Preventive Medicine 102 (2017) 93–99

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Crowdsourced data collection for public health: A comparison with nationally representative, population tobacco use data

John D. Kraemer [a], Andrew A. Strasser [b], Eric N. Lindblom [c], Raymond S. Niaura [d,e,f], Darren Mays [f,*]

[a] Department of Health Systems Administration, School of Nursing & Health Studies, Georgetown University, Washington, DC, United States
[b] Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[c] O'Neill Institute for National and Global Health Law, Georgetown University Law Center, Washington, DC, United States
[d] Schroeder Institute for Tobacco Research and Policy Studies, Truth Initiative, Washington, DC, United States
[e] Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[f] Department of Oncology, Georgetown University Medical Center, Cancer Prevention & Control Program, Lombardi Comprehensive Cancer Center, Washington, DC, United States

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form 30 June 2017
Accepted 5 July 2017
Available online 8 July 2017

Keywords:
Tobacco control
Young adult
Crowdsourcing

ABSTRACT

Introduction. Internet-based crowdsourcing is increasingly used for social and behavioral research in public health, however the potential generalizability of crowdsourced data remains unclear. This study assessed the population representativeness of Internet-based crowdsourced data.

Methods. A total of 3999 U.S. young adults ages 18 to 30 years were recruited in 2016 through Internet-based crowdsourcing to complete measures taken from the 2012–2013 National Adult Tobacco Survey (NATS). Post-hoc sampling weights were created using procedures similar to the NATS. Weighted analyses were conducted in 2016 to compare crowdsourced and publicly-available 2012–2013 NATS data on demographics, tobacco use, and measures of tobacco perceptions and product warning label exposure.

Results. Those in the crowdsourced sample were less likely to report an annual household income of $50,000 or greater, and e-cigarette, waterpipe, and cigar use were more prevalent in the crowdsourced sample. High proportions of both samples indicated cigarette smoking is very harmful and very addictive. Comparable proportions of non-smokers and smokers reported cigarette warning label exposure, however the likelihood of reporting that smoking is very harmful by frequency of warning label exposure was lower among smokers in the crowdsourced sample.

Conclusions. Our findings indicate that crowdsourced samples may differ demographically and may not produce generalizable estimates of tobacco use prevalence relative to population data after post-hoc sample weighting. However, correlational analyses in crowdsourced samples may reasonably approximate population data. Future studies can build from this work by testing additional methodological strategies to improve crowdsourced sampling strategies.

1. Introduction

Internet crowdsourcing, defined as a "distributed problem-solving and production model that leverages the collective intelligence of online communities," is a tool with potential to address public health challenges (Brabham et al., 2014). Crowdsourcing offers an efficient way to obtain information from online respondents more quickly than some conventional data collection tools. Crowdsourcing applications entail varying involvement from participants, including answering questions (e.g., surveys); providing feedback on concepts (e.g., policies, programs); coding data (e.g., images); and creating user-generated content (e.g., communication messages) (Brabham et al., 2014). Researchers are increasingly using crowdsourcing data collection, particularly in social and behavioral sciences (Bohannon, 2016).

Crowdsourcing has value because data collection is efficient and relatively low cost and participants are readily available without geographic constraints (Brabham et al., 2014; Gosling and Mason, 2015; Mason and Suri, 2012). Examples of research using crowdsourcing include tobacco control (Mays et al., 2016a; Mays et al., 2016b; Leas et al., 2016; Brewer et al., 2016), skin cancer prevention (Mays and Tercyak, 2015), and sexual behavior (Syme et al., 2017). Research using crowdsourcing includes observational studies to characterize specific constructs such as health beliefs, and correlational investigations of how exposures such as health messaging relate to outcomes such as beliefs or behavior (Mays et al., 2016a; Mays et al., 2016b; Leas et al., 2016; Brewer et al., 2016; Mays and Tercyak, 2015; Syme et al., 2017). Peer-reviewed papers using a single crowdsourcing platform (Amazon Mechanical Turk) increased from 61 in 2011 to 1120 in 2015 (Bohannon, 2016). Increasing interest has spurred development of methodological tools for researchers (Litman et

* Corresponding author at: Department of Oncology, Georgetown University Medical Center, Cancer Prevention & Control Program, Lombardi Comprehensive Cancer Center, 3300 Whitehaven Street NW, Suite 4100, Washington, DC 20007, United States.
E-mail address: dmm239@georgetown.edu (D. Mays).

http://dx.doi.org/10.1016/j.ypmed.2017.07.006
0091-7435/

al., 2016), and crowdsourcing is appearing in funding opportunities from agencies such as the National Institutes of Health.

Crowdsourcing platforms can replicate data collected using validated behavioral measures and tasks (Briones and Benham, 2016; Crump et al., 2013), and crowdsourcing samples may provide greater demographic diversity than traditional convenience samples (e.g., college students) (Briones and Benham, 2016; Berinsky et al., 2012). There are concerns about generalizability since crowdsourced participants are those with technology access who are motivated to engage in research, and because there may be relatively small numbers of individuals in crowdsourced participant pools who meet specific eligibility criteria at any given time (Brabham et al., 2014; Chandler et al., 2014). There is also some evidence indicating that crowdsourced samples may differ from the population on measures relevant to public health research, such as political beliefs (Chandler and Shapiro, 2016).

