RTI Press

Methods Report

ISSN 2378-7813

September 2021

Modeling the Probability of Fraud in Social Media in a National Cannabis Survey

Lauren M. Dutra, Matthew C. Farrelly, Brian Bradfield, Jamie Ridenhour, and Jamie Guillory

RTI INTERNATIONAL

RTI Press publication MR-0046-2109

RTI International is an independent, nonprofit research organization dedicated to improving the human condition. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Dutra, L. M., Farrelly, M. C., Bradfield, B., Ridenhour, J., and Guillory J. (2021). *Modeling the Probability of Fraud in Social Media in a National Cannabis Survey*. RTI Press Publication No. MR-0046-2109. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2021.mr.0046.2109

This publication is part of the RTI Press Methods Report series.

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

Contents

About the Authors

Lauren M. Dutra, ScD, is a research scientist at RTI International.

Matthew C. Farrelly, PhD, is Chief Scientist and Director of the Center for Health Analytics, Media, and Policy (CHAMP) at RTI International.

Brian Bradfield, BA, is an economist at RTI International.

Jamie Ridenhour, MStat, is a research statistician at RTI International.

Jamie Guillory, PhD, is an RTI contractor with Prime Affect Research.

RTI Press Associate Editor

Jenny Wiley

Acknowledgments

The authors would like to thank Phil Kott, Jill Dever, Kian Kamyab, Josh Goetz, Jessica Pikowski, Carla Bann, and Gary Zarkin. This research was funded by RTI International.

Abstract

Cannabis legalization has spread rapidly in the United States. Although national surveys provide robust information on the prevalence of cannabis use, cannabis disorders, and related outcomes, information on knowledge, attitudes, and beliefs (KABs) about cannabis is lacking. To inform the relationship between cannabis legalization and cannabis-related KABs, RTI International launched the National Cannabis Climate Survey (NCCS) in 2016. The survey sampled US residents 18 years or older via mail (n = 2,102), mail-to-web (n = 1,046), and two social media data collections (n = 11,957). This report outlines two techniques that we used to problem-solve several challenges with the resulting data: (1) developing a model for detecting fraudulent cases in social media completes after standard fraud detection measures were insufficient and (2) designing a weighting scheme to pool multiple probability and nonprobability samples. We also describe our approach for validating the pooled dataset. The fraud prevention and detection processes, predictive model of fraud, and the methods used to weight the probability and nonprobability samples can be applied to current and future complex data collections and analysis of existing datasets.

Introduction

Cannabis legalization is rapidly spreading throughout the United States.[1] In 2010, 27 percent of Americans lived in states with legal recreational and medical cannabis or medical cannabis only; by 2018, this figure had more than doubled to 56 percent.[2–4] In this rapidly evolving legal environment, cannabis use has increased. According to the National Survey on Drug Use and Health (NSDUH), national past-month cannabis use increased significantly between 2002 and 2016 among 18-to-25-year-olds (17.3 percent to 20.8 percent, $P < 0.05$) and adults 26 years old and older (4.0 percent to 7.2 percent, $P < 0.05$).[5]

Validated population-level surveys of cannabis use, such as NSDUH, primarily focus on establishing the prevalence of cannabis use alone or in combination with the use of other substances.[6] For example, whereas NSDUH assesses perceived risk and availability of cannabis, it does not provide additional information on knowledge, attitudes, and beliefs (KABs) about cannabis.[7] In addition, the relationship between cannabis policies and use remains somewhat unclear, partly due to difficulty obtaining individual-level information on cannabis use combined with geographic identifiers.[8] Access to these datasets is often restricted to prevent confidentiality. As a result, the predictors of national cannabis use remain unclear.

To address the lack of national information on the relationship between KABs, cannabis policies, and cannabis use at the time, RTI International launched the National Cannabis Climate Survey (NCCS) in August 2016. The survey combined address-based (probability) and social media (nonprobability) samples to obtain information about the relationship between the cannabis legal environment (recreational and medical legalization, medical only legalization, or neither), KABs, and cannabis use behaviors among the general population and adult cannabis users.

The NCCS combined probability and nonprobability samples to balance the advantages and disadvantages of these two types of samples.[9] Probability samples (e.g., address-based samples; ABS) provide broad coverage of the US household population,[10,11] result in less coverage bias than nonprobability samples,[9] and are generally subject to very little fraud.[12–14] Nonprobability samples, such as social media samples, are efficient for accessing hard-to-reach and rare populations,[15] such as current cannabis users. However, nonprobability samples are susceptible to constantly evolving methods of fraud,[14] such as multiple submissions of a survey by the same individual (often with varying identifying information to attempt to escape detection),[16] manipulating answers to screen into studies ("gaming the survey"),[13,14] and bots,[14] among others.

Several fraud prevention procedures (designed to prevent fraudulent completes of surveys) have been identified for social media samples, including asking participants not to complete surveys more than once or asking if they have previously completed the survey and collecting identifying information, such as e-mail addresses, IP addresses, and zip codes.[12,14,17] In addition, several established fraud detection procedures exist. These procedures, which are applied to remove fraudulent completes after data collection has occurred, often include deduplication (removal of duplicate entries) and cross-validation (confirming that the participant met inclusion criteria).[12–14,16–18]

All of these procedures, however, have limited efficacy in detecting and removing fraudulent completes.[14] A few studies have identified additional methods of identifying fraud after data collection. These studies identified distinguishing characteristics of fraudulent responses and used these characteristics to identify potential fraudulent responses.[14,16,19] Generally, these techniques rely on examining one indicator of fraud at a time, usually through bivariate comparisons. However, as was the case with the NCCS, fraudulent completes can present as patterns of responses across multiple variables, resulting in the need for more sophisticated fraud detection methods than bivariate analyses. To address this issue, we developed a fraud prediction model to calculate the probability that each response was fraudulent based on patterns of responses to key variables. To our knowledge, this is the first publication to use multivariable modeling to calculate the probability of fraud in a social media sample.

This manuscript describes the fraud model, the weighting scheme that we used to calibrate multiple probability and nonprobability samples after eliminating fraudulent responses, and the validation

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

of the survey results. The fraud model presented in this manuscript has two advantages over existing methods of identifying characteristics of fraudulent responses: it combines patterns of responses for multiple variables to determine fraud, and it produces a continuous probability that researchers can use to carefully evaluate the likelihood of fraud for each participant. The fraud model can be applied to existing and future datasets when traditional fraud prevention and detection methods are insufficient. Because of rapidly evolving (and increasingly sophisticated) methods of committing fraud on social media,[14] multivariable methods of identifying fraud are and will continue to be needed.

Materials and Methods

Sample

Between August 2016 and May 2017, we collected data for the NCCS through two ABS household (probability) samples and two social media (nonprobability) convenience samples. The purpose of the survey was to compare cannabis-related KABs across states with three different cannabis legal environments: states with recreational and medical cannabis laws, states with medical cannabis laws only, and states with neither medical nor recreational cannabis laws. We sampled an approximately equal number of addresses from each legal environment by using stratified sampling methods for the ABS samples and quotas for the social media samples. Inclusion and exclusion criteria were identical for all modes of data collection except when noted in the following sections. Participants had to be 18 years of age or older and live in the continental United States. The RTI International Institutional Review Board approved all procedures.

