5-8 were selected based on the highest r_{tet} value. All other criteria were deemed acceptable if $\alpha<0.7$. Afterward, selected criteria were ordered from highest to lowest r_{tet} value to be applied sequentially to the study responses, findings were summarized in a table and a flow chart.

Results

Table 1: Screening of Fraudulent Criteria

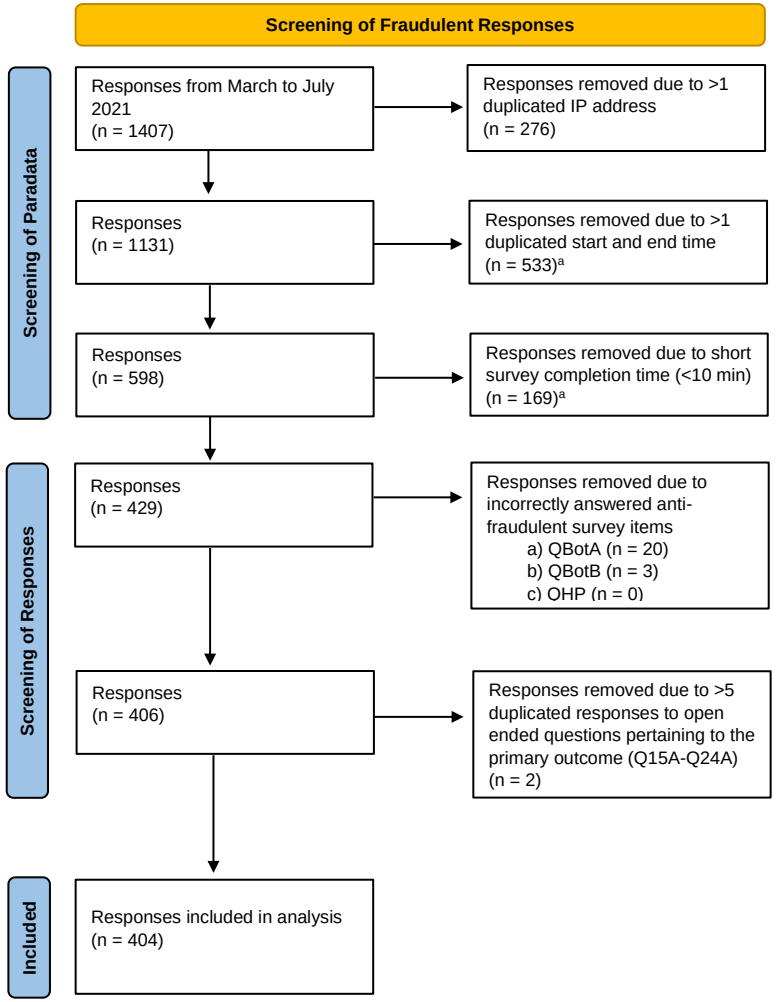
Criteria	n that met criterion	r_{tet}[a]	α after criterion applied	Overall n after applying criterion sequentially[b]	α after applying criterion sequentially
No criteria applied	NA	NA	NA	NA	0.743
1) Screening of paradata					
IP addresses are duplicated[a]	276	N/A	0.728	1131	0.728
Criteria based on survey start and end time					
Start time of survey were duplicated	763	0.183	0.836	N/A	N/A
End time of survey were duplicated	790	0.205	0.836	N/A	N/A
Start or end time of survey were duplicated	848	0.149	0.841	N/A	N/A
Start and end time of survey were duplicated	705	0.239	0.832	598	0.827
Criteria based on survey completion time					
Survey completion time <5 minutes	246	0.310	0.770	N/A	N/A
Survey completion time <10 minutes	630	0.316	0.795	429	0.833
Survey completion time <15 minutes	916	0.284	0.774	N/A	N/A
2) Screening of responses					
Responses removed due to incorrectly answered anti-fraud survey items	78	0.0233	0.747	406	0.835
Straightlining on at least half of the open-ended questions	70	0.622	0.719	404	0.836

[a] Tetrachoric correlation coefficients for each criterion was calculated in comparison to the IP address duplication criterion

[b] Selection of criteria for sequential removal was based on the highest tetrachoric correlation coefficients per criteria group

- The optimal criteria to remove fraudulent responses were 1) duplicated IP addresses, 2) identical start and end times, 3) completion time <10 minutes, 4) anti-fraud questions and 5) straightlining on open-ended responses.
- Most criteria had a positive weak association (range r_{tet}:0.183-0.622) to the IP address criteria. However, almost all individual criteria had a moderate internal consistency (min α=0.719).
- In the end, the five selected criteria yield the highest internal consistency (α=0.836) when applied sequentially, resulting in 404 responses (out of 1407) being included in the final analysis.

Results cont.



[a] Response time seconds were rounded down to the nearest minute

Figure 1. Screening of Fraudulent Responses

Discussion

Strengths

- We screened a large sample size (n=1407), explored numerous criteria, and demonstrated that internal consistency increases with each criterion applied.

Limitation

- Despite the use of various criteria, we cannot be certain that the included data does not contain fraudulent responses. Non-fraudulent responses could have also been lost with each criteria applied.

Conclusions

- Paradata, anti-fraud questions and serial assessment of internal consistency are useful ways to remove fraudulent data from web-based surveys. Future online HIV research should consider other rigorous criteria or tools in removing fraudulent responses, such as validated anti-fraud questions.

Acknowledgements

OPTIONS LAB Collaboratory in HIV/STI treatment and prevention science
UNIVERSITY OF TORONTO FACULTY OF MEDICINE
ST. MICHAEL'S UNITY HEALTH TORONTO Research Training Centre
ONTARIO HIV TREATMENT NETWORK
CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada