Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk

Munmun De Choudhury[†] **and Emre Kıcıman**[‡]
[†] Georgia Institute of Technology, [‡] Microsoft Research
munmund@gatech.edu, emrek@microsoft.com

Abstract

Online social support is known to play a significant role in mental well-being. However, current research is limited in its ability to quantify this link. Challenges exist due to the paucity of longitudinal, pre- and post mental illness risk data, and reliable methods that can examine causality between past availability of support and future risk. In this paper, we propose a method to measure how the language of comments in Reddit mental health communities influences risk to suicidal ideation in the future. Incorporating human assessments in a stratified propensity score analysis based framework, we identify comparable subpopulations of individuals and measure the effect of online social support language. We interpret these linguistic cues with an established theoretical model of social support, and find that esteem and network support play a more prominent role in reducing forthcoming risk. We discuss the implications of our work for designing tools that can improve support provisions in online communities.

Introduction

Social support is an important ingredient in the attainment of improved mental well-being (Kaplan, Cassel, and Gore 1977; Leavy 1983; Billings and Moos 1984; Stroebe and Stroebe 1996). It is reported that supportive interactions can have a "buffering effect" (Cohen and Hoberman 1983); that is, they can be protective against the negative consequences of mental health. Consequently, social support is recognized to be a significant psychosocial coping resource. It, therefore, bears profound implications for health provisions and interventions that attempt to strengthen aspects of support networks for individuals prone to suicidal thoughts (Cohen, Underwood, and Gottlieb 2000).

With the pervasive adoption of social media platforms, online communities have emerged to be a prime mechanism through which individuals with well-being challenges seek and obtain help, advice, and support (Huh and Ackerman 2012). Perceived support in these communities has been found to be linked to improved self-efficacy and well-being (Fox and Jones 2009), including facilitating recovery from health challenges (Newman et al. 2011), as well as in fostering positive behavior change (Munson et al. 2010).

Despite ample evidence given by cross-sectional studies of the positive therapeutic role of online support (Rappaport 1993), researchers recognize that necessary and adequate empirical data has yet to be accumulated to substantiate this claim (Rudd 1993). Consequently, establishing a causal relationship between availability of online social support and mental health outcomes can be challenging. This is because most studies apply retrospective procedures such as regression and classification models, wherein it is difficult to identify comparable subpopulations of users who have the same set of initial symptoms. It is also plausible that certain individual traits are associated with both the access to social support and the occurrence of at-risk tendencies. Since these symptoms and traits are often the best predictors of well-being risk like suicidal ideation (Cohen and Hoberman 1983), lack of knowledge of baseline conditions can have confounding effects in interpreting the link between social support and at-risk states.

A further threat to validity in cross-sectional studies of online social support is the potential bias in the retrospective measurement of social support among distressed individuals. The challenges are compounded by the difficulty in gathering a pre-morbid (or pre-risk) group of individuals who participate in these online communities. Finally, qualitative and observational studies of support also often fail to reveal conditions under which specific types of support can be non-beneficial or even harmful (Thoits 1982).

Contributions. Our work seeks to address these methodological gaps in assessing the role of online social support in future risk to suicidal ideation. To do so, we examine publicly shared longitudinal post and comment data on a number of prominent mental health communities in the social media Reddit. We then identify individuals in this data who proceed to post on a Reddit suicide support forum at a later point in time. Due to the semi-anonymous nature of these communities (De Choudhury and De 2014), the content shared by individuals allows us to obtain high quality, self-reported data around mental health concerns and suicidal ideation, including data that precedes and succeeds expression of suicidal ideation in individuals.

Utilizing comments received on posts in these communities as a proxy for social support, we develop a human-machine hybrid statistical methodology that models and quantifies the effects of the language of these comments in indi-

viduals who do and do not post on the suicide support forum. Applying stratified propensity score matching (Rosenbaum and Rubin 1983) in a iterative fashion, we first identify linguistic features of comments that show significant effects. We obtain human assessments on the presence of suicidal ideation risk markers in posts associated with these features. Then we filter the features that correspond to comparable subpopulations. Finally, we include these assessments in computing local average treatment effect, so as to assess the effects of specific linguistic features of comments in future risk to suicidal ideation.

Findings. Our findings show that not all individuals posting in the Reddit mental health communities are equally likely to be influenced by support received through comments. Those who benefit from online social support tend to show greater social and futuristic orientation, interpersonal awareness, and lower cognitive impairment. We find these observations to align with observations in the psychology literature on suicidal ideation and support (Kaplan, Cassel, and Gore 1977). Next, we qualitatively interpret the context of use of the linguistic features of the comments, that our stratified matching approach identifies to have significant effects on future risk. To do so, we adopt the social support behavioral code framework of characterizing social support (Cutrona and Suhr 1992). We find that linguistic features used to provide esteem or network support tend to reduce one's risk to suicidal ideation in the future. Somewhat surprisingly, features associated with interpersonal acknowledgments tend to be significantly counter-beneficial.

