Analysis of Smoking and Drinking Relapse in an Online Community

Acar Tamersoy
College of Computing
Georgia Institute of Technology
tamersoy@gatech.edu

Duen Horng Chau
College of Computing
Georgia Institute of Technology
polo@gatech.edu

Munmun De Choudhury
College of Computing
Georgia Institute of Technology
munmund@gatech.edu

ABSTRACT

Online communities and social media are known to play an important role in improving health efficacy and well-being. In this paper, we examine the role of such platforms in promoting smoking and drinking cessation. We focus on two support communities on Reddit, StopSmoking and StopDrinking, to analyze relapse events among several thousand individuals. For this purpose, we formulate and identify the key engagement and linguistic characteristics of abstainers and relapsers based on participation in the communities spanning almost nine years, and we employ a robust statistical methodology based on survival analysis to examine how participation and these characteristics relate to likelihood of relapse. Our results show that half of the population is at a high risk of relapse within 1-2 months of cessation attempts; however, individuals who continue to abstain beyond three years tend to maintain high likelihood of sustained abstinence. Furthermore, we find positive affect and increased social engagement to be predictors of abstinence. We discuss the implications of our work in tracking effectiveness of online health communities and for designing health interventions.

CCS CONCEPTS

•Information systems →Information systems applications;

KEYWORDS

addiction; smoking; drinking; abstinence; relapse; health; well-being; social media; Reddit; online health communities

ACM Reference format:
Acar Tamersoy, Duen Horng Chau, and Munmun De Choudhury. 2017. Analysis of Smoking and Drinking Relapse
in an Online Community. In *Proceedings of DH '17, London, United Kingdom, July 2–5, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3079452.3079463

1 INTRODUCTION

Addiction challenges, especially to legal substances like tobacco and alcohol, constitute the third leading cause of preventable death and disability in the U.S. [35]. Together, tobacco and alcohol use are causally related to multiple types of cancer, heart disease, cerebrovascular disease, and other chronic conditions. These substances are also attributed to hundreds of thousands of deaths and over 2.5 million years of potential life lost per year in the U.S. [25].

However, maintaining abstinence from tobacco or alcohol is difficult [39]. Research indicates that 80-90% of those who attempt to quit smoking or drinking relapse within a year of their quit dates [15]. Hence, there is a rich body of research on identifying proximal or short-term precipitants of smoking or drinking activities [27, 35, 41]. However, limited research provides statistical and empirical insights into cues that may be associated with abstinence or relapse in the longer term. This is largely because of the difficulty in recruiting individuals identified with this stigmatized health behavior as well as the practical and monetary challenges of long-term tracking of abstinence and relapse experiences [11, 36].

Use of social media platforms and online communities has been found to be linked to improved self-efficacy and well-being, including facilitating recovery from health challenges [13, 28]. Research has indicated these platforms to provide a constantly available source of information and psychosocial support, as well as have been found to foster positive behavior change [23]. Despite some preliminary work examining the link between social media data and addiction cessation [26, 27, 37], empirical investigations and quantitative evidence on *how* participation in social media communities may support or hinder tobacco/alcohol cessation are limited.

In this paper, we address these gaps in prior work examining how activity in an addiction cessation social media community may be used to analyze smoking and drinking relapse events. Thereby, we explore the efficacy of the community in preventing relapse in the long term. Our motivation lies in the observation that the social environment and other psychological influences play particularly critical roles in long-term abstinence from addictive behaviors [14], which now may be quantified with social media. We focus on two specific research questions (RQs):

RQ 1: *How is participation in social media communities that provide support toward smoking and drinking cessation associated with the risk of relapse? Additionally, based on participation in these communities, can we infer the likelihood of relapse over time?*

RQ 2: *Are engagement (e.g., receiving extensive feedback from others) and linguistic constructs of content shared (e.g., expressing greater positive sentiment) within these communities predictors of likelihood of relapse to smoking/drinking?*

We focus on two prominent smoking and drinking cessation communities on the social media site Reddit: StopSmoking[1] and

[1] www.reddit.com/r/StopSmoking

DH '17, July 2–5, 2017, London, United Kingdom
 ISBN 978-1-4503-5249-9/17/07…$15.00.
DOI: http://dx.doi.org/10.1145/3079452.3079463

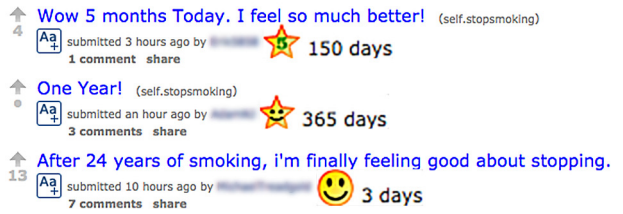

Figure 1: Screenshot from the StopSmoking community. The badge icon next to each post contains information about the posting users' abstinence status, i.e., the number of days the user has self-reported to have been abstaining from smoking. Similar badges exist in the StopDrinking community.

StopDrinking[2]. These two communities are identified as "self improvement communities" on Reddit and are geared toward providing support and motivation to smoking/drinking addiction sufferers. A unique aspect of these communities that makes them suitable for our investigation is that they allow individuals seeking help and support on smoking/drinking cessation to acquire "badges" (see Figure 1). Badges are a mechanism by which individuals can self-report the duration of their smoking/alcohol abstinence. The badges are set up to be updated automatically everyday, unless a user reports a relapse or a change to their abstinence status.

The main contribution of this paper revolves around the study and analysis of relapse and abstinence experiences of about 6 thousand individuals from these two Reddit communities, based on their self-reported badge information. Specifically:

- We devise a methodology to collect longitudinal data on a user's badges in these communities, and thereafter use the badges to identify addiction abstinence or relapse status.
- We formulate and identify the key engagement and linguistic characteristics of abstainers and relapsers based on participation in the communities spanning nine years: 2006–2015.
- We employ a robust statistical methodology based on survival analysis [17] to examine how participation and the characteristics above relate to likelihood of relapse—this method is suitable for analyzing data like ours where the outcome variable is the time until the occurrence of an event of interest (i.e., relapse). To our knowledge, employing this method to study an online community's efficacy in promoting addiction abstinence is novel.

Our results present a number of significant insights that may help researchers better understand the role of participation in online support communities toward tobacco or alcohol relapse and abstinence. We find that the likelihood of experiencing a relapse to smoking/drinking within a day of abstinence is very high; 45%/33% of the individuals in the communities we study are estimated to undergo this event. The median survival time is 25/56 days for smoking/drinking, i.e., half of the population is projected to relapse within about one/two months from start of our study. However, the rate of survival improves significantly beyond three years, suggesting the potential of the communities we study for sustaining cessation among those who do not relapse for a considerable amount of time. Finally, we observe that the linguistic constructs used by the Reddit users in their posts and comments as well as their engagement patterns that capture access to social support are important predictors in preventing relapse.