ABS Samples

We obtained two ABS samples (Figure 1) from RTI International's in-house ABS frame (http://abs.rti.org), which is sourced from the US Postal Service Computerized Delivery Sequence file (CDS). The CDS, which is updated monthly, contains all mail delivery points in the United States, and as is the case with most ABS samples, offers high coverage of the household population for mailed surveys.[20]

Mail 1.0 and Mail-to-Web Samples

The first ABS sample ("Mail 1.0") included 5,000 addresses (Table 1). We mailed these households a paper survey with a $5 incentive; 1,280 participants returned the mail survey. Of the 3,720 households that did not return the paper survey, we sent half of these households (n = 1,860) instructions for accessing the survey by web and a $2 incentive. We received 1,046 "Mail-to-Web" completes. The total number of responses to the Mail 1.0/Mail-to-Web recruitment was 2,326 out of the original 5,000 households sampled, yielding a response rate of 46.5%. Upon receiving the completed Mail 1.0 responses, we found that the age variable was missing from the survey, so the resulting data was discarded. The Mail-to-Web data was not affected by this issue.

Mail 1.1 Sample

We used a second ABS mail sample ("Mail 1.1") to replace the faulty data from the Mail 1.0 survey. Excluding all households in the first ABS sample, we drew a new sample of 4,149 households. We mailed paper surveys to 4,149 households and received 822 completed surveys (19.8 percent response rate). To reduce the cost of the second mail survey, we lowered the initial incentive from $5 to $2; there was no additional incentive included with reminder materials.

Social Media Surveys

Next, we performed two rounds of social media data collection to supplement the number of current adult cannabis users in the ABS sample. For both rounds, we used paid social media ads to target participants and delivered incentives via Amazon gift cards. The ads did not reveal the subject matter of the survey. Because of our interest in policy analyses, we set quotas to recruit an approximately equal number of participants from states with recreational and medical cannabis legalization, medical cannabis only legalization, or neither type of legalization based on the effective dates of state recreational and medical cannabis laws in July of 2016.

Social Media Fraud Prevention

When developing the social media surveys, we included fraud prevention measures that were established in the literature at the time,[13,16,17]

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

Figure 1. Data collection methods

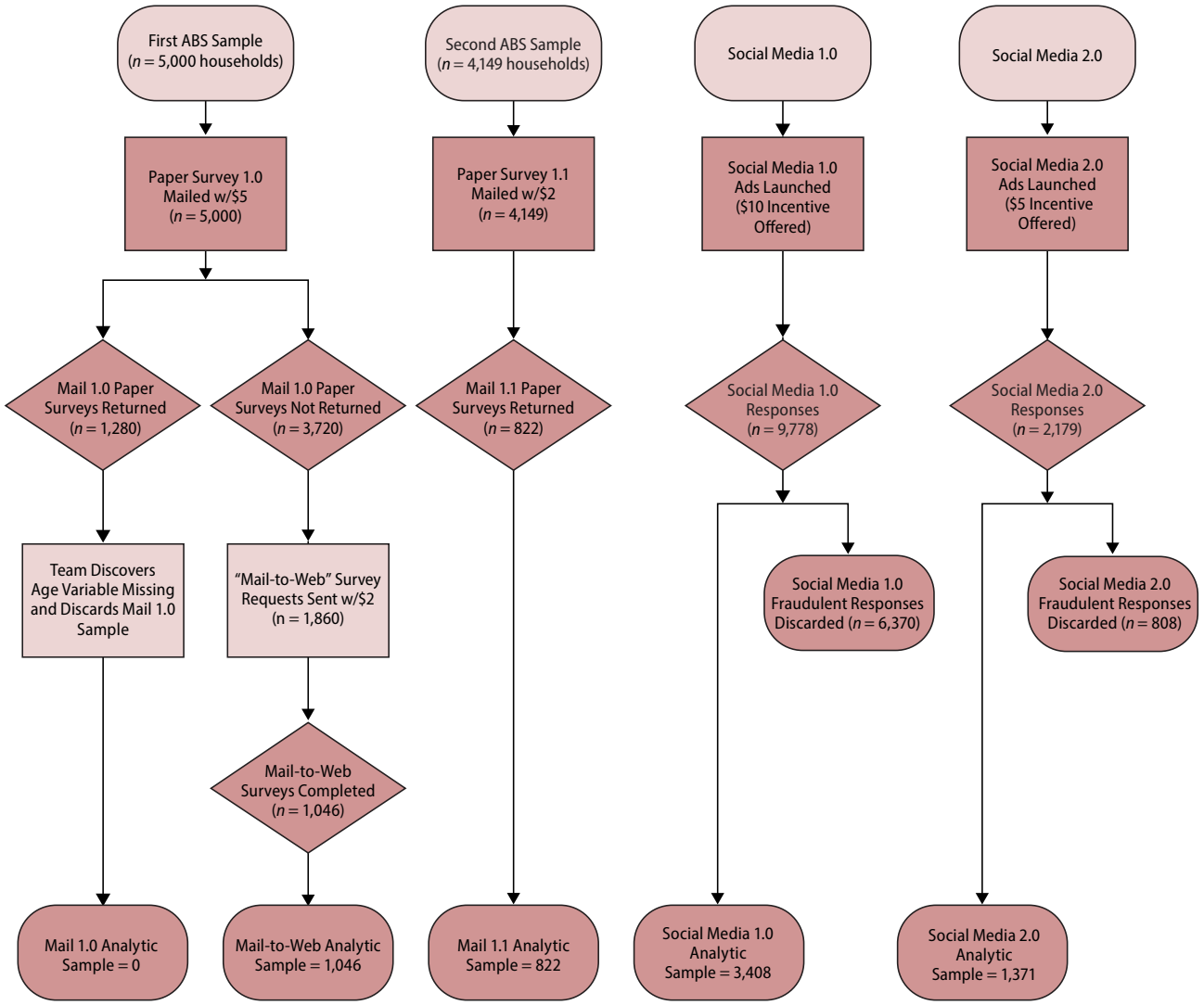


Table 1. Participant counts by sample source

Sample source	Invitations sent	Responses received	Analytic sample	Dates data received
First ABS Sample				
Mail 1.0	5,000	1,280	0[a]	8/18/16–9/29/16
Mail-to-Web	1,860[b]	1,046	1,046	8/18/16–9/29/16
Second ABS Sample				
Mail 1.1	4,149	822	822	5/22/17–6/14/17
Social Media Samples				
SM 1.0	N/A	9,778	3,408	8/18/16–9/29/16
SM 2.0	N/A	2,179	1,371	12/30/16–4/29/17

[a] A key variable was missing from the Mail 1.0 sample, so the returned paper surveys were discarded, yielding an analytic sample of 0.

[b] The 1,860 households that were sent Mail-to-Web instructions (and the 1,046 who responded) were part of the original sample of 5,000 households from the first ABS sample.

https://doi.org/10.3768/rtipress.2021.mr.0046.2109

including obscuring the purpose of the study through the use of distractor questions in the screener, collecting IP addresses, recording timestamps, instructing participants not to complete the survey multiple times (and noting that incentives would be withheld as a result), collecting e-mail address and state of residence, and asking questions assessing inclusion criteria in both the screener and body of the survey. Additional fraud prevention measures applied to the second social media sample are described in subsequent sections.

Social Media 1.0

For the first round of social media data collection (SM 1.0), we used paid advertisements on Facebook to recruit participants, and the incentive was $10. We received a large quantity of responses at odd hours (2:00 to 4:00 a.m. US Central time) from IP addresses outside of the United States and found evidence of link sharing on third-party websites. We collected 9,778 SM 1.0 responses.

Social Media 2.0

We conducted a second round of social media data collection (SM 2.0) to replace suspected fraudulent responses in SM 1.0. We used paid advertisements on Instagram only (to decrease the likelihood of overlap across the two social media samples) and targeted states with low completion rates for SM 1.0. We screened out participants who said that they had completed an RTI survey in the past 3 months to prevent individuals from completing both social media surveys. We collected 2,179 SM 2.0 responses.

