

# Canadian Oncology Nursing Journal

## Revue canadienne de soins infirmiers en oncologie

---

Volume 33, Issue 2 • Spring 2023  
eISSN: 2368-8076



Canadian Association of Nurses in Oncology  
Association canadienne des infirmières en oncologie

# “Taking out the trash”: Strategies for preventing and managing fraudulent data in web-surveys

by Prabdeep Panesar and Samantha J. Mayo

## INTRODUCTION

Given the rapid adoption of the internet and mobile technology, conducting online research has become more accessible for scientists, as they are able to engage numerous participants while collecting data quickly and efficiently. Online research, particularly anonymous web-surveys, allow participants to partake in research at their own convenience while being able to maintain their privacy (Wyatt, 2000). This method further allows researchers to target unique population types in a cost-effective manner, while having increased reach and scalability, making it easier to collect a large sample size, in real-time, and under real-world conditions (Heffner et al., 2021; Teitcher et al., 2015).

Despite the advantages of web-based surveys, ensuring data integrity of such research techniques continues to pose a challenge. Anonymous and incentivized surveys may be particularly vulnerable to multiple responses from individuals, responses from those who falsely claim to meet the eligibility criteria, or robot (bot) submissions. Such fraudulent responses can lead to misleading or biased data impacting the accuracy of the results and significantly undermining the data integrity. This may further affect the scientific and clinical implications derived from the results, therefore negatively influencing the social benefit of the research being conducted. These issues highlight the importance

of implementing strategies to detect and prevent fraudulent responses to maintain data integrity when conducting web-surveys. Therefore, the purpose of this article is to review strategies to prevent the collection of fraudulent responses in web-surveys, as well as methods to manage fraudulent responses after data collection has been completed.

## PREVENTATIVE STRATEGIES

Prior to disseminating web-surveys it is essential for researchers to consider implementing methods to prevent the collection of fraudulent responses or facilitate the identification of fraudulent responses. Anti-deception methods may include implementing specific types of survey questions, utilizing software tools, and additional pre-cautionary measures.

**Survey questions.** During the development of survey questions researchers can include combinations of questions designed to identify inconsistencies. In an online incentivized survey of smoking behaviours, Choi and colleagues (2017) posed multiple questions designed for this purpose. For example, they posed two questions in their survey about daily cigarette use; one as an open text-box response (*On average, how many cigarettes do you smoke each day?*) and another as a multiple-choice question (*How many cigarettes/day do you smoke? – Options including: 10 or less; 11–20; 21–30; 31 or more;* Choi et al., 2017). With these questions, the authors identified that more than a third (97/270) of the collected surveys had incongruent responses and, subsequently, excluded these surveys from the analysis (Choi et al., 2017).

For non-anonymized surveys, including questions to collect participants' identifying information (e.g., name, email) may help researchers identify duplicate or suspicious responses that might be considered for omission.

Indicators reported in previous studies have included: multiple submissions from the same name; miscellaneous letters included at the end of a name being repeated (e.g., Johna, Johnb, Johnc); and email addresses with long strings of random characters (e.g., 5–10) (Pratt-Chapman et al., 2011; Teitcher et al., 2015).

**Software.** Software tools can also help to determine duplicate or suspicious responses and are often available as optional features within popular web-survey platforms. Such software tools include Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA), which can deter bot-related submissions by posing challenging questions only humans are able to complete. Alternatively, software programs can also include hidden questions that only bots can detect and provide a response to (Heffner et al., 2021). Other software tools can track and help identify potentially duplicate or non-legitimate responses from humans (Pozzar et al., 2020). For example, numerous software programs are able to collect paradata such as survey start and stop time and mouse movement on the screen to understand how participants may be navigating the survey; the programs can also collect IP addresses, geolocation, or internet cookies to determine multiple entries (Heffner et al., 2021). With the assistance of such software tools, researchers have identified and excluded responses made by bots, multiple submissions from a single individual, and those that have completed the survey despite being ineligible for the study. For example, when evaluating the start and stop time for study completion, researchers have been able to identify when respondents have completed the survey within an unrealistic timeframe warranting further investigation (Kramer et al., 2014; Salinas, 2023). Also, when collecting IP addresses, investigators were able to

## ABOUT THE AUTHORS

Prabdeep Panesar, BSc, Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, ON

Samantha J. Mayo, RN, PhD, Lawrence S. Bloomberg Faculty of Nursing, University of Toronto, Toronto, ON; Princess Margaret Cancer Centre, University Health Network, Toronto, ON

evaluate whether multiple responses were generated from one individual and further verify whether participants fulfilled the geographical eligibility of the study (Choi et al., 2017).

