Inferring Smoking Status from User Generated Content in an Online Cessation Community

Michael S. Amato PhD, George D. Papandonatos PhD, Sarah Cha MSPH, Xi Wang, PhD, Kang Zhao PhD, Amy M. Cohn PhD, Jennifer L. Pearson PhD, MPH, Amanda L. Graham PhD

1The Schroeder Institute for Tobacco Research and Policy Studies at Truth Initiative, Washington, DC; 2Center for Statistical Sciences, Brown University, Providence, RI; 3School of Information, Central University of Finance and Economics, Beijing, China; 4Department of Management Sciences, The University of Iowa, Iowa City, Iowa; 5 Battelle Memorial Institute, Arlington, VA; 6Department of Oncology, Georgetown University Medical Center, Washington, DC; 7School of Community Health Sciences, University of Nevada, Reno, NV; 8Department of Health, Behavior and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Corresponding Author: Michael S. Amato, PhD, The Schroeder Institute for Tobacco Research and Policy Studies at Truth Initiative, 900 G Street NW, Fourth Floor, Washington, DC 20001, USA. Telephone: 202-454-5767; Fax: 202-454-5785; E-mail: mamato@truthinitiative.org

Abstract

Introduction: User generated content (UGC) is a valuable but underutilized source of information about individuals who participate in online cessation interventions. This study represents a first effort to passively detect smoking status among members of an online cessation program using UGC.

Methods: Secondary data analysis was performed on data from 826 participants in a web-based smoking cessation randomized trial that included an online community. Domain experts from the online community reviewed each post and comment written by participants and attempted to infer the author’s smoking status at the time it was written. Inferences from UGC were validated by comparison with self-reported 30-day point prevalence abstinence (PPA). Following validation, the impact of this method was evaluated across all individuals and time points in the study period.

Results: Of the 826 participants in the analytic sample, 719 had written at least one post from which content inference was possible. Among participants for whom unambiguous smoking status was inferred during the 30 days preceding their 3-month follow-up survey, concordance with self-report was almost perfect (kappa = 0.94). Posts indicating abstinence tended to be written shortly after enrollment (median = 14 days).

Conclusions: Passive inference of smoking status from UGC in online cessation communities is possible and highly reliable for smokers who actively produce content. These results lay the groundwork for further development of observational research tools and intervention innovations.

Implications: A proof-of-concept methodology for inferring smoking status from user generated content in online cessation communities is presented and validated. Content inference of smoking status makes a key cessation variable available for use in observational designs. This method provides a powerful tool for researchers interested in online cessation interventions and establishes a foundation for larger scale application via machine learning.
Introduction

User generated content (UGC) is a valuable but underutilized source of information about individuals who participate in online behavioral interventions. Web-based smoking cessation interventions are used by thousands of smokers each year and yield quit rates comparable to other common forms of cessation interventions. Further improving the effectiveness of web-based cessation interventions by leveraging data from UGC could help reduce smoking rates at the population level.

Previous studies of web-based health interventions have drawn on UGC to identify common topics, themes, and types of social support discussed by participants. Members of web-based cessation communities typically exchange greetings, share experiences, and provide emotional and informational support. Similar behaviors have been observed in web-based communities for individuals dealing with problem drinking or eating disorders, individuals living with HIV/AIDS or Huntington's Disease, and in cancer survivor networks. Specific to web-based cessation communities, relapse and struggling with cravings are highly prevalent themes.

Analysis of UGC has also been used to investigate motivational states and processes of participants in web-based cessation communities. Johnsen et al. performed linguistic analysis on a sample of UGC written by web-based cessation program participants, and found that a prevention focus (ie, avoiding relapse) was more common than a promotion focus (ie, achieving abstinence). Of interest for the current study, the Johnsen study developed a method for directly measuring and quantifying a cessation behavior change construct from passively collected UGC. Although that construct was measured at the aggregate network level, their method could have been applied at the individual level as well. Previous studies have similarly measured individual level variables from passively collected UGC, for example measuring amounts of insightful disclosure in a web-based community for breast cancer survivors.