We discuss the role of social media communities in acting as mediators supporting addiction cessation, and the implications for designing timely, community-centric intervention technologies.

[2]www.reddit.com/r/StopDrinking

2 BACKGROUND AND PRIOR WORK

2.1 Addiction Cessation and Relapse

Factors and precipitants that lead to addiction relapse (e.g., smoking or drinking) have invited the interest of behavioral scientists and addiction researchers for decades [39]. The prevailing theory is that stress and cognitive impairment increase the likelihood of relapse, while social and emotional support tend to act as buffers toward mitigating urges to relapse [19, 29]. Smokers who relapse after a short period report high levels of stress prior to initial abstinence or at one, three, and six months after cessation [39].

However, since there is a direct clinical implication around issuing just-in-time interventions to prevent relapse [20], the majority of existing efforts have focused on identifying the near real-time antecedents of a relapse [35]. Limited research exists in understanding factors that may be associated with *preventing relapse* in the long term. Quantification of these factors is equally important, as they can help evaluate ongoing public health interventions and the design of smoking or drinking cessation programs. An exception is [4] where the size and structure of individuals' social networks were analyzed to find that their connections and interactions relate to reduced smoking tendencies in the long term.

Most of the above studies are, however, retrospective [32]. They identify risk factors in a post-hoc manner based on survey data and retrospective self-reports about mood and observations about relapse episodes. Most of these studies also rely on individuals to actively volunteer and provide self-reported information about their addiction status, making compliance over time not only difficult, but also expensive. Furthermore, since tobacco addiction and alcoholism are stigmatized [11], subject recruitment from the general population is a challenging task. For instance, most prior studies have focused on the 4-5% of smokers who attended smoking cessation clinics or reached out to a counseling hotline [24].

In this work, we leverage participation of individuals in a support community on the social media site Reddit to address some of the above challenges. Longitudinal large-scale data obtained from social media allows us to assess the likelihood of future relapse or abstinence over a long period of time. Moreover, focusing on a semi-anonymous online community, as most of Reddit's communities are, equips us to study a larger and more diverse population interested in obtaining help and advice on cessation. By identifying how participation, engagement, and the nature of content shared relate to relapse, we are further able to explore the role played by an online support community in improving self-efficacy toward long-term abstinence.

2.2 Online Health Communities

People afflicted by medical conditions often find support via online health communities [28]. One study suggests that 30% of U.S. web users have participated in medical or health-related groups [18]

and frequently appropriate online platforms to seek health advice and support in unconventional ways.

Besides support, these communities serve a range of purposes, including seeking advice and connecting with experts and individuals with similar experiences [18]. In this light, approaches to community building have been proposed, e.g., [40], and the role of participation in such communities toward promoting ailment recovery and coping has been examined in a number of different domains, such as diabetes [23]. Other studies have demonstrated that social media provides a way for people to communicate with their contacts about health concerns [30]. Taken together, this body of work supports the notion that people struggling with smoking or drinking cessation may benefit from participation in support communities online, which we examine here.

2.3 Inferring Health Status with Social Media

Recent research in social computing has been able to utilize the abundant and growing repository of social media data to provide a new type of "lens" into inferring health and well-being status of individuals and populations, such as influenza, depression, PTSD, suicide, and so on [6, 9, 10, 33]. A common observation in the above works has been that social interactions and linguistic constructs of content shared by individuals could be utilized toward building robust computational inference frameworks of health risk. Our work builds on this direction by examining to what extent participation, engagement, and attributes of linguistic expression in a social media support community could signal continued abstinence from smoking or drinking related behaviors.

Although limited, there has been some recent work examining social media cues associated with addictive behaviors, including tobacco use and prescription drug use. In an early work, the authors in [26] explored attributes of alcohol references in Facebook profiles of college students using qualitative content coding methods. In [27], for instance, the authors found that among individuals who announced an intent to quit smoking on Twitter, relapsers expressed more negative sentiment compared to those who ceased their smoking behavior during the time of the study. The predictive ability of these cues toward relapse or abstinence was, however, not explored. Culotta's recent work [8] filled this gap; while they identified indicators of smoking cessation attempts, the factors related to long-term abstinence were not studied. The authors in [22] adopted a method similar to [27] to study a prescription drug abuse recovery community. Finally, in our previous work [37], we examined the StopSmoking and StopDrinking communities on Reddit to characterize attributes of short-term (~one month or less) and long-term (~one year and beyond) abstinence from smoking and drinking. But in this work, we did not examine factors that can be predictive of risk to relapse over time, or the efficacy of participation in these online communities in continued abstinence.

While the above pieces of work did demonstrate some predictive capability of the identified cues in inferring addiction relapse or abstinence, their supervised learning-based methodology is inadequate to estimate long-term trajectories of likelihood of relapse or abstinence (see Section 4.2 for details of these limitations). Moreover, prior work like [37] did not consider longitudinal abstinence information; thus could not capture the efficacy of online communities in promoting specific trajectories of abstinence. We extend this body of work by: 1) utilizing self-reported longitudinal information on smoking/drinking abstinence or relapse status of individuals in a support community, and 2) employing a robust statistical methodology, adapted from the survival analysis literature, to explore how participation in the communities we study is related to relapse events over time.

3 DATA

Towards our research goals, we focus on obtaining data from two communities on the social media site Reddit: StopSmoking and StopDrinking, both of which are considered self-improvement communities, or "subreddits" as they are called on Reddit. We refer to them as SS and SD, respectively, through the rest of the paper. Both subreddits host *public* content. As mentioned above, they are support communities for individuals intending to control or stop tobacco/alcohol use, garner thousands of subscribers, and have been examined in our prior work to study patterns of smoking and drinking abstinence [37]. At the time of writing, SS has over 44,000 subscribed users, while SD has over 59,000.

"Badges" as Proxies of Abstinence Progress. As mentioned earlier, a key aspect of these subreddits is that they allow users to acquire "badges" to help track their abstinence progress (see Figure 1). Such badges are subreddit-specific and are displayed next to the username whenever the user posts or comments on the subreddit (ref. Figure 1). Typically, a user makes a badge request to the moderators of the subreddit he or she is interested in through the subreddit's interface or by privately messaging the moderators. Badges are then awarded by the subreddit moderators either manually (SD) or automatically through an application known as "badgebot" (SS). We utilize the information displayed via the badges as a proxy for self-reported ground truth data on abstinence status.

3.1 Data Collection

Our data collection proceeded as follows. We seeded this task by gaining access to a dataset of posts, comments, and associated metadata from SS and SD utilized in our prior work [37]. This seed dataset contained 1,859 SS users (86,835 posts and 766,574 comments) and 1,383 SD users (59,201 posts and 492,573 comments). Since this dataset did not include longitudinal information on the badges or abstinence status of users – data critical to address our RQs, we employed Reddit's official API[3] to devise a method, given below, that extracted longitudinal data going forward from the day of last post in the seed dataset (November 23, 2014).