**Additional strategies.** Fraudulent data may also be deterred by emphasizing the consequences of submitting fraudulent responses and highlighting data surveillance. For example, researchers may benefit from providing explicit disclaimers at the beginning of their surveys including statements such as “participants will not be compensated if suspected of providing fraudulent responses” or that “investigators can contact you by telephone or email to confirm your eligibility for the study” (Teitcher et al., 2015). Moreover, while it may be beneficial to clearly state the inclusion and exclusion criteria for eligible participants, this lets individuals know what responses to avoid when completing screening questions allowing them to access the survey. To avoid this, Choi and colleagues (2017) masked their study’s eligibility criteria and included five screening questions that prevented misrepresentation of eligibility and unambiguous exclusion of ineligible participants. Investigators may further consider presenting incentives as a raffle compared to offering a gift card for each response as the low incentive may outweigh the benefit of providing fraudulent data.

## REFERENCES

- Choi, S. H., Mitchell, J., & Lipkus, I. (2017). Lessons learned from an online study with dual-smoker couples. *American Journal of Health Behavior*, 41(1), 61–66. <https://doi.org/10.5993/ajhb.41.1.6>
- Heffner, J. L., Watson, N. L., Dahne, J., Croghan, I., Kelly, M. M., McClure, J. B., Bars, M., Thrul, J., & Meier, E. (2021). Recognizing and preventing participant deception in online nicotine and tobacco research studies: Suggested tactics and a call to action. *Nicotine & Tobacco Research : Official journal of the Society for Research on Nicotine and Tobacco*, 23(10), 1810–1812. <https://doi.org/10.1093/ntr/ntab077>
- Kramer, J., Rubin, A., Coster, W., et al. (2014). Strategies to address participant misrepresentation for eligibility in Web-based research. *Int J Methods Psychiatr Res.*, 23(1), 120-129. <https://doi.org/10.1002/mpr.1415>
- Pozzar, R., Hammer, M. J., Underhill-Blazey, M., Wright, A. A., Tulsy, J. A., Hong, F., Gundersen, D. A., & Berry, D. L. (2020). Threats of bots and other bad actors to data quality following research participant recruitment through social media: Cross-sectional questionnaire. *Journal of Medical Internet Research*, 22(10), e23021. <https://doi.org/10.2196/23021>
- Pratt-Chapman, M., Moses, J., & Arem, H. (2021). Strategies for the identification and prevention of survey fraud: Data analysis of a web-based survey. *JMIR Cancer*, 7(3), e30730. <https://doi.org/10.2196/30730>
- Salinas, M. R. (2023). Are your participants real? Dealing with fraud in recruiting older adults online. *Western Journal of Nursing Research*, 45(1), 93–99. <https://doi.org/10.1177/01939459221098468>
- Teitcher, J. E., Bockting, W. O., Bauermeister, J. A., Hofer, C. J., Miner, M. H., & Klitzman, R. L. (2015). Detecting, preventing, and responding to “fraudsters” in internet research: Ethics and tradeoffs. *The Journal of Law, Medicine & Ethics : A Journal of the American Society of Law, Medicine & Ethics*, 43(1), 116–133. <https://doi.org/10.1111/jlme.12200>
- Wyatt J. C. (2000). When to use web-based surveys. *Journal of the American Medical Informatic Association: JAMIA*, 7(4), 426–429. <https://doi.org/10.1136/jamia.2000.0070426>

## DATA CLEANING

Despite implementing preventative measures to deter the collection of fraudulent responses, it is essential to analyze the data and remove suspicious submissions following the completion of data collection. After eliminating any cases that are easily identifiable as ineligible, researchers may begin to screen the data for potential indicators of fraudulent data. As previously mentioned, the collected data may further be assessed for conflicting responses to combined questions and/or duplicate or suspicious names or emails (Pratt-Chapman et al., 2011; Teitcher et al., 2015). Similarly, open text-box responses can also be assessed for duplicate or similar responses, as well as answers that do not address the question (Choi et al., 2017; Pratt-Chapman et al., 2011). To avoid the omission of legitimate data, researchers may decide to create decision rules, such as only excluding cases if they fulfill at least two of the listed indicators. Pratt-Chapman and colleagues (2021) executed similar techniques and decision rules to identify fraudulent responses in an online survey of cancer survivors, which was advertised by email and social media. After screening for and excluding responses meeting two or more criteria (e.g., incongruent responses, irregular timestamps), the researchers excluded 1,408 responses from the 1977 total

responses collected. When later comparing the data collected prior to and after removal of fraudulent responses, statistically significant differences were observed across demographic characteristics including age, gender, education, cancer stage, cancer type, health status, and insurance coverage (Pratt-Chapman, 2021). This emphasizes the importance of creating criteria to screen for and identify fraudulent responses to ensure data integrity.

## CONCLUSION

Given the increased chances of collecting fraudulent responses when conducting online research such as web-surveys, it is essential to develop anti-deception protocols and criteria to identify suspicious submissions once data collection is complete. While online research does provide a range of benefits, it is essential to have multiple safeguards in place to ensure data integrity, while also considering the ethical implications of such methods. Overall, better informing researchers about these deceptions and, in accordance, taking pre-cautionary measures, we maintain the scientific and clinical implications derived from such research methods.