Several recent studies have used crowdsourcing to examine questions aimed at informing Food and Drug Administration (FDA) tobacco regulations (Mays et al., 2016a; Mays et al., 2016b; Leas et al., 2016; Brewer et al., 2016; Pearson et al., 2016). Such an efficient data collection approach has potential value for tobacco regulatory science because FDA is charged with supporting regulations using population data and often such data need to be generated within a short timeframe to inform regulations (Husten and Deyton, 2013). Crowdsourcing also provides the capability to reach priority groups to inform tobacco regulations, such as tobacco users and nonusers and specific demographic groups (Husten and Deyton, 2013). Although crowdsourcing is increasingly used for tobacco research, as noted above prior studies have used crowdsourcing for topics ranging from skin cancer prevention (Mays and Tercyak, 2015) to sexual risk behavior (Syme et al., 2017), and its use is increasing overall (Bohannon, 2016). Empirical evidence on how crowdsourcing can be used to inform tobacco regulation as a case study can guide research in other public health domains as well.

This study empirically examined the potential contributions of crowdsourced data in public health research by comparing crowdsourced data to nationally representative U.S. survey data. The study focuses on tobacco use as an example because deidentified, nationally representative data from the National Adult Tobacco Survey (NATS) are publicly available from the U.S. Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2016). This study additionally focused on young adults ages 18 to 30 years because they are defined as a priority population for tobacco research (National Cancer Institute, 2016) and can be accessed through crowdsourcing platforms where participation is limited to adults ages 18 and older. Our aim was to determine if crowdsourced data provides similar estimates of demographics, tobacco use behavior, and tobacco risk perceptions by comparing data collected through Amazon Mechanical Turk to the NATS. Additionally, we aimed to determine if correlational analyses in crowdsourced data are similar to population data, drawing from research indicating exposure to tobacco warning labels can affect risk perceptions (Mays et al., 2016a; Mays et al., 2016b; Leas et al., 2016; Brewer et al., 2016). Our study focused on these specific measures because monitoring population-level trends in tobacco use behavior, perceptions, and exposure to interventions such as warning labels is critical to tobacco regulatory decision-making (Husten and Deyton, 2013). We tested the study aims by comparing data collected through Amazon Mechanical Turk to the NATS on demographics, tobacco use behaviors and perceptions, and exposure to tobacco warnings using parallel measures and sample weighting strategies for crowdsourced data.

2. Methods

2.1. Sampling

2.1.1. National Adult Tobacco Survey

The most recent population-based tobacco use dataset at the time of the study was the 2012–2013 NATS (Center for Disease Control and Prevention, 2015). The NATS is a stratified, random-digit dialed landline and cellular telephone survey of non-institutionalized adults ≥ 18 years old residing in the 50 U.S. states and the District of Columbia (Agaku et al., 2014). From October 2012 to July 2013, 60,192 interviews were conducted (44.9% response rate) including 6682 young adults ages 18 to 30 in the analytic sample (Center for Disease Control and Prevention, 2015; Agaku et al., 2014). NATS data are weighted by inverse probability of selection, adjusted for nonresponse and household characteristics, and raked to population totals on state, age, gender, race/ethnicity, marital status, education, and phone type (Centers for Disease Control and Prevention, 2014).

2.2. Crowdsourced data

Crowdsourced data were collected in April 2016 through Amazon Mechanical Turk, an Internet marketplace where researchers can post "human intelligence tasks" including surveys or other data collection (Crump et al., 2013). After reviewing a brief study description inviting them to take a survey with questions about tobacco use, Mechanical Turk members interested in participating reviewed a more comprehensive description with a link to a consent form and eligibility screener. Young adults ages 18 to 30 were eligible to participate, access to the task was limited to those with accounts registered in the U.S. All study procedures were implemented in English without translation to other languages. To ensure representation of cigarette smokers, smoking status was assessed at screening and smokers were oversampled to reflect 40% of the sample. Post-hoc weighting (described below) was used to adjust to the national smoking prevalence in the NATS. Eligible, consenting individuals proceeded to an online survey consisting of measures described below. The target sample was 4000 respondents—the maximum practical sample that could be achieved with study resources and a target that meets or exceeds those of similar crowdsourced studies to date (Mays et al., 2016a; Mays et al., 2016b; Leas et al., 2016; Brewer et al., 2016). Participants completing procedures were given a $1 monetary credit through Mechanical Turk. The data collection protocol was reviewed and determined to be exempt by Georgetown University's IRB.

2.3. Measures

For comparison to the NATS dataset, the crowdsourced data collection used measures directly from the NATS wherever possible. Any differences are noted below. We selected a subset of measures from the NATS for crowdsourced data collection due to practical constraints on the number of items that could be implemented using crowdsourcing (Mason and Suri, 2012) and based on priority topics for tobacco regulatory science described above (Husten and Deyton, 2013).

2.4. Demographics

Demographics were assessed using NATS items including age, gender, race/ethnicity, education, and marital status (Agaku et al., 2014). NATS measures household income using multi-question probing identifying general income levels (e.g., less than $50,000) and determining specific levels through follow-up questions (e.g., $30,000 to $40,000, $40,000 to $50,000) (Center for Disease Control and Prevention, 2015). This type of measure could not practically be administered online, so a single item asking "What was your household income before taxes last year? Please report the income that is most important to you, whether that is your own income or your parents" was used (Mays et al., 2016a; Mays et al., 2016b). Respondents were grouped into similar income categories across the samples.

2.5. Tobacco use

Measures captured use of cigarettes, electronic cigarettes, hookah (waterpipe tobacco), cigars/cigarillos/filtered little cigars, and smokeless