We describe how our method and findings can provide insights into the positive and negative impacts of online commentary on future mental well-being. We also discuss design and ethical implications of our work in building novel technologies for moderators and volunteers, that seek to improve social support provisions in online communities.

Related Work

Role of Social Support in Health and Well-Being

In the context of mental health, seminal work by Kaplan et al. (Kaplan, Cassel, and Gore 1977) defined social support as "the degree to which an individual's needs for affection, approval, belonging, and security are met by significant others". The study of social support parameters is identified to be a major investigatory tool for examining psychosocial influences upon health and disease (Kaplan, Cassel, and Gore 1977). Rappaport (Rappaport 1993) suggested that, in contrast to psychotherapy, socially shared stories form a kind of group narrative that constitutes a social identity and an avenue for social learning, growth, and behavior change. Cohen and Wills posited that social support "buffers" people from the potentially pathogenic influence of stressful events (Cohen and Hoberman 1983).

Recognizing the role of social support in improved health outcomes, social scientists have developed a helpful categorization schema *Social Support Behavioral Code* (Cutrona and Suhr 1992). This schema was developed by evaluating the frequency of occurrence of 23 communication behaviors intended to be supportive in five categories: *informational support* (providing information or advice), *instrumental support* (expressing willingness to help in a tangible way or actually do so), *esteem support* (communicating respect and confidence in abilities by acts such as complimenting one), *network support* (communicating belonging to a group of people with similar experiences), and *emotional support* (communicating concern, or empathy). We adopt this characterization model of social support in our work.

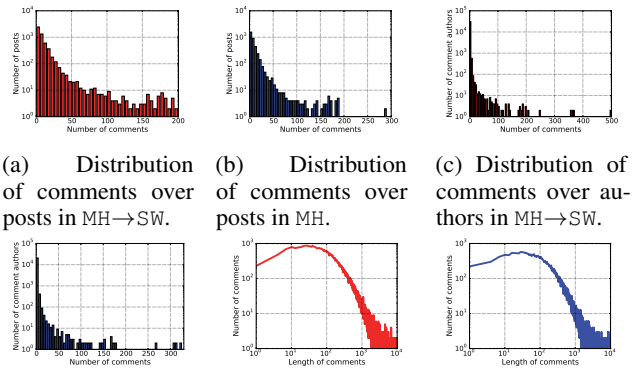
However, efforts to assess the validity and reliability of social support indicators are noted to be lacking in the literature (Thoits 1982). Understanding how specific individuals respond to specific support mechanisms can enable developing tailored ways that improve one's psychological adjustment, efficacy, as well as resistance to and recovery from illness (Burleson et al. 2002). Our work attempts to explore these individual-level differences by studying online mental health support communities and understanding how linguistic attributes as well as types of support enabled through commentary, may relate to future risk or distress.

Online Communities and Social Support

Health Efficacy and Online Support. A rich body of work has examined the important role played by online communities in enabling individuals elicit and provide social support around a variety of health and well-being challenges, ranging from cancer to diabetes (Coulson 2005; Wang, Kraut, and Levine 2012; De Choudhury and De 2014; Andalibi et al. 2016). Online communities have been identified to be powerful platforms where disease-specific guidance and feedback, emotional support, coping and management strategies may be sought (Greene et al. 2011; Newman et al. 2011).

In the realm of mental health, a recent meta-analysis indicates that online support is effective in decreasing depression and increasing self-efficacy and quality of life (Rains and Young 2009). For instance, Oh et al. (Oh et al. 2013) surveyed Facebook users to find that a positive relationship existed between having health concerns and seeking health-related social support. Andalibi et al. (Andalibi et al. 2016) studied how individuals with experience of sexual abuse sought support in different online communities on Reddit for emotional wellness. Our work extends these investigations by examining to what extent we can quantitatively discover links between support and risk to suicidal ideation in different mental health Reddit communities.

Causality. The above research provides valuable insights into whether and how online support relates to and can potentially help improve one's health and well-being. However, we note that significant gaps exist in being able to quantify and measure their influence in future health outcomes. To address this challenge, Winzelberg et al. (Winzelberg et al. 2003) and Lieberman et al. (Lieberman et al. 2003), both conducted clinical trials wherein they assessed the effectiveness of online support communities for individuals with breast cancer. Following a 12 week study comparing two groups with and without access to an online support group BosomBuddies, Winzelberg et al. (Winzelberg et al. 2003) reported that the former groups showed reduced depression, perceived stress,



(a) Distribution of comments over posts in MH→SW.