Based on lessons learned from SM 1.0, SM 2.0 included additional fraud prevention measures, including a lower incentive ($5 Amazon gift card)[16] (Figure 2), screening out participants who missed attention checks and participants with mismatched state and zip code, refreshing the survey link daily to prevent link sharing, restricting access to the survey to daytime hours and IP addresses registered within the United States, and using the Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA), a tool that prevents automated (bot) completion of the survey.[14] CAPTCHA requires evidence of human presence to reach a website, often by requiring the user to select relevant photos from a compilation of images or to type text into a text box. In addition, participants who had Facebook accounts were required to authenticate using Facebook single sign-on, and we screened out participants who reported that they had learned about the study through any method other than "Facebook" or "Instagram."

Social Media Fraud Detection

After completing data collection, fraud detection methods for both samples included identifying duplicates using a combination of e-mail address,[16] timestamps, IP addresses, and identical responses. We also removed responses with 50 percent or more missing responses, excluded IP addresses outside of the United States and those known to be fraudulent or suspicious (using an online database), and excluded survey completions of 5 minutes or less (mean completion time was 20 minutes).

For SM 2.0, we also excluded respondents who answered "Facebook" or "Instagram" as their referral source but did not access the survey from either platform. After completing fraud detection procedures, 8,365 SM 1.0 responses (14 percent decrease in sample size) and 1,371 SM 2.0 responses (37 percent decrease in sample size) remained.

Fraud Model

Because (1) fraud detection methods only resulted in a small decrease in sample size for SM 1.0, (2) we found evidence that the survey's URL had been shared on social media, and through manual examination of the data, (3) we noticed patterns of unusual and contradictory responses in the data, we remained concerned about potential fraud in the sample. We created a fraud model (described in the Analysis section) to identify additional fraudulent responses.

Variables

Variables for Fraud Model

Outcome Variable

Fraud: To develop the fraud model, we first identified survey responses that distinguished between valid (Mail-to-Web) and invalid (SM 1.0 respondents with non-US IP addresses) respondents. Since participants were required to be US residents, we deemed SM 1.0

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

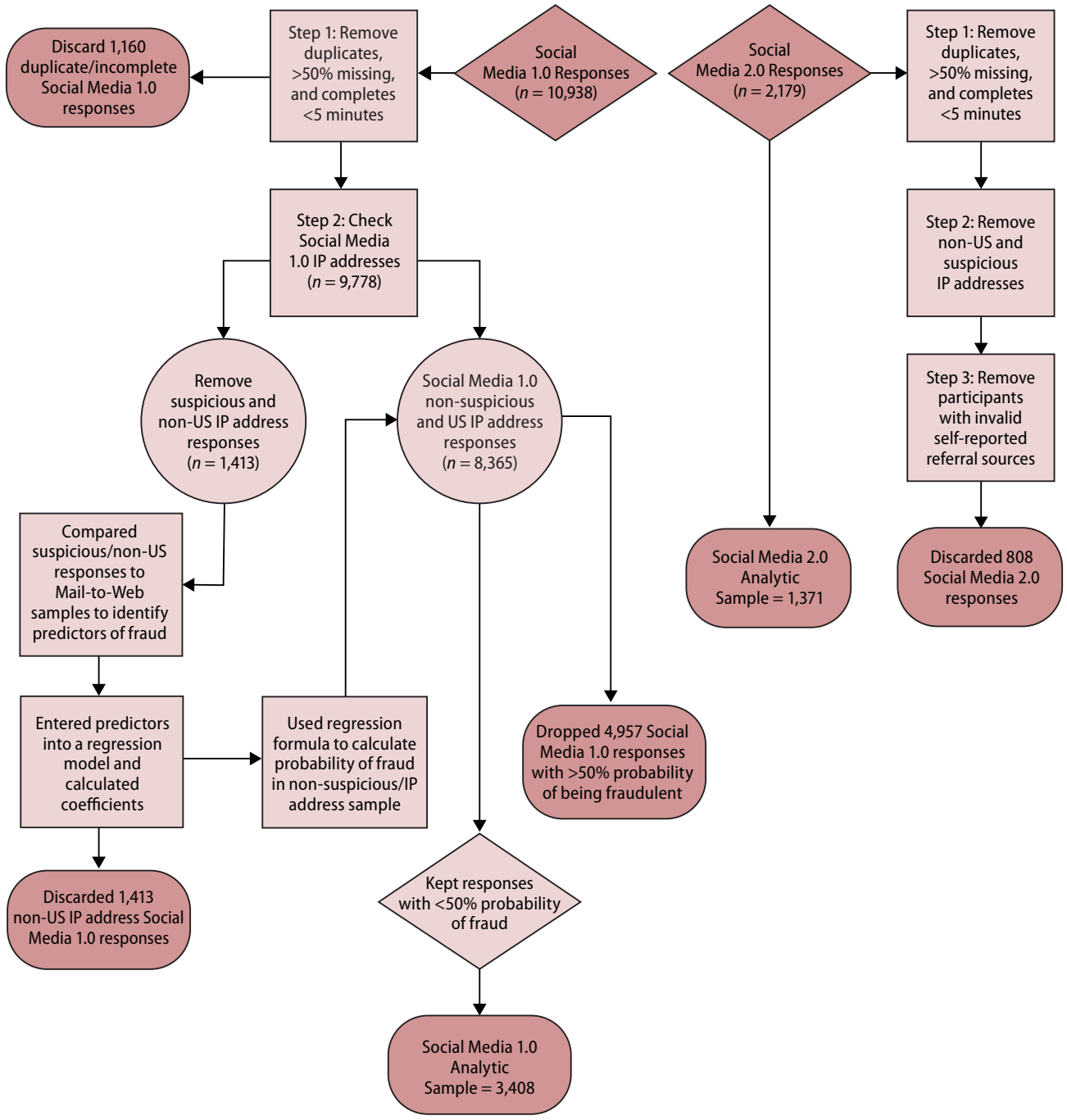
respondents with non-US IP addresses fraudulent (fraud = 1). Because ABS samples are highly reliable, we deemed the Mail-to-Web responses valid (fraud = 0).

Probability of fraud: Continuous probability of fraud was also the outcome of the fraud model.

Predictor Variables

To identify the predictor variables for the fraud model, we started with the full list of variables assessed by the survey, then excluded demographics (because these characteristics tend to vary by mode of data collection[14]) and questions used to estimate cannabis use prevalence (to avoid

Figure 2. Social media fraud prevention and detection measures



 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

biasing estimates used to validate the combined dataset). Using bivariate comparisons, we identified variables that distinguished between fraudulent and nonfraudulent responses. Then, we narrowed down this list to responses that met one or more relevant characteristics from Baker and Downes-LeGuin's list of suspicious survey responses[21]: selection of all responses for a multiple-choice question, selection of unlikely ("bogus") or low probability answers, internally inconsistent responses, and "straight lining" (selecting one answer for all items) in grids.[14] We excluded the following variables that did not meet any of these criteria: method of accessing the internet, social media use, mental health, and voting frequency. We also excluded variables with cell sizes smaller than 10 and/or variables for which 25 percent or fewer participants responded to the item because these items would result in model instability and/or a large number of missing responses for the model. Based on these criteria, we also excluded driving a car within three hours of getting high, usual method of obtaining cannabis, going to work within three hours of getting high, and using cannabis while at work.

