Crowdsourced data may inaccurately represent some population groups

June 22 2017

Credit: George Hodan/public domain

While crowdsourcing, a practice that enables study participants to submit data electronically, has grown in use for health and medical research, a study led by UC San Francisco comparing the online approach to a standard telephone survey has found that certain crowdsourced groups are either over- or underrepresented by age, race/ethnicity, education and physical activity.

Their findings, appearing online June 22, 2017, in the *American Journal of Public Health*, suggest greater attention needs to be given to determining what populations are and are not reachable using remote, electronic data collection platforms, and studies relying on crowdsourced respondents need to define the profile of the people generating those data, the researchers said.

Proactive efforts also are needed to understand and promote inclusion of underrepresented groups within projects using crowdsourced recruitment and data collection, they said.

"Online crowdsourced recruitment leads to systematic underrepresentation of some U.S. adults - such as certain racial and ethnic minorities, those with lower educational attainment, and older adults - and overrepresentation of others," said lead author Veronica Yank, MD, assistant professor of medicine at UCSF.

The growth of Internet-based sampling and data collection offers an important opportunity for cheaper, higher volume collection of health -related data. However, the resulting data may not be generalizable, and if certain groups are over- or underrepresented, the research may generate misleading conclusions when extrapolated to larger populations.

How participants enter a particular study also can greatly influence generalizability of results. Population studies traditionally have spent considerable effort on targeted recruitment of representative samples and statistical adjustments for over- or underrepresentation of subgroups among those enrolled, allowing investigators to determine the degree of confidence about the representativeness of the data.

Crowdsourcing, in which self-selected individuals provide electronic data or feedback, is currently one of the most innovative methods for study population accrual. Social science and psychology researchers

widely use it, and the National Institutes of Health Precision Medicine Initiative will recruit participants this year through the Internet, social media and mobile technologies to form an "All of Us" cohort, the largest study group ever undertaken.

At UCSF, the ongoing Health eHeart Study currently has consented more than 100,000 participants, toward a goal of 1 million, and includes any interested adult with an email address. It harnesses the power of online and mobile technology to gather cardiovascular data through devices such as smartphone apps, ECG smartphone cases and portable blood pressure cuffs.

In their *American Journal of Public Health* study, Yank and her colleagues utilized Amazon Mechanical Turk (MTurk), the world's largest online crowdsourcing platform, to compare demographic and health characteristics of adults recruited through it to those of the U.S. population. They focused on health characteristics that are known risk factors for cardiovascular disease, many of which are suitable for remote measurement and data collection.

MTurk has 500,000 registered anonymous members overall, with about 400,000 in the United States and 15,000 active on any given day. Businesses or other entities who would like members to complete short tasks for compensation post descriptions on the platform website, with compensation of $0.10-$0.25 per 10 minutes. For this study, 2,015 U.S.-based adults at least age 18 completed the survey between July-August 2015.

For comparison, the researchers used 2013 data of 428,211 respondents from the Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System (BRFSS), the world's largest telephone health survey. Administered annually in English by landline or cell phone to adult U.S. residents, the survey gathers cross-sectional data on

demographic and health characteristics and disease risk factors. The selected questions focused on demographics, ethnicity, educational attainment, annual income, employment status, and individual characteristics known to influence cardiovascular morbidity and mortality.

Overall, compared to the BRFSS, the crowdsourced samples tended to be overrepresented in the 20-39 age range and underrepresented in the 40-75 age range. The cardiovascular disease risk profile of crowdsourced participants also differed in well-defined ways from the U.S. population, the researchers found.

Crowdsourced participants were younger, more likely to be non-Hispanic and white, and had higher levels of educational attainment. Those age 40-59 were most representative with regard to smoking, diabetes, hypertension and hyperlipidemia, but even they had significant differences with regard to race/ethnicity, education and physical activity. Crowdsourced data from younger age groups were even less similar, and those age 60 and older were difficult to reach by crowdsourcing.

As a result, policymakers, funders of research and researchers should be explicit about the advantages and limitations of relying on crowdsourced data, especially when underlying sociodemographic characteristics or health variables may influence health outcomes, Yank said. The target population for the research or policy question must be reachable through the electronic platform, knowing certain groups will be underrepresented in the resulting data. The possible need for statistical adjustment for nonrepresentative samples also should be built into each study design.

"These findings have implications for the upcoming national Precision Medicine Initiative, which will use online crowdsourcing as one of its recruitment and data collection approaches for the million Americans it plans to enroll in its cohort," Yank said.

Among the study limitations, the researchers tested only one crowdsourcing platform, with relatively low hourly incentive. The Internet protocol addresses used for determining crowdsourcing respondent locations also might represent sites of employment or other community hubs, which are more likely clustered in urban settings, and the BRFSS has known limitations.

Provided by University of California, San Francisco

Citation: Crowdsourced data may inaccurately represent some population groups (2017, June 22) retrieved 25 April 2024 from https://medicalxpress.com/news/2017-06-crowdsourced-inaccurately-population-groups.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.