A descriptive study of the prevalence and typology of alcohol-related posts in an online social network for smoking cessation

Amy M. Cohn, PhD[1,2], Kang Zhao, PhD[3], Sarah Cha, MSPH[1], Xi Wang, MS[3], Michael S. Amato, PhD[1], Jennifer L. Pearson, PhD[1,4], George D. Papandonatos, PhD[5], Amanda L. Graham, PhD[1,2]

[1] Schroeder Institute for Tobacco Research and Policy Studies at Truth Initiative, Washington, DC

[2] Department of Oncology, Georgetown University Medical Center / Cancer Prevention and Control Program

[3] Department of Management Sciences, The University of Iowa, Iowa City, Iowa

[4] Department of Health, Behavior and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[5] Center for Statistical Sciences, Brown University, Providence, RI

This work is supported by the National Cancer Institute of the National Institutes of Health #R01CA192345 to Graham/Zhao.

Address correspondence to:
Amy Cohn, PhD
Schroeder Institute for Tobacco Research and Policy Studies at Truth Initiative
900 G Street, NW, Fourth Floor
Washington, DC 20001
acohn@truthinitiative.org

Citation:

Cohn, A. M., Zhao, K., Cha, S., Wang, X., Amato, M. S., Pearson, J. L., … Graham, A. L. (2017). A Descriptive Study of the Prevalence and Typology of Alcohol-Related Posts in an Online Social Network for Smoking Cessation. *Journal of Studies on Alcohol and Drugs*, *78*(5), 665–673. https://doi.org/10.15288/jsad.2017.78.665

ABSTRACT

Objective: Alcohol use and problem drinking are associated with smoking relapse and poor smoking cessation success. User-generated content in online social networks for smoking cessation provides an opportunity to understand the challenges and treatment needs of smokers. This study used machine-learning text classification to identify the prevalence, sentiment, and social network correlates of alcohol-related content in the social network of a large online smoking cessation program, BecomeAnEX.org. **Method**: Data were analyzed from n = 814,258 posts (January 2012 to May 2105). Post containing alcohol keywords were coded via supervised machine learning text classification for information about the user's personal experience with drinking, whether the user self-identified as a problem drinker or indicated problem drinking, and negative sentiment about drinking in the context of a quit attempt (i.e., alcohol should be avoided during a quit attempt). **Results**: Less than 1% of posts were related to alcohol, contributed by 13% of users. Roughly a third of alcohol posts described a personal experience with drinking; very few (3%) indicated "problem drinking". The majority (70%) of alcohol posts did not express negative sentiment about drinking alcohol during a quit attempt. Users who did express negative sentiment about drinking were more centrally located within the network compared to those who did not. **Conclusions**: Discussion of alcohol was rare, and most posts did not signal the need to quit or abstain from drinking during a quit attempt. Featuring expert information or highlighting discussions that are consistent with treatment guidelines may be important steps to ensure smokers are educated about drinking risks.

Keywords: Alcohol, smoking cessation, social networks, sentiment, machine learning

INTRODUCTION

Web-based smoking cessation programs play an increasingly central role in tobacco control. They have broad reach to smokers (Pew Internet and American Life Project, 2006), can be accessed 24/7 and in an on-demand format, and are cost effective (An et al., 2010; Graham et al., 2012) and scalable (An, et al., 2010; Bock et al., 2008; Cobb et al., 2005). Large-scale trials report quit rates of 18-20% at 12 months (Graham et al., 2011; Muñoz et al., 2009) and several meta-analyses and systematic reviews provide evidence of effectiveness (Civljak et al., 2013; Graham et al., 2016; Hutton et al., 2011; Shahab and McEwen, 2009). A unique aspect of some web-based cessation programs is their ability to facilitate the exchange of information and support throughout the quitting process via online social networks. Participation in these forms of online social interaction have been shown to have both physical and psychological benefits (Eysenbach et al., 2004; Idriss et al., 2009; Qiu et al., 2011; Wang et al., 2014; Zhao et al., 2014). User generated content in online social networks also provides a unique opportunity to understand the specific challenges and potential treatment needs of smokers.

Alcohol use is one topic that may be especially relevant to explore in an online network for smoking cessation (Cunningham et al., 2006; Cunningham et al., 2008). Alcohol use is robustly associated with heavier smoking, greater nicotine dependence, lower motivation to quit (Cargill et al., 2001), and poor smoking cessation outcomes (Hughes and Kalman, 2006; Leeman et al., 2008). It has also been identified as a primary trigger for smoking relapse (Kahler et al., 2010; McKee et al., 2006). Research shows that nearly half of all smoking-related slips occur during a drinking episode, and even after a quit attempt, individuals are more likely to slip or relapse back to smoking on days in which any alcohol is consumed, compared to non-drinking days, and on heavy drinking days(Leeman, et al., 2008). Because alcohol use is associated with relapse, tobacco cessation treatment guidelines recommend that smokers consider limiting or abstaining from alcohol while quitting (Fiore et al., 2008). Consuming alcohol is normative and routine for many smokers (Falk et al., 2006; Piasecki et al., 2011), including those accessing online smoking cessation programs (Cunningham, et al., 2006), making it likely that discussion of alcohol use during a quit attempt would be common in an online community for smoking cessation.