This study represents a first effort to measure the time course of changes in smoking status for members of a web-based cessation community, relying exclusively on passively collected UGC. This is a novel use of UGC in the study of web-based cessation interventions. The ability to reliably measure smoking status from UGC in online cessation communities could provide a valuable tool for both intervention development and evaluation. For example, intervention designers could make use of such an ability to dynamically tailor interventions to an individual’s specific needs, and evaluators could use UGC as supplementary data when imputing outcomes for individuals lost to follow-up.

Our aims were to: (1) investigate the validity of inferring smoking status from UGC through validation with self-reported smoking status obtained in a randomized clinical trial, and (2) estimate the potential impact of such inference in terms of the proportion of online cessation community members for whom reliable inference may be possible.

Methods

Setting

The study was conducted with users of BecomeAnEX.org, a publicly available web-based cessation program. Launched in 2008, the site was developed in collaboration with the Mayo Clinic Nicotine Dependence Center in accordance with national treatment guidelines. A national mass media campaign and ongoing online advertising have resulted in over 800,000 registrants since its inception. To register on BecomeAnEX, individuals must agree to the site’s Terms of Use and Privacy Policy. The Privacy Policy states that (1) BecomeAnEX collects information about users and their use of the site; (2) Information is used for research and quality improvement purposes only; and (3) Personal information is kept confidential. Thus, de-identified data from all registered users is available for analysis. BecomeAnEX teaches problem-solving and coping skills to quit smoking, educates users about cessation medications, and facilitates social support through a large online social network. The social network is comprised of thousands of current and former smokers who interact via several asynchronous communication channels (eg, blogs, group discussions, private messages). All user actions are date- and time-stamped and stored in a relational database.

Participants

This study investigates UGC produced by a subsample of participants from a randomized control trial (RCT) conducted on BecomeAnEX (ClinicalTrials.gov NCT01544153). The study protocol for the randomized trial was reviewed and approved by Western Institutional Review Board (protocol #20110877). The trial protocol and characteristics of the trial sample have been published elsewhere. The trial was conducted between March 2012 and November 2015. Briefly, participants in the RCT were 3290 new registrants who were randomized to one of four treatment arms in a 2 × 2 design that crossed (1) an enhanced social network integration protocol designed to integrate study participants into the BecomeAnEX social network with (2) free nicotine replacement therapy. All participants had access to the full website and community. Self-reported smoking status was collected 3 months after enrollment (response rate = 62.3%). At least one post or comment was contributed by 1180 study participants (22.3%) during the study period. Given our focus on developing a method to link UGC with abstinence, we elected to combine all treatment arms. The analytic sample was comprised of n = 826 participants who completed the 3-month follow-up survey and wrote at least one post or comment between study enrollment and 30 days after survey completion.

Measures

Self-reported Smoking Status

Self-reported smoking status was assessed 3 months after study enrollment using a standard measure of 30-day abstinence: "In the past 30 days, have you smoked any cigarettes at all, even a puff?" Those who had not smoked were considered abstinent.