3.1.1 Obtaining Longitudinal Data. We created two "user dictionaries" containing the author IDs of the SS and SD users existed in the seed dataset, and built a badge value dataset by performing daily crawls for the next five months, from November 24, 2014 to April 23, 2015, and obtaining the badge values of the users on the date of the crawl. The Reddit API limits crawling historical posts on a subreddit to the past thousand posts, so to capture new SS/SD content, each day we also obtained the most recent thousand posts and their associated comments in SS and SD, and stored the new

[3] www.reddit.com/dev/api

Table 1: Summary statistics of the crawled dataset ("All data") and the dataset used in the statistical models ("Survival data").

	StopSmoking (SS)		StopDrinking (SD)	
	All data	Survival data	All data	Survival data
Users	7,221	2,917	7,224	3,074
#Posts	372,414	163,480	285,055	133,887
#Comments	3,424,350	1,496,799	2,907,379	1,333,245
Earliest post	2006-08-02	2006-08-02	2006-02-18	2006-02-18
Latest post	2015-04-23	2015-04-23	2015-04-23	2015-04-23

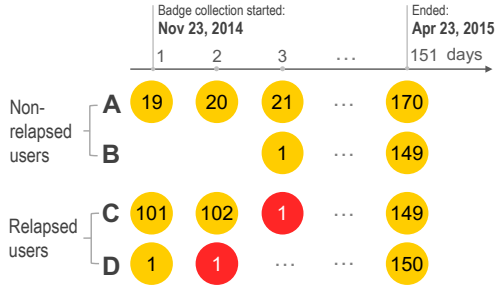

Figure 2: Example badge sequences (rows) obtained from the collection of the daily badge values (values inside the circles) of the users. Users *A* and *B* have strictly increasing badge sequences, indicating successful abstinence, whereas the badge sequences of users *C* and *D* have a drop (102→1) and a repeating badge values of 1, respectively, which indicate a relapse (highlighted in red).

posts or comments in a data batch. Duplicate posts and comments were removed from this batch at the end of the data collection period. For each post and comment, we collected its title, body or textual content, ID, timestamp, and author ID. We included any new user (author of a new post or comment) that we observed during the daily crawls to the corresponding user dictionary, and started collecting their badge values and SS/SD content as well. If the API did not return a badge value for a user, we assigned a special badge value of "NA" to the user. Additionally, we collected the users' historical activity, i.e., posts, comments, and associated metadata, shared in subreddits beyond SS and SD. We henceforth refer to this set of subreddits as OSR (Other SubReddits).

We report the summary statistics of the final crawled dataset in the "All data" columns for SS and SD in Table 1. It is important to note that, per our crawl, each user had at least one recent post or comment in SS/SD, hence had a chance to review their badges recently, therefore our dataset is likely to be free of users who are no longer active in SS/SD and have obsolete badges.

3.1.2 Abstinence Success and Failure from Badges. Now, we discuss how we measure smoking/drinking abstinence success and failure from the longitudinal (daily) badge values of the users. We first used the collection of the daily badge values of a user to establish a *badge sequence* for the user. Figure 2 shows several example

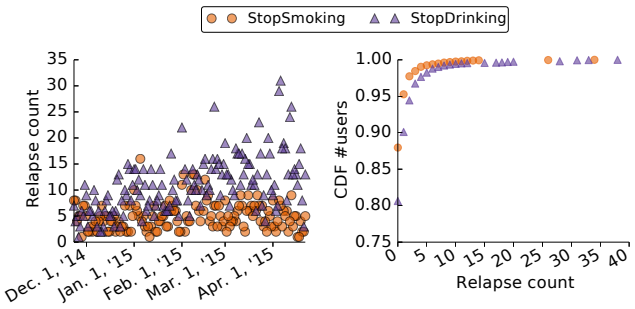

Figure 3: Left: Daily volumes of relapses observed in StopSmoking (SS) and StopDrinking (SD). Right: Cumulative distribution functions (CDFs) of the number of users over the total number of relapses experienced by the users.

badge sequences. We then defined the abstinence and relapse events as follows:

- *Abstinence:* We assumed that the users with strictly increasing badge sequences have successfully abstained from smoking/drinking during our time period of analysis.
- *Relapse:* We assumed that the badge sequences of users who experienced a relapse will be characterized by either (a) an increasing badge sequence with a sudden drop, or (b) a badge sequence with a repeating badge values of 1 (this case captures the users who relapsed on their first day of abstinence).

Our preliminary analysis of the badge sequences revealed a few points to consider before our subsequent statistical analysis.

(1) *Missing badge values.* The badge sequences of 3,342 users/46.28% in SS and 2,994 users/41.45% in SD consisted of only NA values. No badge information means that we do not know about the smoking/drinking abstinence statuses of these users and, hence, they were disregarded.
(2) *Sparse badge values.* As we continued to include new users in our dataset during the daily crawls, for those users admitted shortly before the data collection period ended, we were able to collect only a small number of badge values. To ensure that we have a precise and comprehensive picture of the users' abstinence or relapse history, we omitted the users with an NA badge value and those who had less than 10 badge values.
(3) *Irregularities in values of badge sequences.* Finally, we observed irregularities in the badge sequences of some users. Two prevalent examples were sudden jumps between consecutive badge values (e.g., from the badge value of 30 to 150) and falloffs to large badge values (e.g., from the badge value of 200 to 100). To ensure the integrity of the badge sequences, we omitted the users with badge sequences violating any of the following heuristic rules: for any two consecutive badge values b_t and b_{t+1}, (i) the difference $b_{t+1} - b_t$ should be either negative, 0, 1, or 2, and (ii) if $b_{t+1} - b_t < 0$, then b_{t+1} should be less than or equal to 10.

We report the summary statistics of the filtered dataset in the "Survival data" columns for SS and SD in Table 1. We refer to it as survival data since we leveraged this dataset for our subsequent analyses. Figure 3 shows the daily volumes of relapses and the

cumulative distribution functions (CDFs) of the number of users over relapses experienced by the users. We observe that 2,566 users/87.97% in SS and 2,479 users/80.64% in SD did not relapse during the period of our study. Of those who relapsed, the majority relapsed once (213 users/7.3% in SS and 291 users/9.47% in SD).

4 STATISTICAL METHOD

4.1 Explanatory Variables

We first introduce the variables utilized to analyze smoking and drinking relapse events; they are outlined below and summarized in Table 2. The choice of these variables were framed in the light of prior literature on health recovery and addiction cessation [35, 41], and align with the goals of RQ 1 and RQ 2.

4.1.1 Engagement. Our first set of explanatory variables focus on various aspects of engagement within the SS and SD communities. We consider three dimensions of engagement: self-disclosure, the support received from other users (in-support), and the support provided to other users (out-support).