Despite the proliferation of studies investigating adherence to and outcomes of web-based smoking cessation programs, few have examined the content and structure of online social networks for cessation (Bondy and Bercovitz, 2013; Brandt et al., 2013; Burri et al., 2006; Cobb et al., 2010; Cobb et al., 2013; Myneni et al., 2013; Selby et al., 2010; van Mierlo et al., 2012; Zhang et al., 2013), and none have focused on alcohol-related content within the network. Nicotine replacement therapy and coping with cravings are popular topics in cessation networks (Burri, et al., 2006) and requests for support in these areas are common (Zhang, et al., 2013). Several studies have identified characteristics of key network members who act as leaders in online cessation communities, providing advice and support to others (Cobb, et al., 2010; Selby, et al., 2010; van Mierlo, et al., 2012). These individuals generate the most posts and participate in a wide range of threads; however, they comprise only a small proportion of active members. Only one study of which we are aware has examined smoking and drinking conjointly in an online intervention. Cunningham et al (2008) found that among registered users of an online smoking cessation program, one-third of current daily smokers were problem drinkers, 44% were social drinkers, and only 22% were non-drinkers. In sum, very little is known about user-generated content related to alcohol in online communities for smoking cessation.

Most studies of sentiment and network structure in online cessation programs have relied on manual coding of only a fraction of the available social network content (Brandt, et al., 2013;

Burri, et al., 2006; Zhang, et al., 2013). Manual coding of the hundreds of thousands of exchanges that occur within online social networks is impractical. However, advanced text analytics techniques, such as supervised machine learning can allow for scalable coding of a large amount of unstructured data (Krippendorff, 2012) with relatively little human supervision. Text classification is one analytic technique that examines the content of documents (in this case users' posts) and classifies them into different categories. Text classification requires input from human domain experts at the outset to "train" the computer algorithm. For example, to decide whether a post has provided information about drinking behavior, human experts must first annotate, or code a small number of posts as providing this information, or not. The classification algorithm then extracts various features from these posts (Guo et al., 2009; Zhao, et al., 2014), such as the appearance of certain terms in the post (wine; beer). From human annotations, the computer algorithm then "learns" which features are more predictive of the category of posts. These powerful computational social computing methods have been used to automate the analysis of social ties and sentiment in other areas of health (Chee et al., 2009; Guo, et al., 2009; Zhao et al., 2016; Zhao, et al., 2014), but have yet to be applied broadly to online cessation programs.

This study sought to address five key research questions designed to understand the content and sentiment of alcohol-related posts in an online community for smoking cessation: 1) What is the prevalence of alcohol-related content?; 2) How common are alcohol-related posts that suggest personal experience with drinking (versus virtual celebratory "toasts" for smoking abstinence milestones or general comments about alcohol)?; 3) How common are alcohol-related posts that indicate problem drinking behavior or a user who self-identifies as currently or previously in alcohol recovery?; 4) Is the normative sentiment about drinking during a quit attempt focused on limiting alcohol consumption or on abstaining from alcohol completely (i.e., negative sentiment), in alignment with tobacco dependence treatment guidelines?; and 5) Is sentiment about drinking during a quit attempt associated with social network position? That is, are those who express negative sentiment about drinking during a quit attempt more centrally connected, and thus influential within the network, or are they more likely to be users who are only peripherally connected? Our analyses blend powerful machine-learning computational methods with traditional statistical approaches to examine these questions.

METHOD

Data Source

We analyzed longitudinal data from BecomeAnEX.org (EX), a web-based smoking cessation program developed and managed by Truth Initiative (formerly American Legacy Foundation). Launched in 2008, EX was developed in collaboration with Mayo Clinic Nicotine Dependence Center in accordance with the Clinical Practice Guidelines for Treating Tobacco Dependence (Fiore, et al., 2008). A national mass media campaign (McCausland et al., 2011) and ongoing online advertising have resulted in over 700,000 registered users since its inception. Core elements of the site include setting a quit date, tracking cigarettes and identifying smoking triggers, building a support system, and providing information about pharmacotherapy. EX also includes a large online community of current and former smokers who connect via multiple communication channels: private messages, public posts on member profile pages (“message boards”), group discussions, and blog posts. Blog posts and group discussions elicit many-to-many communications whereas message board posts and private messages elicit one-to-one communications. Communication via blogs (and comments), message boards, and threaded group discussions are all public communications that can be accessed by all EX users; private messages occur only between two users.

The dataset used in the current study spanned January 1, 2012 to May 31, 2015 and included records of 814,258 online activities by 9,377 users, including both posting and reading events. This time period was selected due to the community's migration from a different platform prior to this period, which resulted in a slightly different user experience. Our analyses focus on this time frame given the stability of the social network feature set. The content of private messages was not included in the dataset to protect privacy. The study protocol for these analyses was reviewed and approved by Chesapeake Institutional Review Board (protocol #00010302).