Content-Inferred Smoking Status

Five former smokers who are active, long-standing members of the BecomeAnEX Community were recruited to serve as Domain Experts for the study. Domain Experts attempted content inference of smoking status for every post in the BecomeAnEX community written by participants between study enrollment and 30 days after they completed their 3-month follow-up survey. Each post was individually coded by two domain experts. A study team member overseeing the annotations served as a tiebreaker for any posts where the two original coders disagreed. Coding was guided by two questions. In the first question (Q1), each post was coded for the author’s smoking status as of the moment it was written (“What is the poster’s status at the time they wrote this post?”). Available codes were “Clearly smoking,” “Clearly not smoking,” or “Unclear.” Domain experts were instructed to use inference and make their best guess based the text and subtext of each post, but to use the “unclear” code whenever they did not feel confident that a reliable judgment could be made. In the second question...
(Q2), every post coded as “Clearly not smoking” was also coded for the number of days that the author claimed to have been abstinent (“If answer to Q1 is ‘clearly not smoking’ is the length of their quit obvious? If yes, write in length of quit using terms used by the poster.”) Duration estimates (i.e., length of quit) relied on a convention among BecomeAnEX members to publicly state the number of days they have been smoke free. Domain experts were instructed to code Q2 as “No” if they did not feel that a confident duration estimate could be made. If a post indicated that a participant had quit on the same day, that was coded as “1 day.” See Table 1 for example posts.

Because not all participants completed their follow-up survey exactly 90 days after their enrollment date, the date of each participant’s posts was centered on the date of their follow-up survey; the day of self-report for all participants was “Day 0” (T0), with most participants enrolling at approximately “Day −90”.

Statistical Analyses

Analysis 1: Comparison to Self-reported 30-day Abstinence

Analysis 1 sought to establish the validity of using UGC to infer smoking status by comparing the results of domain expert coding with self-reported 30-day abstinence. These analyses focused on a subset of content that was written during the 30 days immediately preceding the date that each participant completed their 3-month follow-up survey (i.e., Day −30 through Day 0). This range was chosen because it was the same time period participants were asked to consider for the self-report item assessing 30-day abstinence.

For each participant, an individualized smoking status timeline (30 days long, spanning Day −30 through Day 0) was constructed using the following rules: (1) All days on the timeline were initialized to a default state of “unclear.” (2) For each post with a content-inferred status of “smoking,” the corresponding day on the timeline was set to “smoking.” (3) For each post with a content-inferred status of “abstinent,” the corresponding day on the timeline was set to “abstinent” as was (4) the inferred number of prior days of abstinence. For example, if a participant wrote the following ten days before their self-report date, “11 days smoke free and still going strong,” their personal timeline would indicate abstinence on Days −21 through Day −10. If two posts yielded conflicting inferences for a single date on a participant’s personal timeline, the inference for that date was set to “unclear.”

Based on their personal timelines, participants were then categorized into three mutually exclusive groups: (1) abstinence inferred for all 30 days, (2) smoking inferred for any days, and (3) abstinence inferred for some days, but not all, with zero smoking days inferred. For groups 1 and 2, content inference suggested an unambiguous smoking status (abstinent and smoking, respectively). Content-inferred 30-day abstinence was compared with self-reported 30-day abstinence using Cohen’s Kappa to assess the validity of the content-inference methodology. Kappa coefficients have their range constrained by differences in prevalence between the dichotomous measures under investigation, and caution should be exercised in their interpretation when the associated sign test is significant. In the absence of prevalence differences, standard cutoffs for measuring agreement have been established by Landis and Koch, which rate them as follows: 0.80–1.00 = Almost Perfect, 0.60–0.80 = Substantial, 0.40–0.60 = Moderate, 0.20–0.40 = Fair, 0.00–0.20 = Slight, and <0.00 = Poor.

For participants in group 3, for whom content inference yielded ambiguous classification, we conducted an exploratory analysis to discover the most accurate classification scheme. Participants in group 3 were classified as abstinent or smoking based on the number of days for which abstinence was inferred. For this exploratory analysis, we investigated threshold number of days that maximized concordance with self-reported smoking status.

Analysis 2: Impact Across All Time Points and Participants

Building on the results of analysis 1, analysis 2 investigated the proportion of the analytic subsample for whom inference was possible. The analysis included all time points, from study enrollment (typically Day −90) through Day 30. Among all 826 participants, we determined the number for whom each of the following could be identified: (1) one or more days of smoking, (2) one or more days of abstinence, (3) slips, defined as a period of smoking after a period of abstinence.