Literature indicates that self-disclosure can be an important therapeutic ingredient and is linked to improved physical and psychological well-being [5]. In SS/SD, the majority of the posts have a self-disclosing nature, including reflections of feelings, thoughts, and experiences related to quitting [38] (see Figure 1), whereas through the comments the users provide feedback or encouragement to the author of the original post. As such, we capture self-disclosure by considering the users' tendency to submit posts (relative to comments) and define the corresponding variable as the ratio of the number of posts to the total number of posts and comments the user has in SS/SD.

Addiction literature also indicates social support to act as an important mediator of stress during smoking/drinking urges [35]. We consider two forms of social support: in-support and out-support. For both, we consider the users' commentary activities in SS/SD (as a response to a post or another comment) as the primary mechanism of providing feedback and support in these communities. Specifically, we define in-support to be the average number of comments received per post submitted by the user. As the initiator of the discussion in the post, we assume that all the comments on the post contribute towards the in-support of its author (even if some of the comments are directed to other comments). We capture out-support by considering the users' tendency to respond to other users' posts and comments (relative to the number of users who responded to them). Specifically, if set T includes the users to which the corresponding user responded and set F includes the users from which he or she received responses, we define the out-support of the user to be the ratio of the cardinality of set T to the sum of the cardinalities of set T and set F.

This set of explanatory variables therefore contains three variables and we refer to them as *engagement variables*.

4.1.2 Language. Our second set of explanatory variables focus on the linguistic attributes of a user's posts and comments in SS/SD and OSR. The Linguistic Inquiry and Word Count (LIWC: www.liwc.net) is a database containing 74 psychologically meaningful linguistic categories and the word patterns associated with each category. Prior work has used LIWC to characterize individuals at risk for postpartum depression [9], and smokers on Twitter who are at risk for relapse [27]. We introduce a count variable for each of the 12 LIWC categories we deemed the most relevant based on prior work (see Table 2), representing the total number of times that any of the words in the corresponding category appear in the user's posts or comments.

Table 2: List of explanatory variables used in the statistical models for StopSmoking (SS) and StopDrinking (SD). "OSR" stands for subreddits other than SS/SD.

Engagement variables:
self-disclosure SS/SD
in-support SS/SD
out-support SS/SD
Language variables:[a]
first person singular, first person plural, second person, third person pronoun (abbreviated as i*th* pp.) words counts in SS/SD
"body", "health" words counts in SS/SD
past, present, future tense words counts in SS/SD
positive affect (PA), negative affect (NA), "swear" words counts in SS/SD
addiction words count in OSR

[a] Grouped for brevity, those for SS/SD are LIWC-related.

To examine if smoking or drinking-related content in OSR can potentially help characterize smoking and drinking relapse events, we adapt the addiction-related smoking and drinking lexicons from prior work [37]. These two lexicons were compiled from Urban Dictionary[4]. We consider a single count variable (referred to as addiction words count), representing the total number of times that the words in the lexicon appear in the user's posts or comments.

Together, this set of explanatory variables contains 13 variables and we refer to them as *language variables*.

4.2 Survival Analysis

To characterize smoking and drinking relapse events in our data, we leverage the survival analysis techniques [17].

Why Survival Analysis? As achieving long-term abstinence of tobacco or alcohol is challenging [39], relapse to smoking or drinking is a behavior change that can happen anytime, even after years of cessation. However, in studies of human subjects, it is often the case that the study period is not long enough to observe whether the event of interest (relapse in our case) has happened or not. Consequently, the analysis of the probability of "survival" (e.g., prevention of relapse) during the study period as a dichotomous variable (relapsed vs. not relapsed) using conventional statistical techniques (e.g., a linear regression technique or a chi-squared test) fails to account for non-comparability between subjects whose relapse is observed during the study period versus not [17]. Also, simply ignoring subjects who do not experience the event of interest has been noted to produce biased underestimates of survival [34]. Therefore, we borrow techniques from the survival analysis literature for the purposes of our study.

In the survival analysis literature, if the event does not happen before the study ends, the subjects are considered to be *right-censored*

[4] www.urbandictionary.com

at the last assessment time [17]. Another important concept is that of the *survival function* $S(t)$, which denotes the probability that an individual survives at least to time t. The Kaplan-Meier method is a widely used nonparametric technique to graphically construct the unconditional survival function without covariates [17]. It is important to note that this method provides an *estimation* of the survival function if the underlying data is censored (as in our case), but the estimated function is still useful for forecasting purposes [2]. We leverage the Kaplan-Meier method to examine how participation in SS/SD is associated with the risk of relapse (RQ 1).

4.3 Cox Regression

We also employ Cox regression [7] to examine associations between time to first relapse and our explanatory variables (RQ 2). The Cox regression is a statistical technique to analyze survival data where time to event is formulated as a function of possible prognostic factors [12]. It has the advantages of being flexible, allowing the event risk to change over time, and of being semi-parametric, without the need to specify the survival time function [12]. The response variable in Cox regression is typically represented as a pair of values: time to event and a status indicator denoting whether the event of interest has happened or not. If the event of interest does not happen before the study ends, then the time of the event is considered to be *right-censored* at the last assessment time (i.e., while exact the time of the event is unknown, it is known to be at least as long as the follow-up period) [17]. Here, we leverage the users' badge values to determine their response variable values in the regression model. E.g., consider the following examples:

(a) If user A had the badge value of 30 when they experienced the first relapse, then their values for the response variable would be the pair (time to event = 30, relapsed = "yes").
(b) In contrast, if user B did not experience a relapse and had the badge value of 150 on the last day of our observation period, then their values for the response variable would be the pair (time to event = 150, relapsed = "no"), denoting that user B's relapse time is right-censored.
(c) A key point to consider in our case is that users may join SS/SD at any time during their cessation period and, thereby, specify any value for their initial badge in SS/SD. E.g., if user C has been abstaining from smoking/drinking for 200 days and decides to join SS/SD, they would pick 200 as their initial badge value. In this case, we consider user C as a *delayed entry* [17] to our study. The Cox regression supports such delayed entries as the user C; the response variable is then represented as a triplet of values: starting time of the observation, ending time of the observation, and a status indicator as before. Thus, for user C the response variable would be the triplet (observation start = 200, observation end = 300, relapsed = "yes").

4.4 Statistical Models

To understand the explanatory powers of our independent variables, we consider three Cox regression models: the Engagement model, the Language model, and the Engagement + Language model, which consist of the engagement, language, and engagement and language variables, respectively. The first model constitutes our baseline; prior work has indicated that long-term social engagement has a positive impact on the psychological states of individuals [9]. The Language model is motivated from prior work investigating the role of linguistic attributes in describing or predicting health challenges from social media [22, 27], and through the Engagement + Language model, we examine the additional role of engagement in characterizing smoking and drinking relapse events. In these models, we log-transform the language variables (which denote counts) to correct for outliers and skewness. Note that, when computing the values for the variables of non-relapsed users, we consider the data observed for these users during the whole study period, whereas for relapsed users, we only consider the data observed until the time of their relapse. An inherent assumption in Cox regression is the proportional hazards assumption, which states that the coefficients in the regression model should not change with time. We ensured that our explanatory variables satisfy this assumption with the statistical test proposed in [16].