Data Reduction – Text Classification

Manual Annotation: We recruited three long-standing members of the EX community to serve as domain experts to first manually annotate (i.e., code) a sub-set of posts, which would later be used for machine-learning coding. Two experts in alcohol and smoking cessation research worked closely with the domain experts to develop and refine an annotation/coding guide based on the key research questions we sought to address. For the first step to creating an annotation guide, we generated a list of alcohol-related keywords (e.g., drink, alcohol, beer; Appendix 1) and searched for these keywords in all public posts using an automated search process. We retrieved 19,547 posts that contained at least one of the alcohol-related keywords. Domain experts manually coded samples of the posts in batches of 200 posts; weekly group meetings were held following each batch to review annotations, clarify disagreements, and agree on revisions to the guide. This process continued iteratively until a clear set of guidelines was finalized with consensus among domain experts and the researchers, and kappa of $\geq$ 0.70 (Fleiss kappa) was reached for each of the four domains. The final annotation guide included the following content domains (CD), each coded as yes/no:

CD1. "Did this post mention anything related to alcohol?"

CD2. "Did the author describe personal experience with drinking in the post?"

CD3. "Did the author describe him/herself as a problem drinking/alcoholic/in recovery or was problem drinking mentioned in the post?"

CD4. "Does the post express negative sentiment about alcohol as it related to quitting?"

Following the practice period, the three domain experts received a random selection of 1,850 posts with alcohol-related keywords and completed annotations for the four content domains listed above. The rating given by the majority of domain experts (2 out of 3) was used as the final code when a rating disagreement arose. The final inter-rater reliability measured by Cohen's Kappa for each of the four content domains was adequate: CD1=0.74, CD2=0.85, CD3=0.88, CD4=0.71.

Machine Learning: Four machine-learning-based binary classifiers, one for each of the content domains, were then trained from the set of 1,850 manually annotated posts. Of these 1850 posts, domain experts identified 672 as being related to alcohol use. A classifier uses machine learning algorithms to automatically detect the label of a post (yes/no for each of the content domains). For each classifier, we first extracted different types of characteristics (e.g., features), including *meta-features* and *text features*. *Meta-features* were characteristics (but not content) of the post including *post length* (in word count) and *post type* (blog post/comment, message board post, group discussion post/reply). *Text features* included *unigrams* (the frequency of each word in a post), *bigrams* (the frequency of each possible two-word sequence in a post), and *Term-Frequency-Inverse Document Frequency (TF-IDF) score*. For example, "smoking cessation" and "cessation program" would be two possible bigrams in a post containing the phrase "smoking cessation program." TF-IDF score is a numeric value that is the product of term frequency, which measures how frequently a term appears in a single post, and inverse document frequency, which measures how rare a phrase is across all posts (Manning et al., 2008).

Next, we evaluated the performance of various classification algorithms on different combinations of feature sets (unigram, bigram, or TF-IDF scores) using standard 10-fold cross validation (Picard and Cook, 1984). This form of validation uses 90% of the posts to train a classifier, and the remaining 10% to test the classifier's performance. The validation and training sets were rotated in 10 different trials. The performance of the different classification algorithms on each of the four content domain questions was evaluated using accuracy, F1 score, and AUC. Accuracy is the percentage of posts whose labels were predicted correctly by the classifier. F1 score is also an accuracy measure defined as

$$F1 = \frac{2*true\ positive}{2*true\ positive+false\ negative+false\ positive}.$$

F1 ranged from 0 to 1, with higher values indicating better performance. AUC is the area under the Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate. AUC also has a range of 0 to 1, with 1 representing the perfect classifier based on the classifications provided by the human domain experts.

If no algorithm dominated others on all three metrics, the best-performing algorithm with the highest AUC was used, as AUC is more robust against skewed prior distributions (Gao et al., 2007). Finally, the best performing classifier algorithms was applied to 17,697 posts (19,547 minus1,850) that contained alcohol-related keyword(s), but were not annotated, and a label of *Yes* or *No* was given to each post for each of the four content domains.

Data Reduction – Social Network Variables

A user's social network position in the EX community was calculated by examining each user's reading and posting behaviors and constructing a social network metric based on the flow of information either “toward” or “away” from that user. In this approach, each node represents a user: a directed tie from user A to user B indicates that a post or message published by user A was accessed by user B. One way to measure how central or important an individual is within such a social network is to use in-degree and out-degree centralities (Krippendorff, 2012). In-degree reflects the total number of EX members a user has been exposed to by reading their posts; out-degree reflects the total number of EX members who have read the content posted by a user (Zhao, et al., 2016). Higher values for in/out-degree reflect greater centrality in the social network.

Data Analysis

We first examined the accuracy of algorithms and results of machine learning coding of the four content domains. Next we calculated the frequency and prevalence of any posts containing alcohol-related content (CD1), and alcohol-related content that mentioned personal experience with drinking (CD2) or problem drinking (CD3), and negative sentiment about drinking in the context of a quit attempt (CD4). Univariate analysis of variance (ANOVA) tests were then conducted to examine whether expressing negative sentiment about drinking (yes/no) was associated with higher or lower levels of network connectivity, as measured by in-degree and out-degree centralities (dependent variables). To meet model assumptions, dependent variables were evaluated for non-normality and both square root and logarithmic transformations were examined to correct for skewness of data. The log transformation best minimized skewness for both in-degree and out-degree, and was used in analyses below.