Results

Personal timelines for all participants who self-reported as “abstinent” on their 30-day follow-up survey are shown in Figure 1. Personal timelines for participants who self-reported as “smoking” on their 30-day follow-up survey are shown in Figure 2. Each row in the figures represents the personal timeline of a single participant, while time is shown on the x-axis. As can be seen from the figures, most UGC was created shortly after enrollment. Relatively few participants (n = 95, 11.5% of the analytic sample) created content during the 30 days preceding the date of their 3-month follow-up survey. That 30-day period, which is the time period participants were asked to consider for the survey question assessing 30-day abstinence, is represented in the figures by the yellow band.

Table 1. Example Posts

<table>
<thead>
<tr>
<th>Inferred Status</th>
<th>Post Content</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>i am new at this as well my quit date is 11/11/12...i could really use a quit buddy and some one i could talk to when the urge starts getting tough... new here and will post more once i have some time to look around a bit more. but i am going to quit smoking in 4 days.... i will not be swayed by that date. o.k. now everyone, well shoot i am almost ashamed but anyway i did slip up and smoked a cigarette today, but that will be my last to slip!!!! i am not proud of this but i quit for 10 days, had a trigger and started smoking again. i am going out of town for a week so i will set a new quit date when i return.;</td>
<td>NA</td>
</tr>
<tr>
<td>Abstinent</td>
<td>after 43 years of smoking, as of today, i haven’t smoked for 32 days!!! too cool!!! okay it is day three still haven’t smoked. y’all have been great! i have been reading like crazy, thank you so much. do not want to get ahead of myself but today is and will make #5 i am just so happy i am doing this. day 11 for me and still going strong, congrats to everybody who has quit smoking.</td>
<td>32 days</td>
</tr>
</tbody>
</table>

For participants in group 3, for whom content inference yielded ambiguous classification, we conducted an exploratory analysis to discover the most accurate classification scheme. Participants in group 3 were classified as abstinent or smoking based on the number of days for which abstinence was inferred. For this exploratory analysis, we investigated threshold number of days that maximized concordance with self-reported smoking status.
Analysis 1: Comparison to Self-Reported 30-day Abstinence

Of the 95 participants who authored at least one post during the 30 days immediately preceding T0, smoking status on one or more of the thirty days preceding T0 was inferred for 82 participants.

Abstinence Inferred for All Days
There were 23 participants for whom “abstinent” status was inferred for all thirty days preceding T0. At the 3-month study follow-up, all 23 of those participants self-reported 30-day abstinence.

Smoking Inferred for any Days
There were 15 participants for whom “smoking” status was inferred on at least one of the thirty days preceding T0. At the 3-month study follow-up, only 1 of those participants reported abstinence; the remaining 14 reported having smoked.

Abstinence Inferred for Some Days, But Not All
There were an additional 44 participants for whom “abstinent” status was inferred on some, but not all, of the 30 days preceding T0. At the 3-month study follow-up, 34 (77%) reported 30-day abstinence while the remaining 10 reported having smoked.

Concordance With Self-reported 30-day Abstinence
Among the 38 participants for whom content inference suggested an unambiguous smoking status (“abstinent” on all 30 days or “smoking” on at least 1), content-inferred estimates of smoking prevalence did not differ from self-report (39.5% vs. 36.8%, \( p = 1.0 \)) and concordance was almost perfect (kappa = 0.94, 95% CI = 0.84 to 1.00). Concordance among the 44 participants for whom content inference did not reveal an unambiguous smoking status was maximized at kappa = 0.26 (95% CI = 0 to 0.53) by classifying participants as “abstinent” if abstinence was inferred for at least 3 days, and classifying participants as “smoking” if abstinence was inferred on fewer than 3 days. Content-inferred estimates of smoking prevalence were significantly higher than self-report (43.2% vs. 22.7%, \( p = .039 \)). See Table 2 for cross tabs.