The resulting variables included in the model were:

- **Military health insurance:** Using military, CHAMPUS, TriCare, or the VA insurance for **most** medical care (1) (as opposed to Medicare, Medicaid, Indian Health Service, other, none, or "don't know"; 0); this is a low probability response.[22]
- **Parent or guardian of a child (or children) of all ages:** Endorsing being a guardian of child(ren) ages 12 or younger, 13 to 17, and 18 to 21 (1) versus two or fewer of these options (0); this response represents selection of all items in a multiple-choice question.
- **Self-employed:** Endorsing self-employed (1) occupational status, as opposed to employed for wages, out of work, a homemaker, a student, retired, unable to work, or prefer not to answer (0); this is a low probability response.[23]
- **High while taking survey** (1), as opposed to not high (0); this is a low probability response.
- **Accessing survey through "a mailed letter someone gave to me"** (1), which was not possible (low probability answer). The other response options were feasible: via a mailed letter sent to my home, a Facebook ad or sponsored NewsFeed story, sent to me by Facebook or another way, or another way (0).
- **Types of tobacco used in the past 30 days:** A count of the number of products endorsed from the following: (1) cigarettes; (2) vapes; (3) cigars; (4) chewing tobacco, snuff, dip, or snus; (5) and hookah or waterpipe); this pattern reflects selection of all responses for a multiple-choice question.
- **Marijuana consumption modes in the past 30 days:** A count of the following products: (1) edible marijuana; (2) personal vaporizer, e-joint, or volcano to smoke dry marijuana plant matter (such as leaves, buds, or flower); (3) personal vaporizer, e-joint, or volcano to smoke marijuana as hash, hash water, hash oil, or marijuana concentrates (dabs); and (4) smoke a blunt (marijuana or hash in a cigar or blunt wrap); this response represents selection of all responses for a multiple-choice question.
- **Daily versus occasional cannabis use:** Rating daily cannabis use as better (1), safer (1), and more morally acceptable or correct (1) rather than vice versa (0 for all comparison groups), which are inconsistent responses.
- **Recreational versus medical use:** Rating recreational cannabis use as better (1), safer (1), and more morally acceptable or correct (1) than medical cannabis, which represent inconsistent responses.
- **Legal to drive high:** Reporting "yes" (1) (versus "no" or "don't know"; 0) to whether it is legal to drive after using marijuana in the participant's state; this is a low probability response.
- **Cannabis more harmful to society than alcohol:** Selecting marijuana (1) as more harmful to society than alcohol if widely available, as opposed to rating alcohol as more harmful, the two substances as equally harmful, or don't know (0). Based on the existing literature,[24] these responses represent low probability and/or inconsistent answers.

Weighting Variables

The following variables were used to create weights that calibrated the subsamples of the NCCS:

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

- **Gender** was defined as female, male, or other category.
- **Age** was self-reported number of years old.
- **Race/ethnicity** was coded as non-Hispanic white, non-Hispanic Black/African American, Hispanic, or non-Hispanic other race.
- **Education** was coded as never attended school or only kindergarten, grades 1–8, grades 9–11, grade 12 (high school graduate) or GED, some college but no degree, associates degree (AA, AS), college graduate (BA, BS), some graduate or professional school, or graduate or professional degree.
- **State cannabis legal status** was defined by participant's self-reported state of residence, according to the following categories: recreational and medical cannabis legal, medical cannabis only legal, or neither.
- **Political philosophy** response options included very conservative, somewhat conservative, moderate—neither liberal nor conservative, somewhat liberal, very liberal, or none of the above.
- **Internet access** was measured as reporting dial-up service, DSL service, cable modem service, fiber optic service, mobile broadband plan, satellite, or some other service (1) versus no internet service (0).
- **Social media use** was categorized as responding "yes" to the question, "Are you on social media, such as Facebook, Instagram or Twitter" (1) versus responding "no" (0).
- **Dwelling type:** This information was obtained from the CDS, and we defined the variable as apartment, multifamily, or high-rise building (1) versus a single-family home (0).
- **Rural postal delivery route:** This information was obtained from the CDS, and the variable was defined as a rural postal delivery route (1) versus all other types of delivery routes (0).

Validation Variables

The following variables were used to validate the sample and its estimates of cannabis use:

- **Ever cannabis use** was assessed by the question, "Have you ever, even once, used marijuana in any form?" We assigned participants who reported having ever used cannabis a value of 1 for this variable and all others a value of 0.
- **Current cannabis use** was defined as reporting last using marijuana "within the past 30 days" (1); all other participants were noncurrent users (0).

Analyses

All analyses were conducted in Stata 16.0 (https://www.stata.com/).

Fraud Model Development

First, we used chi-square analyses, *t* tests, and ANOVAs to identify differences in survey responses between the Mail to Web (fraud = 0) and SM 1.0 responses with non-US IP addresses (fraud = 1).[14] Next, we regressed fraud on our predictor variables. We used logistic regression (as opposed to further bivariate comparisons) because it enabled us to combine responses to multiple questions to produce a probability of fraud for each individual in the sample. The logistic regression model was

$$ln\left(\frac{p(fraud)}{1 - p(fraud)}\right) = \beta_0 + \sum_{i=1}^{m} BiXi,$$

where m is equal to the number of predictor variables in the model. We used the resulting model to obtain beta values for each of the predictor variables in the model. We refer to this equation as the fraud prediction formula. Once we had calculated the formula, we dropped the non-US IP address SM 1.0 responses from the sample.

The next step was to use the fraud prediction formula to identify additional fraudulent responses among the US IP address SM 1.0 responses. We used the formula to calculate the probability of fraud for these respondents by multiplying the value of each beta coefficient in the formula by each participant's value for X for all predictor variables in the model and summing these values to calculate y, which was equal to $ln\left(\frac{p(fraud)}{1 - p(fraud)}\right)$, for each respondent. After calculating y, we solved for *p(fraud),* which is the probability that each SM 1.0 response is fraudulent. We set a cutoff of 50 percent or greater probability of fraud for dropping participants from this sample.

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

Sensitivity Analyses

To ensure that 50 percent was the correct cutoff value, we conducted sensitivity analyses using values of 33, 50, and 66 percent or greater probability of fraud as cutoff values for SM 1.0 participants with US IP addresses. Using the same predictor variables included in the fraud prediction formula, we compared the characteristics of each of the samples obtained from the three cutoff values to the characteristics of the Mail-to-Web sample (valid) and non-US IP address SM 1.0 responses (fraudulent) to identify the best cutoff value.

Weighting

After choosing a cutoff value for fraud, we used weights to calibrate all of the NCCS subsamples to each other and the resulting pooled sample to the US population. The weighting procedures we used were a modified version of an existing approach applied to an Oregon cannabis survey.[25] Our weighting procedures also represent an updated and final version of the preliminary weighting scheme used on the NCCS before the fraud model was developed.[9] Generally, our approach involved a descriptive comparison of demographic and geographic characteristics and predictors of cannabis use across the probability and nonprobability samples, sample matching using the R MatchIt package, multiple propensity score models, comparing the demographics of the social media and ABS samples across these models, and comparing the prevalence of several measures of cannabis use and opinions for the NCCS and previous surveys.[9,25] The final weighting scheme was based on the differences that we observed between the mail and social media samples, the similarities we observed between the Mail-to-Web and social media samples, and the finding that political philosophy was a better predictor of attitudes toward cannabis use than cannabis use itself.[25] We also used the SUDAAN 11 (https://sudaansupport.rti.org/) WTADJX procedure for calibration.

Validation

After determining the cutoff for fraud, dropping all remaining fraudulent SM 1.0 responses, and weighting the pooled dataset, we validated the sample[9] by comparing NCCS estimates for ever and current cannabis use with similar estimates in the published literature,[9] specifically estimates obtained from the 2016 NSDUH,[7] the 2017 Yahoo News/Marist Poll,[26] and the 2016 Gallup Poll.[27] We attempted to locate social media or online surveys of cannabis use but were unable to locate any.