RESULTS

Machine Learning Classification

Classifier performance is reported in Table 1. Standard classification schemes worked well for CD1 and CD2 classifiers, as evidenced by their satisfactory performance as measured by accuracy, F1 score and AUC. However, for CD3 and CD4, posts with positive labels ("yes") were rare (3.30% for CD3 and 13.46% for CD4 among annotated posts). For these two classification tasks, standard classification schemes were characterized by low recall for the positive ("yes") labels, as well as low AUC, even though the overall accuracy rates were higher than 85%. To improve CD3 and CD4 classifiers for the positive labels, we also adopted under-sampling techniques for CD3 and CD4 to mitigate the effect of low prior probabilities. Under-sampling the negative label can generate a training dataset that is more balanced between both positive ("yes") and negative ("no") labels, so that a classifier can better learn differences between the two. In addition, for CD3, we found that using a list of high frequency keywords and phrases (Appendix 2) improved performance beyond using all words and phrases (i.e., unigrams). For CD4, because the focus was on sentiment towards alcohol use - instead of the sentiment of the whole post - we leveraged a popular method for aspect-based sentiment analysis (Feldman, 2013). We weighted unigrams by their distance to alcohol-related keywords in a post, assuming that words appearing closer to these keywords in texts contribute more to the author's sentiment towards alcohol use. With these modifications to standard text classification schemes, we were able to improve the performance of CD3 and CD4 classifiers. Overall, all classifiers achieved satisfactory performance with over 0.80 for all three performance metrics.

Research Question 1

After applying the four classifiers to all 17,697 unannotated posts with at least one alcohol-related keyword (i.e., drink), our CD1 classifier identified 6,527 posts that were correctly related to alcohol use (i.e, drink alcohol). Adding the 672 posts that were annotated as "positive" for alcohol content by domain experts, we found a total of n = 7,199 alcohol-related posts (6,527 plus 672). The final number of posts that contained alcohol-related content (n = 7,199) differed from the number of posts that contained all possible mentions of alcohol keywords (n = 19,547) because some "mentions" were not related to alcohol. For example, while the word "drink" was an alcohol key word, posts that had the word "drink" were dropped from the final dataset if they referred to drinking coffee, drinking water, soda, or any other non-alcoholic beverage. Alcohol-related posts represented just under 1% of all posts made during the study period (n_{total} = 814,258 posts). The alcohol-related posts were made by 13% of users (n_{total} = 9,377 users) who made any post in the community during the study period.

Research Questions 2 and 3

Among the alcohol-related posts, 33.02% (n = 2,377) described a personal experience with drinking (e.g., *"I drank a lot of beer Sunday night, stayed up late & smoked double the cigarettes I usually smoke"; "I'm also making a goal to not drink until I'm completely confident that I will not smoke!!!"*). Among the alcohol-related posts, only 3.65% (n = 263) were identified as being related to problem drinking. (e.g., *I am going to go to an AA meeting today"; "I'm drinking way too much and blacking out."*)

Research Question 4

Among all alcohol-related posts, 33.07% (n = 2,381) expressed negative sentiment about drinking alcohol when quitting smoking; that is, the user focused on limiting or abstaining from drinking alcohol completely during a quit attempt (e.g., *"No more drinking for a good while, as that's a huge trigger"; "I absolutely promise you that drinking alcohol at this early stage of your quit especially when you are struggling is self-sabotage"*). Alcohol-related posts indicating negative sentiment were roughly equally distributed across three the sub-networks, with 33% occurring in blog posts, 31% in message boards, and 31% in group discussions.

Research Question 5

For the final research questions, we were interested in examining whether users who posted negative sentiment about alcohol use during a quit attempt were more likely to be highly connected and active users on the website compared to users who did not post negative sentiment about alcohol. If highly connected users regard drinking alcohol as an obstacle to quit success and communicate this to others on the website, this would suggest a possible mechanism for “spreading” information about alcohol’s negative effects on cessation success in future interventions.

The EX social network during the study period included 2.58 million ties, and 71,251 nodes with at least one tie. Figure 1 shows differences in network characteristics as a function of negative sentiment about drinking. Compared to those who did not express negative sentiment (the green curve), those who expressed negative sentiment about drinking (the purple curve) showed greater centrality both in-degree [*F*(1,1230) = 107.55; M=701.09±845.06 vs. M=281.72±394.19], and out-degree [*F*(1,1242) = 111.66; M=852.61±1174 vs. M=308.20±451.92].

DISCUSSION

Our analyses showed that alcohol-related content was present in 1% of all posts, and made by 13% of all contributing users. A third of the alcohol-related posts described personal experience with drinking, while only 3% indicated some level of past or current problem drinking behavior by the user. The majority of posts did not express negative sentiment about drinking alcohol during a quit attempt. However, users who did express negative sentiment about drinking during a quit attempt were more highly connected to others in the network, with roughly double the number of network ties compared to those who did not express negative sentiment.

It was surprising that alcohol-related content was so rare. Research shows about a quarter to a third of smokers enrolling in smoking cessation trials (Fridberg et al., 2014; Leeman, et al., 2008; O'Malley et al., 2009) or calling a state quit line report moderate to hazardous drinking behavior (Toll et al., 2012). It is possible that members of an online smoking cessation community are reluctant to discuss the topic of alcohol, especially if they perceive that alcohol is not commonly discussed in the community or that users, particularly more influential ones, hold negative sentiment about drinking and may disparage or judge users for relapsing because of a drink. Alternatively, smokers who are concurrent problem drinkers may be seeking other, more intensive forms of cessation treatment and may not be represented in an online cessation program. Our findings are somewhat consistent with other work of alcohol-related posts on Twitter. For example, one study by West al (2006) found that less than 1% of Twitter posts culled over a 36-day period indicated problem drinking behavior, while another study found that 7.8% of Twitter posts culled over a similar time frame indicated alcohol-related content using a select set of five alcohol-related keywords ("alcohol", "beer", "liquor," "vodka," and "hangover") (Cavazos-Rehg et al., 2015). It is important to note that these studies differed from ours in that they focused on a select group of alcohol-related keywords and coded posts over a much shorter time frame.