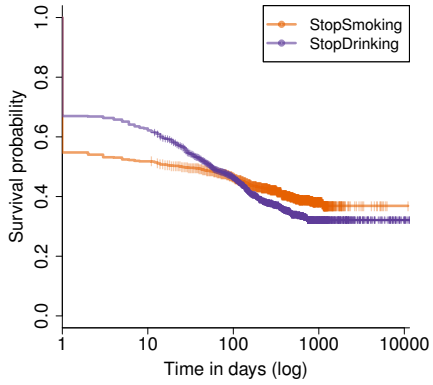

Figure 4: Survival functions obtained for StopSmoking (SS) and StopDrinking (SD) using the Kaplan-Meier method.

5 RESULTS

5.1 RQ 1: Likelihood of Relapse

Per our RQ 1, we begin by examining how the extent of participation in the SS and SD communities relates to estimates of smoking/drinking relapse and abstinence. To that end, Figure 4 shows the survival functions obtained for SS/SD using the Kaplan-Meier method.[5] Both SS and SD have an initial drop-off with 55% and 67% of the users estimated to be at risk of relapse beyond the first day of abstinence.

We also obtain the median survival time from our Kaplan-Meier estimator, which is the time at which 50% of the users are estimated to have relapsed. The median survival time for SS is 25 days (95% confidence interval (CI) = [1, 127]), whereas for SD it is considerably longer with 56 days (95% CI = [35, 102]). These short median survival times of SD and SS align with established studies in the addiction literature [31]. In a way, we find social media-based empirical evidence that bolsters the known fact that

[5] Again, note that this method provides an estimation of the survival functions for our censored data, hence the difference in the number of users at risk of relapse in this analysis and the number of relapsed users according to our relapse criteria in Section 3.1.2.

smoking or drinking cessation is difficult, and the experiences of individuals who participate in the Reddit support communities align with observations about the same made in clinical populations [35].

However, we find that the probability of survival (not experiencing a relapse event) 500 days after being on SS is 40%, while the same for the SD community is 34%. Therefore, although a significant fraction of the populations on both communities are expected to relapse in the short term, survival trend shows a stable pattern in the longer term. In other words, beyond 1000 days, the likelihood of experiencing a relapse is low in both communities.

Survival curves can also be used to estimate the likelihood that a user who has not experienced a relapse event at a specific time point will continue to abstain from smoking/drinking for an additional length of time (calculated by dividing the probability of survival at time t_j by the probability of the same at time t_i, where $j > i$). For example, the probability that a user in SD who did not relapse by 50 days would continue to do so for another 50 days is 0.46/0.51 = 90.2%. If the user does not relapse in 500 days, the probability of continuing the same for another 500 days is 0.32/0.34 = 94.1%. So, as the time of abstinence increases, the likelihood of ever experiencing a relapse event decreases. This analysis provides an alternative explanation of the above observation. In general, our findings align with prior work in the smoking/drinking addiction literature where symptomatic recovery patterns has been examined [35].

What is interesting, however, is the noticeable difference in the survival probabilities for SS and SD. We observe that the SD users are more likely to maintain abstinence beyond any number of days up to about 100 days, after which the SS users become more likely to maintain abstinence. This finding may be explained by the fact that while there is considerably high concomitance between the health behaviors of smoking and drinking [35], smokers tend to relapse at a faster rate than alcoholics; however, those smokers who have maintained abstinence for a while have a greater likelihood than alcoholics to continue to quit post cessation [1].

Overall, we conclude that in the context of RQ 1, participation in the SS and SD communities can lend us valuable insights into patterns and estimates of the likelihood of relapse over time, both in the short term and in the long term.

5.2 RQ 2: Predictor Variables

Recall that the goal of RQ 2 is to examine how attributes of engagement as well as linguistic constructs derived from content shared on SS/SD are associated with and predictive of the likelihood of relapse in the future.

5.2.1 Assessing Goodness of Fit. First, we evaluate the goodness of fits of our models using *deviance*. Briefly put, deviance is a measure of the lack of fit to data, hence lower values are better. It is calculated by comparing a model with the saturated model—a model with a theoretically perfect fit, which we consider to be the intercept-only model and refer to as the *Null* model. Table 3 provides a summary of the different model fits. Compared to the Null models, we observe that all three of our models provide considerable explanatory power with significant improvements in deviances in both SS and SD. The difference between the deviance of a Null model and the deviances of the other models approximately follows a χ^2 distribution, with degrees of freedom (df) equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of the Engagement model with that of the Null model in SS, we see that the information provided by the engagement variables has significant explanatory power: $\chi^2(3, N = 2,917) = 4,235.95 - 4,184.84 = 51.11, p < 10^{-10}$. This comparison with the Null model is statistically significant after Bonferroni correction ($\alpha = \frac{0.05}{3}$ as we consider three models). We observe similar deviance results for the Language and Engagement + Language models in SS and SD, with the latter model possessing the best fit and highest explanatory power.

Table 3: Summary of different model fits (D is model deviance). Null is the intercept-only model. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{3}$).

	StopSmoking (SS)				StopDrinking (SD)			
Model	D	df	χ^2	p	D	df	χ^2	p
Null	4,235.9	0			7,619.1	0		
Engagement	4,184.8	3	51.11	$< 10^{-10}$	7,529.2	3	89.81	$< 10^{-18}$
Language	4,123.9	13	112.0	$< 10^{-17}$	7,484.2	13	134.8	$< 10^{-21}$
Engagement + Language	4,104.1	16	131.8	$< 10^{-19}$	7,424.2	16	194.8	$< 10^{-32}$

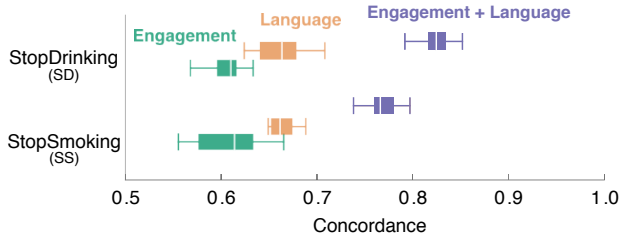

Figure 5: Boxplots for 10-fold cross-validated concordance scores of the statistical models. The Engagement + Language model gives significant predictive power with a mean concordance of 0.77 in StopSmoking (SS) and 0.82 in StopDrinking (SD).