Results

Bivariate Results

For bivariate comparisons of SM 1.0 respondents with non-US IP addresses and Mail-to-Web respondents, non-US IP address SM 1.0 respondents were significantly more likely than Mail-to-Web respondents to report military health insurance, having children in all three age groups captured by the survey, being self-employed, being high while taking the survey, reporting receiving a mailed survey from someone else, number of types of tobacco used, number of modes of cannabis used, being more accepting of daily cannabis use than occasional use, being more accepting of recreational cannabis use than medical use, believing it is legal to drive high, and believing that cannabis is more harmful to society than alcohol ($P < 0.001$; Table 2).

Fraud Prediction Formula

Regressing the fraud variable on our predictor variables yielded the following fraud prediction formula:

$$ln\left(\frac{p(fraud)}{1 - p(fraud)}\right) = \beta_0 + 23.97(\text{Military insurance}) + 301.64(\text{Children of all ages}) + 14.37(\text{Self-employed}) + 1651.79(\text{High}) + 382.87(\text{Letter from someone else}) + 9.00(\text{Number of types of tobacco}) + 4.37(\text{Number of modes of cannabis}) + 20.37(\text{Daily cannabis use better than occasional}) + 15.38(\text{Daily cannabis use safer than occasional}) + 11.72(\text{Daily cannabis use more right than occasional}) + 36.58(\text{Recreational cannabis use better than medical}) + 7.37(\text{Medical cannabis use more dangerous than recreational}) + 14.17(\text{Recreational cannabis use more right than medical}) + 373.09\ (\text{Legal to drive high}) + 12.38(\text{Cannabis more harmful than alcohol})$$

Multiplying the values of X by the beta values from the above equation for each SM 1.0 participant with a US IP address and solving for *p(fraud)*, we identified 6,370 participants with a 50 percent or higher probability of fraud.

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

Table 2. Analysis of the characteristics of the Mail-to-Web and non-US IP address SM 1.0 responses in the National Cannabis Climate Survey

	Mail-to-Web sample (n = 1,045)[a]		Non-US IP Address (Fraudulent) SM 1.0 responses (n = 1,413)[b]	
	n	Mean	n	Mean
Military health insurance[c]	42	4.1%	712	50.4%
Having kids in all 3 age groups[d]	21	2.0%	1,216	86.1%
Self-employed	115	11.3%	913	64.6%
High while taking survey	18	1.7%	1,366	96.7%
Received mail survey from someone else	7	0.7%	1,020	72.2%
Polytobacco use[e]	.	0.25	.	4.92
Polycannabis use[f]	.	0.45	.	5.89
Discrepancy between opinions about daily/occasional use[g]	58	5.6%	565	40.0%
Discrepancy between opinions about recreational/medical use[h]	42	4.0%	430	30.4%
Legal to drive high	53	5.1%	1,345	95.3%
Cannabis is more harmful than alcohol	81	7.8%	674	51.1%

[a] There was a significant difference between the Mail-to-Web sample and fraudulent SM 1.0 completes for all variables ($P < 0.001$).

[b] Fraudulent SM 1.0 completes are participants who completed the survey and had IP addresses from outside of the United States.

[c] Participants who endorsed using military, CHAMPUS, TriCare, or the VA for **most** of their medical care.

[d] Participants who endorsed being the parent or guardian of a child (or children) ages 12 or younger, 13 to 17, and 18 to 21 (3 separate items).

[e] Number of the following products that the participant reported using in the past 30 days: (1) cigarettes; (2) vapes; (3) cigars; (4) chewing tobacco, snuff, dip, or snus; (5) and hookah or waterpipe.

[f] Number of the following products that the participant reported using in the past 30 days: (1) edible marijuana; (2) personal vaporizer, e-joint, or volcano to smoke dry marijuana plant matter; (3) personal vaporizer, e-joint, or volcano to smoke marijuana as hash, hash water, hash oil, or concentrates; and (4) blunt.

[g] Participant had unintuitive responses for rating daily versus occasional cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

[h] Participant had unintuitive responses for rating medical versus recreational cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

Sensitivity Analyses

The sensitivity analysis confirmed our use of the 50 percent cutoff. Using the 33 percent cutoff, the nonfraudulent sample significantly differed from the fraudulent sample for all variables and from the Mail-to-Web sample for eight variables (Table A.1 in the Appendix). Using the 50 and 66 percent cutoff values, the nonfraudulent sample significantly differed from the fraudulent sample for all variables and from the Mail-to-Web sample for five variables (Tables A.2 and A.3). Because the samples resulting from 50 and 66 percent cutoff values performed equally well in resembling the Mail-to-Web sample and differing from the fraudulent sample, and the 66 percent cutoff resulted in a much smaller sample size (2,650), we chose to use the 50 percent cutoff (3,408) to preserve statistical power.

Survey Weights

The weighting scheme incorporated six characteristics: gender, age, race/ ethnicity, education, cannabis legal status, and political philosophy.[28] First, we adjusted for differential nonresponse across sampling strata, dwelling type, and rural postal delivery route in the Mail 1.1 sample.[28] Then, we used the SUDAAN 11 WTADJX procedure to calibrate the samples to population estimates and each other.[25,28] We used gender, age, race/ethnicity, and education to calibrate Mail 1.1 respondents to continental US population totals from the American Community Survey (ACS).[29] We then used cannabis legal environment, age, education, gender, and political philosophy to calibrate the Mail-to-Web respondents to Mail 1.1 participants who reported having internet access and the SM 1.0 and SM 2.0 samples to Mail-to-Web participants who reported being on social media.

https://doi.org/10.3768/rtipress.2021.mr.0046.2109

We assumed that the weighted groups of Mail 1.1 respondents and Mail-to-Web respondents with social media, SM 1.0 respondents, and SM 2.0 respondents represented the same subpopulation and that the weighted groups of Mail 1.1 respondents with internet but without social media and Mail-to-Web respondents without social media represented the same subpopulation. We then computed effective cohort sample sizes for each of these groups (sample size divided by unequal weighting effect for each group). We combined respondents with internet but without social media (two groups) and respondents with social media (four groups), using effective cohort sample sizes for both combinations, resulting in one group that could be analyzed as the population of interest.[28]

Validating the Sample

To validate the sample, we compared weighted estimates for cannabis use in the NCCS to the results of other publicly available surveys of adults 18 and over in the United States[9] (Table 3).

NSDUH relies on a stratified, multistage area probability sample and is conducted via in-person interviews,[7] while the Yahoo! and Gallup surveys included random samples of landline and mobile phones and were conducted by phone.[26,27] Ever use was higher in the NCCS sample and subsamples compared with the other data sources, but these values approached those found in the Yahoo News survey. For current use, NCCS estimates fell between the estimates obtained from probability and nonprobability samples.

Discussion

This analysis used a fraud regression model, in combination with other fraud prevention and detection methods, to identify and eliminate probable fraudulent completes in a social media sample. This analysis also described the weighting and validation methods used for this study.

Several lessons can be gleaned from the data collection and described methods. The first is the importance of the prevention of fraud in social media, which has become increasingly common over time. We had few fraud issues when we included fraud prevention methods in our social media data collection (SM 2.0).

Table 3. Comparison of key variables for NCCS versus validated samples

Source	Ever cannabis use, % (SE)	Current cannabis use, % (SE)
NCCS combined sample[a]	58.3% (1.4)	17.0% (1.3)
NCCS ABS (Mail 1.1 and web)	56.1% (2.9)	15.9% (2.9)
NCCS SM 1.0 and 2.0 completes[a]	60.0% (1.3)	17.9% (1.0)
2016 NSDUH data[b]	47.0% (0.35)	10.9% (0.18)
2017 NSDUH data[b]	48.2% (0.36)	11.5% (0.19)
2017 Yahoo News/Marist Poll[c]	52% (NR)	22% (NR)
2016 Gallup Poll[d]	43% (NR)	13% (NR)

Abbreviations: NR = not reported.