The low prevalence of negative sentiment about drinking alcohol during a quit attempt was also surprising, especially because this is at odds with the tobacco dependence treatment guidelines that encourage smokers to avoid or eliminate smoking triggers, including alcohol (Fiore, et al., 2008). However, our result is consistent with analysis of alcohol-related posts on Twitter, where anti-alcohol tweets were found to be much less common than pro-alcohol tweets (Cavazos-Rehg, et al., 2015). Given that users who did express negative sentiment about drinking were more likely to be centrally connected in the network, we may also presume that they are more likely to be abstinent from smoking (Cobb, et al., 2010). If this is indeed the case, the higher prevalence of negative sentiment among well-connected users may reflect that those users are confident ex-smokers who see any alcohol use as a threat to the success of one's quitting and are able to see the clear benefits of abstaining from alcohol completely during the quit smoking process. Because any level of drinking increases the risk for smoking relapse, especially problem drinking (Dawson, 2000; Leeman, et al., 2008), it is important to augment user-generated content with didactic "expert-generated" information that adequately addresses the links between smoking and drinking. The role of the Community Administrator could be leveraged to seed discussions about the risks of drinking alcohol and encourage users to share their experiences with managing alcohol during their quit attempt. Further, it may be useful to prominently feature content about alcohol to make more apparent the challenges about drinking, particularly so that less well-connected users can make an informed decision about whether or when to drink.

This study had several limitations. First, machine text classifiers achieved satisfactory but not perfect accuracy. This could be due the type of qualitative data being extracted, which was highly nuanced even for domain experts. It is possible that we were unable to code all possible posts referencing alcohol use, particularly if the wording used in the post was highly nuanced or colloquial (i.e., "tossing a few back at happy hour"). Second, annotations for CD4 focused only on negative sentiment about drinking during a cessation attempt and did not code for positive or neutral sentiment. We were specifically interested in the extent to which user-generated content aligned with tobacco treatment guidelines (i.e., expressed negative sentiment that alcohol should be limited or avoided), but may have missed important insights about the extent to which community members expressed overtly positive sentiment about alcohol. Future work should examine the extent which both negative and positive sentiment about alcohol use during a quit attempt occur in the network. Third, we did not link the creation of an alcohol-related post to the user's social network position at the time of the post, nor examine whether or to what degree network position may have changed over time as a function of posting about alcohol-related content. Finally, data on user characteristics were not available so we are unable to characterize the individuals included in these analyses. The goal of our paper was to understand the extent to which alcohol was present in user-generated content rather than identifying the individual characteristics of those who posted about alcohol use. This is an important topic for future research.

This study adds to the literature in several ways. This is the first and largest study that has mined the prevalence and sentiment of alcohol-related content in an online smoking cessation program. At an aggregated level, prior studies have examined associations between mental health status or psychological distress with adherence to online smoking cessation programs and characteristics associated with users (Lukowski et al., 2015; Talati et al., 2016; Vickerman et al., 2015). Second, this study used machine learning for text classification as an innovative and novel approach to study alcohol-related conversations at a scale that is not possible with manual coding (Krippendorff, 2012). Such scalability on large-scale dataset reduces sampling bias and enables analyses at both the community and individual level. Third, this work is consistent with SAMHSA initiatives to integrate substance use and tobacco dependence treatments (Santhosh et al., 2014), and priorities of the NIH Collaborative Research on Addictions to understand factors associated with addiction comorbidities through multidisciplinary work (National Institutes of Health, 2016). Our findings provide an important foundation for understanding the interplay between smoking and alcohol use among members of an online smoking cessation program, and identifying potential touchpoints for intervention.

CONCLUSIONS

This study breaks new ground as the first exploration of alcohol-related content in an online social network for smoking cessation. We have blended social network analyses with rigorous methods in machine learning to focus on alcohol use as an important comorbidity of tobacco dependence. Our analyses provide a methodological framework that could be applied to other important topics related to smoking cessation (e.g., use of pharmacotherapy or other nicotine products). Our overarching goal is to build a rich understanding of the individual, interpersonal, and network level influences on tobacco use behavior that may be amenable to intervention. Our findings underscore the notion that individuals who are centrally connected in a social network can play an important role in the spread of information. With the evolution of Web 2.0 technologies and the increasingly customizable array of communication channels, our findings lay important groundwork for exploring ways to better address the comorbid use of alcohol and tobacco.

ACKNOWLEDGEMENTS

This work is supported by the National Cancer Institute of the National Institutes of Health #R01CA192345 to Graham/Zhao.

CONFLICTS OF INTEREST

AMC, SC, MSA, JLP, and ALG are employees of Truth Initiative, which runs the BecomeAnEX.org smoking cessation website.