5.2.2 Concordance Analysis. Next, we report the 10-fold cross-validated concordance scores of our Cox regression models to evaluate their predictive power. Briefly put, concordance is a generalization of the area under the receiver operating characteristic (ROC) curve and it measures how well a model discriminates between different responses. Specifically, it is the fraction of the pairs of observations in the data where the observation with the higher survival time has the higher probability of survival predicted by the model [17]. Generally speaking, a concordance of greater than 0.5 indicates a good prediction ability (the value of 0.5 denotes no predictive ability). Here, we first randomly split our dataset into 10 folds and then considered each fold one by one: we trained our models on the remaining 9 folds and computed the concordance scores of the models on the fold under consideration. This led to 10 concordance scores for each model, generated from the same set of folds. Figure 5 shows the boxplots for these concordance scores.

Table 4: Results of Cox regression examining the associations between time to first smoking/drinking relapse and the explanatory variables. "OSR" stands for subreddits other than StopSmoking (SS)/StopDrinking (SD). HR is Hazards Ratio.

	StopSmoking (SS)		StopDrinking (SD)	
Explanatory variable	HR	[95% CI]	HR	[95% CI]
self-disclosure SS/SD	0.87	[0.34, 2.23]	0.22 **	[0.10, 0.48]
in-support SS/SD	1.03	[0.98, 1.08]	1.02 *	[1.01, 1.04]
out-support SS/SD	0.30 **	[0.15, 0.62]	0.17 **	[0.10, 0.29]
1st pp. singular SS/SD	1.55 *	[1.07, 2.23]	1.27	[0.97, 1.66]
1st pp. plural SS/SD	1.10	[0.84, 1.42]	0.95	[0.81, 1.11]
2nd pp. SS/SD	0.89	[0.72, 1.11]	0.98	[0.84, 1.14]
3rd pp. SS/SD	0.90	[0.70, 1.14]	0.93	[0.81, 1.06]
"body" SS/SD	0.99	[0.76, 1.31]	1.04	[0.87, 1.23]
"health" SS/SD	1.02	[0.81, 1.27]	0.80 **	[0.68, 0.93]
past tense SS/SD	0.68 **	[0.53, 0.88]	0.80 *	[0.65, 0.98]
present tense SS/SD	1.28	[0.90, 1.83]	1.41 **	[1.09, 1.83]
future tense SS/SD	1.01	[0.79, 1.31]	0.95	[0.80, 1.13]
PA SS/SD	0.69 **	[0.52, 0.91]	0.83	[0.66, 1.05]
NA SS/SD	0.99	[0.75, 1.32]	1.12	[0.92, 1.37]
"swear" SS/SD	0.99	[0.75, 1.33]	0.90	[0.75, 1.09]
addiction OSR	0.70 **	[0.63, 0.78]	0.80 **	[0.75, 0.85]

** $p < 0.01$, * $p < 0.05$

We observe that the best performing model in both SS and SD is Engagement + Language, which possesses a significant predictive power with a mean concordance of 0.77 and 0.82 in SS and SD, respectively.

Summarily, we conclude that both engagement and language variables include valuable signal relating to the likelihood of relapse or abstinence in the SS/SD communities, compared to either of the categories alone. How do and by how much do these engagement and language variables relate to the risk of relapse? To address this, we present a discussion of the different notable predictors in the next subsection.

5.2.3 Predictors of Relapse and Abstinence. In Table 4, we present expanded results of our best-performing Cox regression model (Engagement + Language), reporting hazard ratios (HRs) and 95% confidence intervals (CIs) of different explanatory variables in this model. In the interest of brevity, we only report these results for our full model (i.e., Engagement + Language) that uses all engagement and language variables. Examining this full model also allows us to evaluate the relative contributions of all of the variables toward estimating relapse likelihood. Here, the hazard ratio for an explanatory variable denotes the risk of a user relapsing with one unit increase in the value of the corresponding explanatory variable (value of the log of the variable for the log-transformed language variables). A hazard ratio smaller than 1 indicates a decreased daily risk of relapse, while a hazard ratio larger than 1 indicates an increased daily risk of relapse.

The contribution of the different explanatory variables to the two characterization tasks is notable. We observe from Table 4 that the language variables are particularly important variables that characterize smoking and drinking relapse events. Below, we highlight the results for some of the prominent language variables, including examples of the most common phrases to provide missing context. As this is a purely correlational analysis, we make no claims as to the (latent) causal mechanisms underlying these findings.

First person singular pronouns are associated with high risk of smoking relapse (HR=1.55, meaning that the risk of relapse to smoking increases by 55% with one unit increase in the value of the log of the first person singular words count SS variable). This category contains words such as "i" and "me"; e.g., post excerpts from SS users who eventually relapsed: "[...] and makes *me* more confident with *my* decision to completely quit, *i* appreciate you taking the time to direct *me* towards it"; "distracting *myself*. tomorrow *i* plan on [...] so *i*'m focused on anything but smoking."; "*i*'m [...] craving a smoke all day, and now that [...]; *i* don't have anything to distract *me* anymore". We presume that since use of first person singular pronouns indicates high self-attentional focus and psychological distress [39], risk of relapse may be heightened due to experience of stress or depressive episodes as indicated in the addiction literature [35]. Additionally, *lower* use of *second person pronouns* (flipping the ratio to denote the decrease in value, HR=1/0.89=1.12 for SS) and *third person pronouns* (HR=1/0.90=1.11 for SS) are indicative of lowered social interaction with the greater community and linked to increased risk of relapse [5] (though, these interactions are not statistically significant).

Past tense words are associated with low risk of smoking/drinking relapse (HR=0.68 for SS; HR=0.80 for SD). This category contains words such as "had" and "felt"; e.g., a comment excerpt from an SS user who maintained abstinence: "i *had* a dream where i smoked one cig, i *felt* incredible sad that my progress *was gone*". This observation is supported by the literature that reflecting on past experiences is known to improve decision-making abilities among addiction quitters, including improving self-control and reducing impulsivity to relapse urges [35]. Additionally, *present tense words* are associated with high risk of drinking relapse (HR=1.41). This category contains words such as "know", "seem"; e.g., a comment excerpt from an SD user who eventually relapsed: "i *don't know* about withdrawals but many cups of tea and lots of candy *seem* to help the cravings". Literature has indicated that focus on the here and now, as captured by the use of present tense, tends to be linked to lowered cognitive functioning and increased mental health challenges—both of which show comorbidity with addiction [24].