Note: All NCCS estimates in this table are weighted according to the weights designed for the combined sample

[a] Social media responses deemed fraudulent (either during initial fraud detection procedures or through application of the fraud regression model) are not included in this table.

[b] Substance Abuse and Mental Health Services Administration (SAMHSA). Results from the 2017 National Survey on Durg Use and Health: detailed tables. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2018 [cited 2020 Oct 8]. Available from: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.pdf

[c] Marist Poll. Yahoo News/Marist Poll: weed & the American family. 2017 [cited 2020 Oct 2]. Available from: http://maristpoll.marist.edu/wp-content/misc/Yahoo%20News/20170417_Summary%20Yahoo%20News-Marist%20Poll_Weed%20and%20The%20American%20Family.pdf

[d] McCarthy J. One in eight US adults say they smoke marijuana. Gallup; 2016 Aug 8 [cited 2020 Oct 2]. Available from: https://news.gallup.com/poll/194195/adults-say-smoke-marijuana.aspx

The second lesson is the ability to use patterns of similarities and differences between fraudulent and nonfraudulent responses to clean datasets plagued by fraud. The predictive model of fraud described in this manuscript provides an advantage over bivariate analyses[14] by using information obtained from several variables to determine fraud, as opposed to examining the variables one at a time. Also, the model calculates fraud as a probability.

Our use of fraud prevention methods and validation increased our confidence in the quality and accuracy of the resulting dataset. The estimates obtained from the combined ABS and social media sample produced cannabis prevalence estimates similar to but higher than those of other surveys in the field at the time. Because of differences between the surveys, most notably in data collection methods, it is appropriate for the results from the NCCS to resemble, but not exactly match, those obtained from these other surveys.[31–34] NSDUH uses in-person interviews, and Yahoo! and Gallup used telephone surveys to collect data. Responses tend to differ by survey mode due to social desirability and varying perceptions of anonymity.[31] In fact, research suggests that substance users are unrepresented in samples obtained via data collection methods[35] such as interviews[36] and landline surveys.[37] In addition, NCCS did not use the same item as NSDUH to assess ever (lifetime) use of cannabis; NSDUH asks, "Have you ever, even once, used marijuana or hashish?"[38] Our use of quotas to sample participants from different cannabis legal environments likely also affected the prevalence of cannabis use in the study.

Limitations

This study has several limitations that should inform the interpretation of its results. The fraud model we created relied on several assumptions: (1) all non-US IP address SM 1.0 responses were fraudulent (fraud = 1) and all Mail-to-Web responses were not (fraud = 0); (2) all variables that differed significantly ($P < 0.05$) between non-US IP address SM 1.0 responses and Mail-to-Web responses could be used to predict fraudulent responses among SM 1.0 participants with US IP addresses; and (3) the likelihood of a nonfraudulent complete scoring a high probability of fraud was very low. It is possible that one or more of these assumptions is incorrect, but we based these procedures on extensive analyses of the data. Another limitation is that fraud prevention and detection procedures have improved greatly since the NCCS. However, widespread fraud still occurs, and methods of committing social media fraud are constantly evolving to adapt to improved measures of fraud prevention and detection. The process for developing a fraud model described in this manuscript can be applied to existing and future data collections despite changes in fraud and fraud prevention technology.

Conclusion

This paper outlines fraud prevention and detection measures that can be applied to future data collections. In addition, this manuscript outlines a fraud detection model that can be applied to existing social media datasets riddled with fraud. This manuscript also outlines methods of combining probability and nonprobability samples using weights. Overall, this analysis provides methods for resolving common issues encountered during and after data collection.

 https://doi.org/10.3768/rtipress.2021.mr.0046.2109

References

1. McGinty EE, Niederdeppe J, Heley K, Barry CL. Public perceptions of arguments supporting and opposing recreational marijuana legalization. Prev Med 2017;99:80–6. https://doi.org/10.1016/j.ypmed.2017.01.024
2. US Census Bureau. Table 1. Annual estimates of the resident population for the United States, regions, states, and Puerto Rico: April 1, 2010 to July 1, 2018 (NST-EST2018–01). 2018 Dec 19 [cited 2018 Apr 1]. Available from: https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html
3. National Conference of State Legislatures. State medical marijuana laws. 2019 Mar 5 [cited 2019 Apr 1]. Available from: https://www.ncsl.org/research/health/state-medical-marijuana-laws.aspx
4. National Conference of State Legislatures. Marijuana overview: legalization. 2018 Dec 14 [cited 2019 Apr 1]. Available from: https://www.ncsl.org/research/civil-and-criminal-justice/marijuana-overview.aspx
5. Substance Abuse and Mental Health Services Administration (SAMHSA). Key substance use and mental health indicators in the United States: results from the 2016 National Survey on Drug Use and Health (HHS Publication No. SMA 17–5044, NSDUH Series H-52). Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2017 [cited 2019 Mar 28]. Available from: https://www.samhsa.gov/data/
6. National Survey on Drug Use and Health. About the survey: project goals. About NSDUH. 2020 [cited 2020 Oct 2]. Available from: https://nsduhweb.rti.org/respweb/about_nsduh.html
7. Center for Behavioral Health Statistics and Quality. 2016 National Survey on Drug Use and Health: detailed tables. Substance Abuse and Mental Health Services Administration; 2017 [cited 2020 Oct 2]. Available from: https://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs-2016/NSDUH-DetTabs-2016.pdf
8. Centers for Disease Control and Prevention. Data hosting: National Survey on Drug Use and Health (NSDUH). Other Restricted Data. 2020 Aug 11 [cited 2020 Oct 2]. Available from: https://www.cdc.gov/rdc/b1datatype/nsduh.htm
9. Dever J. Combining probability and nonprobability samples to form efficient hybrid estimates: an evaluation of the common support assumption. Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference. 2018 [cited 2018 Oct 1]. Available from: https://copafs.org/wp-content/uploads/2020/05/COPAFS-A4_Dever_2018FCSM.pdf
10. Poushter J. Not everyone in advanced economies is using social media. FACTANK News in the Numbers; 2017 Apr 20 [cited 2021 Feb 12]. Available from: https://www.pewresearch.org/fact-tank/2017/04/20/not-everyone-in-advanced-economies-is-using-social-media/
11. Hruska J, Maresova P. Use of social media platforms among adults in the United States—behavior on social media. Societies (Basel) 2020;10(1):27. https://doi.org/10.3390/soc10010027
12. Konstan JA, Simon Rosser BR, Ross MW, Stanton J, Edwards WM. The story of subject naught: a cautionary but optimistic tale of internet survey research. J Comput Mediat Commun 2005;10(2):00. https://doi.org/10.1111/j.1083-6101.2005.tb00248.x
13. Grey JA, Konstan J, Iantaffi A, Wilkerson JM, Galos D, Rosser BR. 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):1928–37. https://doi.org/10.1007/s10461-015-1033-y
14. Dewitt J, Capistrant B, Kohli N, Rosser BR, Mitteldorf D, Merengwa E Addressing participant validity in a small internet health survey (The Restore Study): protocol and recommendations for survey response validation. JMIR Res Protoc 2018;7(4):e96. https://doi.org/10.2196/resprot.7655
15. Tourangeau R. Defining hard-to-survey populations. In: Tourangeau R, Edwards B, Johnson TP, Wolter KM, Bates N, editors. Hard-to-survey populations. Cambridge, UK: Cambridge University Press; 2014. pp. 3–20. https://doi.org/10.1017/CBO9781139381635.003
16. Bowen AM, Daniel CM, Williams ML, Baird GL. Identifying multiple submissions in Internet research: preserving data integrity. AIDS Behav 2008;12(6):964–73. https://doi.org/10.1007/s10461-007-9352-2

17. Nosek BA, Banaji MR, Greenwald AG. E- research: ethics, security, design, and control in psychological research on the Internet. J Soc Issues 2002;58(1):161–76. https://doi.org/10.1111/1540-4560.00254

18. Mustanski BS. Getting wired: exploiting the Internet for the collection of valid sexuality data. J Sex Res 2001;38(4):292–301. https://doi.org/10.1080/00224490109552100

19. Baker R, Downes-Le Guin T. Separating the wheat from the chaff: ensuring data quality in internet samples. In: Trotman M, editor. Proceedings of the fifth international conference of the Association for Survey Computing: the challenges of a changing world. Association for Survey Computing; 2007. p. 157–166.