REFERENCES

An, L. C., Betzner, A., Schillo, B., Luxenberg, M. G., Christenson, M., Wendling, A., . . . Kavanaugh, A. (2010). The comparative effectiveness of clinic, work-site, phone, and Web-based tobacco treatment programs. *Nicotine & Tobacco Research, 12*, 989-996.

Bock, B., Graham, A., Whiteley, J., & Stoddard, J. (2008). A Review of Web Assisted Tobacco Interventions (WATIs). *Journal of medical Internet research, 10*, e39.

Bondy, S. J., & Bercovitz, K. L. (2013). "Hike up yer Skirt, and Quit." What Motivates and Supports Smoking Cessation in Builders and Renovators. *International journal of environmental research and public health, 10*, 623-637.

Brandt, C. L., Dalum, P., Skov-Ettrup, L., & Tolstrup, J. S. (2013). "After all–It doesn't kill you to quit smoking": An explorative analysis of the blog in a smoking cessation intervention. *Scandinavian journal of public health*, 1403494813489602.

Burri, M., Baujard, V., & Etter, J.-F. (2006). A qualitative analysis of an internet discussion forum for recent ex-smokers. *Nicotine & Tobacco Research, 8*, S13-S19.

Cargill, B. R., Emmons, K. M., Kahler, C. W., & Brown, R. A. (2001). Relationship among alcohol use, depression, smoking behavior, and motivation to quit smoking with hospitalized smokers. *Psychology of addictive behaviors, 15*, 272.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S. J., & Bierut, L. J. (2015). "Hey Everyone, I'm Drunk." An evaluation of drinking-related Twitter chatter. *Journal of studies on alcohol and drugs, 76*, 635-643.

Chee, B. W., Berlin, R., & Schatz, B. R. (2009). *Measuring population health using personal health messages*. Paper presented at the AMIA.

Civljak, M., Stead, L. F., Hartmann-Boyce, J., Sheikh, A., & Car, J. (2013). Internet-based interventions for smoking cessation. *The Cochrane Library*.

Cobb, N. K., Graham, A. L., & Abrams, D. B. (2010). Social network structure of a large online community for smoking cessation. *American journal of public health*, *100*, 1282-1289.

Cobb, N. K., Graham, A. L., Bock, B. C., Papandonatos, G., & Abrams, D. B. (2005). Initial evaluation of a real-world Internet smoking cessation system. *Nicotine & Tobacco Research, 7*, 207-216.

Cobb, N. K., Mays, D., & Graham, A. L. (2013). Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline. *Journal of the National Cancer Institute. Monographs, 2013*, 224-230.

Cunningham, J. A., Selby, P., & van Mierlo, T. (2006). Integrated online services for smokers and drinkers? Use of the check your drinking assessment screener by participants of the Stop Smoking Center. *Nicotine & tobacco research, 8*, S21-S25.

Cunningham, J. A., van Mierlo, T., & Fournier, R. (2008). An online support group for problem drinkers: AlcoholHelpCenter. net. *Patient education and counseling, 70*, 193-198.

Dawson, D. A. (2000). Drinking as a risk factor for sustained smoking. *Drug and alcohol dependence, 59*, 235-249.

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., & Stern, A. (2004). Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj, 328*, 1166.

Falk, D. E., Yi, H., & Hiller-Sturmhofel, S. (2006). An epidemiologic analysis of co-occurring alcohol and tobacco use and disorders. *Alcohol Res Health, 29*, 162-171.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*, 82-89.

Fiore, M., Jaen, C. R., Baker, T., Bailey, W., Benowitz, N., Curry, S. e., . . . Healton, C. (2008). Treating tobacco use and dependence: 2008 update. *Rockville, MD: US Department of Health and Human Services*.

Fridberg, D. J., Cao, D., Grant, J. E., & King, A. C. (2014). Naltrexone improves quit rates, attenuates smoking urge, and reduces alcohol use in heavy drinking smokers attempting to quit smoking. *Alcoholism: Clinical and Experimental Research, 38*, 2622-2629.

Gao, J., Fan, W., Han, J., & Philip, S. Y. (2007). *A General Framework for Mining Concept-Drifting Data Streams with Skewed Distributions.* Paper presented at the SDM.

Graham, A. L., Carpenter, K. M., Cha, S., Cole, S., Jacobs, M. A., Raskob, M., & Cole-Lewis, H. (2016). Systematic review and meta-analysis of internet interventions for smoking cessation among adults. *Substance abuse and rehabilitation, 7*, 55.

Graham, A. L., Chang, Y., Fang, Y., Cobb, N. K., Tinkelman, D. S., Niaura, R. S., . . . Mandelblatt, J. S. (2012). Cost-effectiveness of internet and telephone treatment for smoking cessation: an economic evaluation of The iQUITT Study. *Tobacco control*, tobaccocontrol-2012-050465.

Graham, A. L., Cobb, N. K., Papandonatos, G. D., Moreno, J. L., Kang, H., Tinkelman, D. G., . . . Abrams, D. B. (2011). A randomized trial of Internet and telephone treatment for smoking cessation. *Archives of internal medicine, 171*, 46-53.

Guo, L., Tan, E., Chen, S., Zhang, X., & Zhao, Y. E. (2009). *Analyzing patterns of user content generation in online social networks.* Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.

Hughes, J. R., & Kalman, D. (2006). Do smokers with alcohol problems have more difficulty quitting? *Drug and Alcohol Dependence, 82*, 91-102.