Positive affect words are associated with low risk of smoking relapse (HR=0.69). This category contains words such as "fun" and "yay"; e.g., comment excerpts from SS and SD users who maintained abstinence: "so *proud*! *best* of *luck* to you, stay *strong*!"; "[...] and *strong*. *thanks* you guys. i *love* you all. stay *strong*."; "*great* man! *thanks* for dropping in and [...]! you *inspire* me". Our finding is supported by the literature that has found that experience of positive emotions, including regulatory efforts to alleviate negative mood states is strongly linked to smoking cessation and relapse prevention [3, 27]. In contrast, use of *negative affect words* increases the likelihood of drinking relapse (HR=1.12, though this interaction is not statistically significant). Literature indicates increased negative affect to be associated with symptoms such as mental instability, helplessness, loneliness: factors known to trigger addiction urges [21].

Next, *"health" words* are associated with low risk of drinking relapse (HR=0.80). This category contains words such as "medic*" and "alcohol*"; e.g., a comment excerpt from an SD user who maintained abstinence: "i [...] and got *medicine* designed to help *alcoholics detox* from *alcohol* safely". Recognizing the needs of one's health and well-being, as indicated by the use of these words, is known to lead to better lifestyle choices and improvement in self-regulation and self-efficacy [22].

Addiction words are also associated with low risk of smoking or drinking relapse (HR=0.70 for SS; HR=0.80 for SD). One explanation behind this observation could be that some users tend to use other subreddits (OSRs) to receive feedback about the various challenges related to quitting; e.g., a post excerpt submitted to the subreddit *Anxiety* by an SS user: "i had a couple of panic attacks, and decided to quit *smoking* since i figured they were from [...]". Moreover, as with the discussion of health topics, awareness of one's addiction challenges and risk has been known to increase one's cognitive control and therefore reduce risk of relapse [19].

Finally, examining the *engagement variables*, we observe that self-disclosure significantly reduces the risk of drinking relapse (HR=0.22). Also, in-support is associated with high risk of smoking/drinking relapse (HR=1.03 for SS, though this interaction is not statistically significant; HR=1.02 for SD). We conjecture this might be because the users who received greater support from the SS/SD communities are those who are more vulnerable to relapse. Alternatively, it could also be the support-seeking nature of the content shared by users struggling to maintain abstinence, which attracts responses from the greater community. Finally, we observe that out-support is associated with low risk of smoking/drinking relapse (HR=0.30 for SS; HR=0.17 for SD). Prior work has indicated that social engagement has a positive impact on the psychological states of individuals [9]. Hence, we conjecture that greater feedback to other users helps keep individuals more motivated and focused towards their respective self-improvement goals.

6 DISCUSSION AND CONCLUSION

Our results show that participation in the smoking and drinking support communities we study may not be linked to abstinence in the short term—half of the population is estimated to relapse to smoking/drinking within 25/56 days post-cessation. However, the relatively smaller proportion of individuals who *do* survive past the initial few months are estimated to experience sustained abstinence over a long period of time (beyond three years). In essence, while for short-term abstinence our findings call into question the effectiveness of the social media communities, we find that in the course of time these platforms do provide individuals a place where they can improve their regulation and efficacy toward preventing risks of relapse. Direct comparison between our study sample from Reddit and clinical populations would be inappropriate. Nevertheless, our observations align with the literature that indicate that although these behaviors are highly relapse-prone, individuals who have abstained sufficiently long tend to have a considerably lowered probability of resuming their pre-cessation choices [1].

We also discovered several characteristics of engagement and language that indicate increased or decreased chance of relapse. Higher self-attentional focus and detachment from the social realm (pronoun use), and focus on the present increase the risk of relapse. On the other hand, reflection on one's health and addictive behaviors, expression of positive emotions, self-disclosure, and increased desire to provision support to others (engagement variables) heighten the likelihood of abstinence. We also demonstrated the predictive capability of these variables in estimating the communities' cessation behaviors over time. We believe these findings can have notable impact on several points of scientific and practical consideration. We discuss them below.

6.1 Scientific and Practical Implications

Clinical Research. Given the predictive capability of our survival analysis-based method, early warning systems could be developed to analyze participation patterns on the platform. These systems could engage appropriately if the relapse likelihood in the broader community increases beyond a certain level. Moreover, we found that the likelihood of abstinence and relapse can be projected and tracked over time. This could help clinicians better understand people's experiences and strategies around maintaining long-term abstinence from tobacco or alcohol.

Designing Health Interventions. We may design interventions using the engagement variables and linguistic constructs that are associated with increased likelihood of abstinence. By identifying a link between variables that increase risk of relapse and an individual's Reddit activity, moderators could pair them up with peers in the community for support. Social support and higher levels of social capital have been known to help individuals fight addiction urges [14]. In fact, finding "people like me" is a primary stated reason for user participation in online communities [13]. Encouraging or actionable content from others may also be promoted in their activity timelines; positive feedback may improve self-regulation toward abstinence and mediate urges to relapse, whereas instrumental content may help individuals identify and cope with the challenges and struggles that characterize cessation attempts.

Understanding and Tracking Community Efficacy. Our computational approach also demonstrated the ability to proactively identify a community's efficacy toward promoting addiction cessation, including factors linked to such efficacy. Therefore, we believe our methods and the insights we gleaned may be used to create enabling reflective interfaces for community moderators or involved volunteers, so as to not only understand how participation in these platforms supports their goals of self-improvement, but also to make provisions to quantify and improve their effectiveness. These reflective interfaces could take the form of dashboards of "community morale", showing temporal trends of the same in an aggregated manner, with interactive capabilities that allow digging deeper into linguistic and social cues that relate to specific patterns. Reflective interfaces that leverage our methodology of community efficacy tracking, thus will allow moderators to direct attention to those parts of the community where the need of support is greater.

6.2 Limitations and Future Directions

We acknowledge some limitations in our work. Just like survey approaches, our dataset also suffers from the challenges of falsified reporting in their badges. Relatedly, although we used self-reported information on people's abstinence status, our method or findings

do not make diagnostic or treatment-related claims—we cannot be sure if the individuals actually relapsed or abstained from smoking/drinking. The survival analysis method gives likelihood values of the relapse event for the cohort analyzed and does not make individual inferences, hence it cannot be used to predict if or when a specific individual is going to relapse.

Focusing on a large, prominent support community like SS or SD allowed us to analyze abstinence and relapse events over a diverse population; however, we caution against broad generalizations. SS and SD recognize themselves as "self-improvement communities", thus they tend to attract individuals who are already actively considering quitting smoking/drinking. We also cannot causally attribute abstinence or relapse to our explanatory variables, due to the lack of information on whether the users sought support or intervention through offline means. Quantifying the extent to which participation in an online support community complements efforts toward addiction abstinence is an interesting direction for future work.

Competing Interests. The authors have declared that no competing interests exist.

REFERENCES

[1] Thomas H Bien and Roann Burge. 1990. Smoking and drinking: a review of the literature. *Substance Use & Misuse* 25, 12 (1990), 1429–1454.

[2] Don Bryant. 1972. Recent Developments in Manpower Research. *Personnel Review* 1, 3 (1972), 14–31.