20. American Association for Public Opinion Research. Address-based sampling. Prepared for the AAPOR Council by the Task Force on Address-based Sampling; 2016. Available from: https://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-(2).pdf

21. Hilbert D, Redmiles D. Separating the wheat from the chaff in Internet-mediated user feedback expectation-driven event monitoring. ACM SIGGROUP Bulletin 1999;20(1):35–40. https://doi.org/10.1145/327556.327611

22. Berchick E, Barnett J, Upton R. Health insurance coverage in the United States: 2018. Current Population Reports P60–267(RV). US Department of Commerce, US Census Bureau; 2019 [cited 2020 Oct 2]. Available from: https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-267.pdf

23. Hipple S, Hammond L. Self-employment in the United States. Bureau of Labor Statistics; 2016 [cited 2016 Mar]. Available from: https://www.bls.gov/spotlight/2016/self-employment-in-the-united-states/pdf/self-employment-in-the-united-states.pdf

24. Allen JA, Davis KC, Duke JC, Nonnemaker JM, Bradfield BR, Farrelly MC. New product trial, use of edibles, and unexpected highs among marijuana and hashish users in Colorado. Drug Alcohol Depend 2017;176:44–7. https://doi.org/10.1016/j.drugalcdep.2017.03.006

25. Kott PS. A partially successful attempt to integrate a web-recruited cohort into an address-based sample. Surv Res Methods 2019;13(1).

26. Marist Poll. Yahoo News/Marist Poll: weed & the American family. 2017 [cited 2020 Oct 2]. Available from: http://maristpoll.marist.edu/wp-content/misc/Yahoo%20News/20170417_Summary%20Yahoo%20News-Marist%20Poll_Weed%20and%20The%20American%20Family.pdf

27. McCarthy J. One in eight US adults say they smoke marijuana. Gallup; 2016 Aug 8 [cited 2020 Oct 2]. Available from: https://news.gallup.com/poll/194195/adults-say-smoke-marijuana.aspx

28. Ridenhour J, Kott P. Using calibration weighting in samples with non-probability components. Proceedings of the Joint Statistical Meetings. 2018 [cited 2020 Oct 1]. Available from: https://ww2.amstat.org/meetings/jsm/2018/onlineprogram/ActivityDetails.cfm?SessionID=215233

29. United States Census Bureau. American Community Survey (ACS). 2021. Available from: https://www.census.gov/programs-surveys/acs

30. Center for Behavioral Health Statistics and Quality. Results from the 2017 National Survey on Drug Use and Health: detailed tables. Substance Abuse and Mental Health Services Administration; 2018 [cited 2020 Oct 8]. Available from: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.pdf

31. Supple AJ, Aquilino WS, Wright DL. Collecting sensitive self-report data with laptop computers: impact on the response tendencies of adolescents in a home interview. J Res Adolesc 1999;9(4):467–88. https://doi.org/10.1207/s15327795jra0904_5

32. Dillman DA. Why choice of survey mode makes a difference. Pub Health Rep 2006;121(1):11–13. https://doi.org/10.1177/003335490612100106

33. Keeter S. From telephone to web: the challenge of mode of interview effects in public opinion polls. 2015 May 13 [cited 2021 Aug 2]. Available from: https://www.pewresearch.org/methods/2015/05/13/from-telephone-to-the-web-the-challenge-of-mode-of-interview-effects-in-public-opinion-polls/

34. Bowyer B, Rogowski J. Mode matters: evaluating response comparability in a mixed-mode survey. Political Sci Res Methods 2017;5(2):295–313. https://doi.org/10.1017/psrm.2015.28

35. Johnson TP. Sources of error in substance use prevalence surveys. Int Sch Res Notices 2014;2014:923290. https://doi.org/10.1155/2014/923290

36. Lyons Reardon ML, Burns AB, Preist R, Sachs-Ericsson N, Lang AR. Alcohol use and other psychiatric disorders in the formerly homeless and never homeless: prevalence, age of onset, comorbidity, temporal sequencing, and service utilization. Subst Use Misuse 2003;38(3-6):601–44. https://doi.org/10.1081/JA-120017387

37. Delnevo CD, Gundersen DA, Hagman BT. Declining estimated prevalence of alcohol drinking and smoking among young adults nationally: artifacts of sample undercoverage? Am J Epidemiol 2008;167(1):15–9. https://doi.org/10.1093/aje/kwm313

38. Center for Behavioral Health Statistics and Quality. 2016 National Survey on Drug Use and Health (NSDUH): final approved CAI specifications for programming (English version). Substance Abuse and Mental Health Services Administration; 2015; Available from: https://www.samhsa.gov/data/sites/default/files/NSDUHmrbCAIquex2016v2.pdf

Appendix. Supplementary Tables

Table A.1. Comparison of the characteristics of the mail-to-web and fraudulent (non-US IP address) and nonfraudulent (US IP address and 33 percent cutoff applied from fraud model) Social Media 1.0 responses in the National Cannabis Climate Survey

	Mail-to-web (n = 1,045)[a]		Fraudulent Social Media 1.0 responses (n = 1,413)[b]		Nonfraudulent Social Media 1.0 responses based on 33% cutoff (n = 4,866)[c]		p-value for mail-to-web versus nonfraudulent sample
	n	Mean	n	Mean	n	Mean	
Military health insurance[d]	42	4.1%	712	50.4%	147	3.1%	0.1273
Having kids in all 3 age groups[e]	21	2.0%	1,216	86.1%	123	2.5%	0.2908
Self-employed	115	11.3%	913	64.6%	502	10.5%	0.4482
High while taking survey	18	1.7%	1,366	96.7%	523	10.8%	0.0000
Received mail survey from someone else	7	0.7%	1,020	72.2%	103	2.2%	0.0000
Current tobacco products used[f]	.	0.25	.	4.92	.	1.55	0.0000
Current modes of cannabis used[g]	.	0.45	.	5.89	.	2.12	0.0000
Discrepancy between opinions about daily/occasional use[h]	58	5.6%	565	40.0%	888	18.3%	0.0000
Discrepancy between opinions about recreational/medical use[i]	42	4.0%	430	30.4%	557	11.4%	0.0000
Legal to drive high	53	5.1%	1,345	95.3%	596	12.3%	0.0000
Cannabis is more harmful than alcohol	81	7.8%	674	51.1%	680	14.1%	0.0000

[a] There was a significant difference between the mail-to-web sample and fraudulent Social Media 1.0 completes for all variables ($P < 0.001$).

[b] Fraudulent Social Media 1.0 completes are participants who completed the survey and had IP addresses from outside of the United States.