Hutton, H. E., Wilson, L. M., Apelberg, B. J., Tang, E. A., Odelola, O., Bass, E. B., & Chander, G. (2011). A systematic review of randomized controlled trials: Web-based interventions for smoking cessation among adolescents, college students, and adults. *Nicotine & Tobacco Research, 13*, 227-238.

Idriss, S. Z., Kvedar, J. C., & Watson, A. J. (2009). The role of online support communities: benefits of expanded social networks to patients with psoriasis. *Archives of Dermatology, 145*, 46-51.

Kahler, C. W., Spillane, N. S., & Metrik, J. (2010). Alcohol use and initial smoking lapses among heavy drinkers in smoking cessation treatment. *Nicotine & Tobacco Research*, ntq083.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*: Sage.

Leeman, R. F., McKee, S. A., Toll, B. A., Krishnan-Sarin, S., Cooney, J. L., Makuch, R. W., & O'Malley, S. S. (2008). Risk factors for treatment failure in smokers: relationship to alcohol use and to lifetime history of an alcohol use disorder. *Nicotine & Tobacco Research, 10*, 1793-1809.

Lukowski, A. V., Morris, C. D., Young, S. E., & Tinkelman, D. (2015). Quitline outcomes for smokers in 6 states: rates of successful quitting vary by mental health status. *Nicotine & Tobacco Research*, *17*, 924-930.

Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval (1st edition): Cambridge University Press. USA.

McCausland, K. L., Curry, L. E., Mushro, A., Carothers, S., Xiao, H., & Vallone, D. M. (2011). Promoting a web-based smoking cessation intervention: implications for practice. *Cases Public Health Commun Mark, 5*, 3-26.

McKee, S. A., Krishnan-Sarin, S., Shi, J., Mase, T., & O’Malley, S. S. (2006). Modeling the effect of alcohol on smoking lapse behavior. *Psychopharmacology, 189*, 201-210.

Muñoz, R. F., Barrera, A. Z., Delucchi, K., Penilla, C., Torres, L. D., & Pérez-Stable, E. J. (2009). International Spanish/English Internet smoking cessation trial yields 20% abstinence rates at 1 year. *Nicotine & Tobacco Research, 11*, 1025-1034.

Myneni, S., Cobb, N. K., & Cohen, T. (2013). *Finding meaning in social media: content-based social network analysis of QuitNet to identify new opportunities for health promotion.* Paper presented at the MedInfo.

National Institutes of Health. (2016). Collaborative Research on Addictions at the National Institutes of Health, Strategic Plan 2016-2021.

O'Malley, S. S., Krishnan-Sarin, S., McKee, S. A., Leeman, R. F., Cooney, N. L., Meandzija, B., . . . Makuch, R. W. (2009). Dose-dependent reduction of hazardous alcohol use in a placebo-controlled trial of naltrexone for smoking cessation. *International Journal of Neuropsychopharmacology, 12*, 589-597.

Pew Internet and American Life Project. (2006). Most internet usres start at a search engine online when looking for health information online Retrieved August 24, 2016, from http://www.pewinternet.org/files/old-media//Files/Reports/2006/PIP_Online_Health_2006.pdf.pdf

Piasecki, T. M., Jahng, S., Wood, P. K., Robertson, B. M., Epler, A. J., Cronk, N. J., . . . Sher, K. J. (2011). The subjective effects of alcohol–tobacco co-use: An ecological momentary assessment investigation. *Journal of abnormal psychology, 120*, 557.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79*, 575-583.

Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., . . . Portier, K. (2011). *Get online support, feel better--sentiment analysis and dynamics in an online cancer survivor community.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.

Santhosh, L., Meriwether, M., Saucedo, C., Reyes, R., Cheng, C., Clark, B., . . . Schroeder, S. A. (2014). From the sidelines to the frontline: How the Substance Abuse and Mental Health Services Administration embraced smoking cessation. *American journal of public health, 104*, 796-802.

Selby, P., van Mierlo, T., Voci, S. C., Parent, D., & Cunningham, J. A. (2010). Online social and professional support for smokers trying to quit: an exploration of first time posts from 2562 members. *Journal of medical Internet research, 12*, e34.

Shahab, L., & McEwen, A. (2009). Online support for smoking cessation: a systematic review of the literature. *Addiction, 104*, 1792-1804.

Talati, A., Keyes, K., & Hasin, D. (2016). Changing relationships between smoking and psychiatric disorders across twentieth century birth cohorts: clinical and research implications. *Molecular psychiatry*.

Toll, B. A., Cummings, K. M., O'Malley, S. S., Carlin-Menter, S., McKee, S. A., Hyland, A., . . . Celestino, P. (2012). Tobacco quitlines need to assess and intervene with callers' hazardous drinking. *Alcoholism: Clinical and Experimental Research, 36,* 1653-1658.

van Mierlo, T., Voci, S., Lee, S., Fournier, R., & Selby, P. (2012). Superusers in social networks for smoking cessation: analysis of demographic characteristics and posting behavior from the Canadian Cancer Society's smokers' helpline online and StopSmokingCenter. net. *Journal of medical Internet research, 14,* e66.

Vickerman, K. A., Schauer, G. L., Malarcher, A. M., Zhang, L., Mowery, P., & Nash, C. M. (2015). Quitline Use and Outcomes among Callers with and without Mental Health Conditions: A 7-Month Follow-Up Evaluation in Three States. *BioMed research international, 2015,* 817298.