[3] Timothy P Carmody. 1989. Affect Regulation, Tobacco Addiction, and Smoking Cessation. *Journal of Psychoactive Drugs* 21, 3 (1989), 331–342.

[4] Nicholas A Christakis and James H Fowler. 2008. The collective dynamics of smoking in a large social network. *New England Journal of Medicine* 358, 21 (2008), 2249–2258.

[5] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social Communication* (2007), 343–359.

[6] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. AAAI.

[7] D. R. Cox and D. Oakes. 1984. *Analysis of Survival Data*. Chapman & Hall.

[8] Aron Culotta. 2016. Towards identifying leading indicators of smoking cessation attempts from social media. In *Proceedings of the IEEE International Conference on Healthcare Informatics*. IEEE, 7–9.

[9] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 626–638.

[10] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

[11] James C Dean and Gregory A Poremba. 1983. The alcoholic stigma and the disease concept. *Substance Use & Misuse* 18, 5 (1983), 739–751.

[12] Edward Faught, Mei Sheng Duh, Jennifer R Weiner, Annie Guerin, and Marianne C Cunnington. 2008. Nonadherence to antiepileptic drugs and increased mortality Findings from the RANSOM Study. *Neurology* 71, 20 (2008), 1572–1578.

[13] Susannah Fox. 2011. *The social life of health information 2011*. Pew Internet & American Life Project Washington, DC.

[14] Sandro Galea, Arijit Nandi, and David Vlahov. 2004. The social epidemiology of substance use. *Epidemiologic Reviews* 26, 1 (2004), 36–52.

[15] Elizabeth A Gilpin, John P Pierce, and Arthur J Farkas. 1997. Duration of smoking abstinence and success in quitting. *Journal of the National Cancer Institute* 89, 8 (1997), 572–576.

[16] Patricia M. Grambsch and Terry M. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 3 (1994), 515–526.

[17] Frank E Harrell. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

[18] Grace J Johnson and Paul J Ambrose. 2006. Neo-tribes: The power and potential of online communities in health care. *Commun. ACM* 49, 1 (2006), 107–113.

[19] Lee Ann Kaskutas, Jason Bond, and Keith Humphreys. 2002. Social networks as mediators of the effect of Alcoholics Anonymous. *Addiction* 97, 7 (2002), 891–900.

[20] Thomas E Kottke, Renaldo N Battista, Gordon H DeFriese, and Milo L Brekke. 1988. Attributes of successful smoking cessation interventions in medical practice: a meta-analysis of 39 controlled trials. *JAMA* 259, 19 (1988), 2882–2889.

[21] Edward Lichtenstein and Russell E Glasgow. 1992. Smoking cessation: what have we learned over the past decade? *Journal of Consulting and Clinical Psychology* 60, 4 (1992), 518.

[22] Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1511–1526.

[23] Lena Mamykina, Andrew D Miller, Elizabeth D Mynatt, and Daniel Greenblatt. 2010. Constructing identities through storytelling in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1203–1212.

[24] G Alan Marlatt, S Curry, and JR Gordon. 1988. A longitudinal analysis of unaided smoking cessation. *Journal of Consulting and Clinical Psychology* 56, 5 (1988), 715.

[25] Ali H Mokdad, James S Marks, Donna F Stroup, and Julie L Gerberding. 2004. Actual causes of death in the United States, 2000. *Journal of the American Medical Association* 291, 10 (2004), 1238–1245.

[26] Megan A Moreno, Dimitri A Christakis, Katie G Egan, Libby N Brockman, and Tara Becker. 2011. Associations between displayed alcohol references on Facebook and problem drinking among college students. *Archives of Pediatrics & Adolescent Medicine* 166, 2 (2011), 157–163.

[27] Elizabeth L Murnane and Scott Counts. 2014. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

[28] Mark W Newman, Debra Lauterbach, Sean A Munson, Paul Resnick, and Margaret E Morris. 2011. It's not that I don't have problems, I'm just not putting them on Facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*. ACM, 341–350.

[29] Raymond S Niaura, Damaris J Rohsenow, Jody A Binkoff, Peter M Monti, Magda Pedraza, and David B Abrams. 1988. Relevance of cue reactivity to understanding alcohol and smoking relapse. *Journal of abnormal psychology* 97, 2 (1988), 133.

[30] Hyun Jung Oh, Carolyn Lauckner, Jan Boehmer, Ryan Fewins-Bliss, and Kang Li. 2013. Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in Human Behavior* 29, 5 (2013), 2072–2080.

[31] Deborah J Ossip-Klein, George Bigelow, Sydney R Parker, Susan Curry, S Hall, and S Kirkland. 1986. Task Force 1: Classification and assessment of smoking behavior. *Health Psychology* (1986).

[32] Maria E Pagano, Karen B Friend, J Scott Tonigan, and Robert L Stout. 2004. Helping other alcoholics in Alcoholics Anonymous and drinking outcomes: Findings from Project MATCH. *Journal of Studies on Alcohol* 65, 6 (2004), 766–773.

[33] Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. *PLoS Currents* 6 (2014).

[34] Philip Rowe. 2007. *Essential statistics for the pharmaceutical sciences*. John Wiley & Sons.

[35] Saul Shiffman. 1982. Relapse following smoking cessation: a situational analysis. *Journal of Consulting and Clinical Psychology* 50, 1 (1982), 71–86.

[36] Saul Shiffman, Sally A Shumaker, David B Abrams, Sheldon Cohen, Arthur Garvey, Neil E Grunberg, and Gary E Swan. 1986. Task Force 2: Models of smoking relapse. *Health Psychology* (1986).

[37] Acar Tamersoy, Munmun De Choudhury, and Duen Horng Chau. 2015. Characterizing Smoking and Drinking Abstinence from Social Media. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 139–148.

[38] Greg Wadley, Wally Smith, Bernd Ploderer, Jon Pearce, Sarah Webber, Mark Whooley, and Ron Borland. 2014. What people talk about when they talk about quitting. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design*. ACM, 388–391.

[39] Alexandra B Whitworth, Felix Fischer, Otto M Lesch, Amanda Nimmerrichter, Harald Oberbauer, Thomas Platz, Adriaan Potgieter, Henriette Walter, and W Wolfgang Fleischhacker. 1996. Comparison of acamprosate and placebo in long-term treatment of alcohol dependence. *Lancet* 347, 9013 (1996), 1438–1442.

[40] Paul Wicks, Michael Massagli, Jeana Frost, Catherine Brownstein, Sally Okun, Timothy Vaughan, Richard Bradley, and James Heywood. 2010. Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research* 12, 2 (2010).

[41] Xiaolei Zhou, James Nonnemaker, Beth Sherrill, Alicia W Gilsenan, Florence Coste, and Robert West. 2009. Attempts to quit smoking and relapse: factors associated with success or failure from the ATTEMPT cohort study. *Addictive Behaviors* 34, 4 (2009), 365–373.