Analysis 2: Impact Across All Time Points and Participants

Of the 826 participants in the analytic sample, 719 had written at least one post between enrollment and Day 30 from which content inference was possible. These 719 participants represent 14% of all participants randomized to the RCT. The personal timelines of 36 participants featured one or more instances of conflicted
inference, days on which one post suggested abstinence but a different post suggested smoking. Instances of conflicted inferences were considered “unclear” for all analyses.

Identifying Smoking and Slips
Across all time points, 505 participants wrote at least one post from which “smoking” status was inferred (median number of posts = 2). Posts that indicated smoking tended to be written within a few days of study enrollment (median amount of time until first “smoking” post was 2 days after enrollment). A slip, defined as a post indicating abstinence followed by a subsequent post indicating smoking, was identified for 35 participants.

Identifying Abstinence
Across all time points, 431 participants wrote at least one post from which “abstinent” status was inferred (median number of posts = 2). On average, participants wrote their first abstinence post 2 weeks after study enrollment (median = 14 days). Among posts from which abstinence was inferred, the median duration of abstinence was 5 days.

Table 2. Comparison of Content-Inferred and Self-Reported 30-day Abstinence at month 3

<table>
<thead>
<tr>
<th></th>
<th>Self-Reported</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smoking</td>
<td>Abstinent</td>
<td></td>
</tr>
<tr>
<td>Unambiguous Cases (n = 38)</td>
<td>14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Content Inferred</td>
<td>Smoking</td>
<td>Abstinent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-Reported</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>7</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Abstinent</td>
<td>3</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

1Participants were classified as “abstinent” if abstinence was inferred for 3 or more days.
Discussion

We found that UGC produced by members of an online smoking cessation community contains information that can be used to reliably infer smoking status. The high concordance between our content inference methodology and self-reported 30-day abstinence for the unambiguous cases supports the validity of those inferences. This method has the potential to expand the types of research questions that can be investigated in observational studies of passively collected online data, by allowing inference of smoking status for some (but not all) participants from user-generated data. However, the low concordance for ambiguous cases, as well as the relatively small proportion of total participants for whom unambiguous inference at 3 months post-enrollment was possible, suggest that this method cannot replace more traditional methods (eg, self-report, biochemical verification) for study outcome evaluation.

Observational studies are increasingly common in the emerging area of online cessation interventions. Although RCTs remain the gold standard for causal inference and evaluation of intervention and exposure effects, observational designs that leverage passively collected data have several strengths. Observational studies can often be conducted at lower cost than RCTs, and can be more efficient for exploratory research questions and discovery. Prospective observational designs can provide evidence for causal relationships where random assignment is not possible due to practical or ethical considerations. Specifically in the domain of online interventions, where real-world implementations are constantly evolving in response to technological and environmental developments, the speed with which observational data may be collected and analyzed can provide intervention designers with invaluable evidence for evaluation and decision making that would take months or years to obtain from an RCT.

These findings are inherently limited to only participants who chose to contribute UGC. For participants who exclusively engaged with the community through passive activities (ie, reading UGC created by others), no inferences were possible. We relied on manual content inference performed by domain experts. To effectively scale the method for application to the millions of posts created by online cessation community members each year, additional tools such as machine-learning–based classification or crowd source classification content inference performed by domain experts. To effectively scale the method for application to the millions of posts created by online cessation community members each year, additional tools such as machine-learning–based classification or crowd source classification would be necessary.

Web-based smoking cessation interventions have been shown to be effective in promoting abstinence, generating quit rates that are comparable to other forms of intervention such as telephone quit lines and face to face interventions. However, additional research and methodological innovations hold the potential to increase efficacy further still, by allowing innovations to intervention design. It is hoped that the availability of this method will contribute a valuable tool towards the advancement of that research agenda.

Funding

The study was funded by the National Cancer Institute of the National Institutes of Health (#R01CA192345, Graham/Zhao, Principal Investigators).

Declaration of Interests

None declared.

References