Wang, X., Zhao, K., & Street, N. (2014). Social support and user engagement in online health communities *Smart Health* (pp. 97-110): Springer.

Zhang, M., Yang, C. C., & Gong, X. (2013). *Social support and exchange patterns in an online smoking cessation intervention program.* Paper presented at the Healthcare Informatics (ICHI), 2013 IEEE International Conference on.

Zhao, K., Wang, X., Cha, S., Cohn, A. M., Papandonatos, G., Amato, M. S., . . . Graham, A. (2016). A multi-relational social network analysis of an online community for smoking cessation. *Journal of medical Internet research, 18,* e233.

Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P., & Portier, K. (2014). Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association, 21,* e212-e218.

APPENDICES

1. List of keywords used to retrieve alcohol-related posts (case insensitive):

AA, alcohol, alcoholic, alcoholics anonymous, alcopop, annihilated, bar, beer, bender, beverage, binge, black russian, blitzed, bottle, booze, brew, brewski, bud, budweizer, buzz, buzzed, cabernet, chardonnay, cider, cocktail, cold one, coors, drink, drinkin, drinkin', drinking, drunk, faded, g and t, gin, glass, hair of the dog, hammered, hangover, heineken, henny, hooch, hosed, hungover, inebriated, inhibition(s), intoxicated, keg, liquor, loaded, looped, malt, margarita, merlot, nightcap, pina colada, pinot grigio, pint, plastered, pub, quit killer, rum, sauce/d, smashed, soused, spirit, stewed, tanked/up, tequila, tipsy, under the influence, vodka, wasted, wine, wrecked

2. List of keywords and phrases used for CD3 classifier (case insensitive):

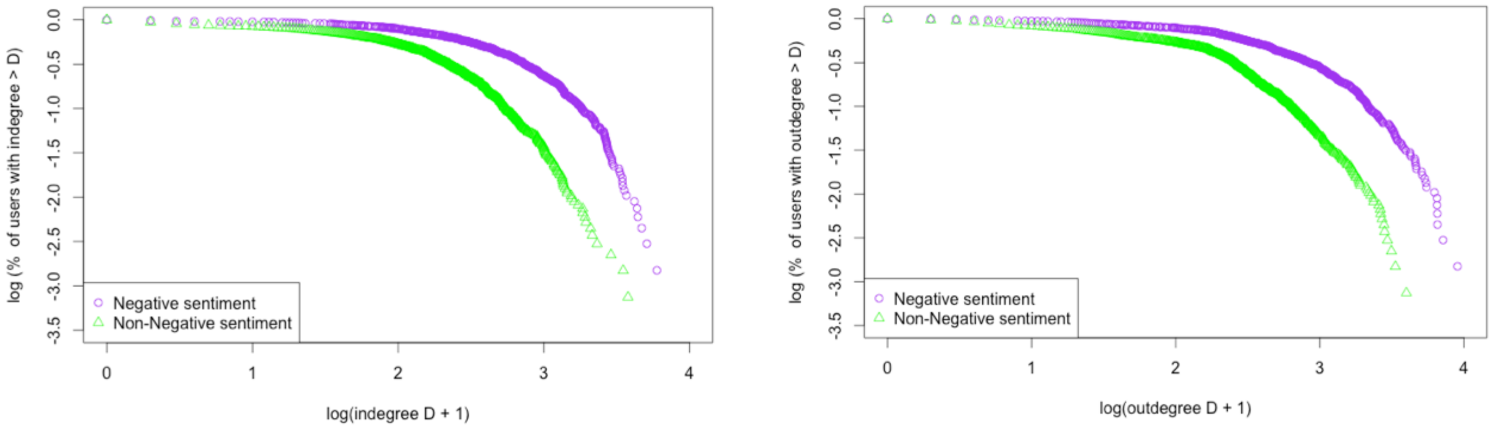
12 step (12-step), AA (A.A.), alcohol being (a) problem, alcoholic, battle alcohol, black(ed/ing) out(s) (blackout), DUI (D.U.I), get (got/getting) wasted, problem(s) with alcohol

Table 1. The best performing classifier for each content domain.

Content Domain	Features	Algorithm	Accuracy	F1	AUC
CD1: “Did this post mention anything about alcohol?”	Meta features, unigrams	J48 Decision tree	0.86	0.86	0.89
CD2: “Did the author describe personal experience with drinking in the post?”	Meta features, unigrams	AdaBoost w/ Naïve Bayesian weak learners	0.81	0.81	0.81
CD3: “Did the author describe him/herself as a problem drinker/alcoholic/in recovery, or was problem drinking mentioned in the post?”	Frequency of 17 words/phrases related to problem drinking	Naïve Bayesian w/ under-sampling	0.96	0.95	0.87
CD4: “Does the post express negative sentiment about alcohol as it related to quitting?”	Meta features, unigrams weighted by TF-IDF scores and distances to alcohol-related keywords	Random forest w/ under-sampling	0.83	0.81	0.81

FIGURE CAPTION

Figure 1. Complementary cumulative distributions (a) in-degree centrality and (b) out-degree centrality for users who expressed negative sentiment towards alcohol use (purple line) vs those who did not (green line).



(a)

(b)