[c] Nonfraudulent Social Media 1.0 completes are participants with US IP addresses who remained in the sample after eliminating all participants with a 33% or greater probability of being valid based on the fraud regression model. There was a significant difference between the fraudulent and the nonfraudulent Social Media 1.0 samples for all variables in the table ($P < 0.001$).

[d] Participants who endorsed using military, CHAMPUS, TriCare, or the VA for **most** of their medical care.

[e] Participants who endorsed being the parent or guardian of a child (or children) ages 12 or younger, 13 to 17, and 18 to 21 (3 separate items).

[f] Number of the following products that the participant reported using in the past 30 days: (1) cigarettes; (2) vapes; (3) cigars; (4) chewing tobacco, snuff, dip, or snus; (5) and hookah or waterpipe.

[g] Number of the following products that the participant reported using in the past 30 days: (1) edible marijuana; (2) personal vaporizer, e-joint, or volcano to smoke dry marijuana plant matter; (3) personal vaporizer, e-joint, or volcano to smoke marijuana as hash, hash water, hash oil, or concentrates; and (4) blunt.

[h] Participant had unintuitive responses for rating daily versus occasional cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

[i] Participant had unintuitive responses for rating medical versus recreational cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

https://doi.org/10.3768/rtipress.2021.mr.0046.2109

Table A.2. Comparison of the characteristics of the Mail-to-Web and fraudulent (non-US IP address) and nonfraudulent (US IP address and 50 percent cutoff applied from fraud model) Social Media 1.0 responses in the National Cannabis Climate Survey

	Mail-to-web sample (n = 1,045)[a]		Fraudulent Social Media 1.0 responses (n = 1,413)[b]		Nonfraudulent Social Media 1.0 responses based on 50% cutoff (n = 3,408)[c]		p-value for mail-to-web versus nonfraudulent samples
	n	Mean	n	Mean	n	Mean	
Military health insurance[d]	42	4.1%	712	50.4%	98	2.9%	0.0793
Having kids in all 3 age groups[e]	21	2.0%	1,216	86.1%	65	1.9%	0.8270
Self-employed	115	11.3%	913	64.6%	312	9.2%	0.0556
High while taking survey	18	1.7%	1,366	96.7%	119	3.5%*	0.0006
Received mail survey from someone else	7	0.7%	1,020	72.2%	19	0.6%	0.6797
Current tobacco products used[f]	.	0.25	.	4.92	.	0.59*	0.0000
Current modes of cannabis use[g]	.	0.45	.	5.89	.	1.01*	0.0000
Discrepancy between opinions about daily/occasional use[h]	58	5.6%	565	40.0%	442	13.0%*	0.0000
Discrepancy between opinions about recreational/medical use[i]	42	4.0%	430	30.4%	204	6.0%*	0.0072
Legal to drive high	53	5.1%	1,345	95.3%	152	4.5%	0.4014
Cannabis is more harmful than alcohol	81	7.8%	674	51.1%	226	6.6%	0.2130

[a] There was a significant difference between the mail-to-web sample and fraudulent Social Media 1.0 completes for all variables ($P < 0.001$).

[b] Fraudulent Social Media 1.0 completes are participants who completed the survey and had IP addresses from outside of the United States.

[c] Nonfraudulent Social Media 1.0 completes are participants with US IP addresses who remained in the sample after eliminating all participants with a 50% or greater probability of being valid based on the fraud regression model. There was a significant difference between the fraudulent and nonfraudulent Social Media 1.0 samples for all variables in the table ($P < 0.001$).

[d] Participants who endorsed using military, CHAMPUS, TriCare, or the VA for **most** of their medical care.

[e] Participants who endorsed being the parent or guardian of a child (or children) ages 12 or younger, 13 to 17, and 18 to 21 (3 separate items).

[f] Number of the following products that the participant reported using in the past 30 days: (1) cigarettes; (2) vapes; (3) cigars; (4) chewing tobacco, snuff, dip, or snus; (5) and hookah or waterpipe.

[g] Number of the following products that the participant reported using in the past 30 days: (1) edible marijuana; (2) personal vaporizer, e-joint, or volcano to smoke dry marijuana plant matter; (3) personal vaporizer, e-joint, or volcano to smoke marijuana as hash, hash water, hash oil, or concentrates; and (4) blunt.

[h] Participant had unintuitive responses for rating daily versus occasional cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

[i] Participant had unintuitive responses for rating medical versus recreational cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

https://doi.org/10.3768/rtipress.2021.mr.0046.2109

Table A.3. Comparison of the characteristics of the Mail-to-Web and fraudulent (non-US IP address) and nonfraudulent (US IP address with 66 percent cutoff applied from fraud model) Social Media 1.0 responses in the National Cannabis Climate Survey

	Mail-to-web sample (n = 1,045)[a]		Fraudulent Social Media 1.0 responses (n = 1,413)[b]		Nonfraudulent Social Media 1.0 responses based on 66% cutoff (n = 2,650)[c]		p-value for mail versus nonfraudulent responses
	n	Mean	n	Mean	n	Mean	
Military health insurance[d]	42	4.1%	712	50.4%	68	2.6%	0.0291
Having kids in all 3 age groups[e]	21	2.0%	1,216	86.1%	31	1.2%	0.0801
Self-employed	115	11.3%	913	64.6%	227	8.6%	0.0166
High while taking survey	18	1.7%	1,366	96.7%	35	1.3%	0.3758
Received mail survey from someone else	7	0.7%	1,020	72.2%	6	0.2%	0.0975
Current tobacco products used[f]	.	0.25	.	4.92	.	0.27	0.2650
Current modes of cannabis use[g]	.	0.45	.	5.89	.	0.69	0.0000
Discrepancy between opinions about daily/occasional use[h]	58	5.6%	565	40.0%	318	12.0%	0.0000
Discrepancy between opinions about recreational/medical use[i]	42	4.0%	430	30.4%	116	4.4%	0.6219
Legal to drive high	53	5.1%	1,345	95.3%	116	4.4%	0.3569
Cannabis is more harmful than alcohol	81	7.8%	674	51.1%	82	3.1%	0.0000

[a] There was a significant difference between the mail-to-web sample and fraudulent Social Media 1.0 completes for all variables ($P < 0.001$).

[b] Fraudulent Social Media 1.0 completes are participants who completed the survey and had IP addresses from outside of the United States.

[c] Nonfraudulent Social Media 1.0 completes are participants with US IP addresses who remained in the sample after eliminating all participants with a 33% or greater probability of being valid based on the fraud regression model. There was a significant difference between the fraudulent and the nonfraudulent Social Media 1.0 samples for all variables in the table ($P < 0.001$).

[d] Participants who endorsed using military, CHAMPUS, TriCare, or the VA for **most** of their medical care.

[e] Participants who endorsed being the parent or guardian of a child (or children) ages 12 or younger, 13 to 17, and 18 to 21 (3 separate items).

[f] Number of the following products that the participant reported using in the past 30 days: (1) cigarettes; (2) vapes; (3) cigars; (4) chewing tobacco, snuff, dip, or snus; (5) and hookah or waterpipe.

[g] Number of the following products that the participant reported using in the past 30 days: (1) edible marijuana; (2) personal vaporizer, e-joint, or volcano to smoke dry marijuana plant matter; (3) personal vaporizer, e-joint, or volcano to smoke marijuana as hash, hash water, hash oil, or concentrates; and (4) blunt.

[h] Participant had unintuitive responses for rating daily versus occasional cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

[i] Participant had unintuitive responses for rating medical versus recreational cannabis use on one or more of the following scales: good/bad, dangerous/safe, or wrong/right.

https://doi.org/10.3768/rtipress.2021.mr.0046.2109

www.rti.org/rtipress **RTI Press publication MR-0046-2109**