(b) Distribution of comments over posts in MH.

(c) Distribution of comments over authors in MH→SW.

(d) Distribution of comments over authors in MH.

(e) Distribution of comment lengths (in characters) for users in MH→SW.

(f) Distribution of comment lengths (in characters) for users in MH.

Figure 1: Characteristics of Reddit commentary data.

and cancer-related trauma.

These methods, however, cannot be adopted in naturalistic settings; e.g., in understanding the role of online social support in future health outcomes, based on historical observational data. Moreover, participation in online support communities is inherently self-selected in nature. This can lead to confounding effects if the effectiveness of support is measured through traditional statistical approaches like regression models (Cohen and Wills 1985). We seek to close this gap by borrowing methods from the causal inference literature, in not only examining whether social support can help or exacerbate risk, but also what specific attributes of this support are likely to be less or more beneficial to specific subpopulations.

Data

Identifying At-Risk Individuals. As a starting point of our data collection, we obtained access to a Reddit dataset of mental health posts from De Choudhury et al. (De Choudhury et al. 2016). This dataset included 79,833 posts from 44,262 unique users, that were shared between February 11 and November 11 2014 on 14 mental health subreddits (henceforth referred to as MH) and a prominent suicide support community on Reddit (r/SuicideWatch, henceforth referred to as SW). Example mental health subreddits include r/depression, r/mentalhealth, r/bipolarreddit, r/ptsd, r/psychoticreddit. All of these subreddits host public content, and have been examined and validated by mental health experts in prior work to be communities of mental health and suicidal ideation support (De Choudhury and De 2014).

Following the method developed in (De Choudhury et al. 2016), we constructed two user classes (Table 1). We first seek to identify a set of Reddit users who initially (say time period t_1) post about mental health concerns (in the mental health subreddits, or MH) only, and would later (say time period t_2) be observed to post about suicidal ideation in the SuicideWatch community (or SW). We also identify a second set of users for whom we would not have any observation

	MH		MH→SW	
	t_1	t_2	t_1	t_2
MH	✓	✓	✓	✓or ×
SW	×	×	×	✓

Table 1: Construction of user classes MH→SW and MH.

of posting in SW in t_1 or t_2, despite their postings on the MH subreddits in t_1. For the purposes of our investigation, we consider t_1 to span from Feb 11 2014 to Aug 11 2014, and t_2 from Aug 12 2014 to Nov 11 2014, as also considered in (De Choudhury et al. 2016). The first set of users thus constitutes the class at risk of suicidal ideation in the future (henceforth referred to as MH→SW), while the latter is the class of users not at risk (henceforth referred to as MH). We were able to identify 440 users in the MH→SW group based on this approach. For the MH users, to balance our class sizes going into the ensuing propensity score matching framework, we obtained an equal number of users (440) randomly sampled from the 28,831 users who never posted in SW in t_1 or t_2. As also noted in (De Choudhury et al. 2016), a caveat of this approach is that some of these 28,831 users may have gone on to post on SW outside of the observable period of our dataset (i.e., after t_2); however the large timespan of our data gives sufficient confidence in the purity of the classes derived.

Commentary Data. Recall, the goal of this paper is to assess the role that social support, as manifested in social media, may play in an individual's risk to suicidal ideation in the future. We consider comments made on Reddit posts of the above identified 440 MH→SW and another 440 MH users to be proxies of social support. Prior work has situated commentary in online communities to be mechanisms through which support is extended to help seekers (White and Dorman 2001).

For each of the 880 users spanning both the classes presented above, we grouped their posts, and then employed the official Reddit API (http://www.reddit.com/dev/api) to obtain the entire comment thread (the last 1000 comments) of each post. The comment threads included the text of each comment associated with a post, the author (username) associated with each comment, and its timestamp in UTC. For the 440 MH→SW users, we obtained 62,024 comments that were written by 32,362 unique users, while for the 440 in MH, there were 41,894 comments written by 21,358 unique users. Figure 1 gives descriptive statistics of the commentary data corresponding to the two user classes.

Method

To study how receiving social support, in the form of comments, impacts an individual's future risk to suicidal ideation (or likelihood of being in MH→SW), we seek to isolate the effects of comments from the influence of other factors that might confound the effect of comments on outcomes. The gold standard for this purpose is a randomized controlled trial, where individual posts are assigned to receive a specific comment, independent of other factors. Of course, random-

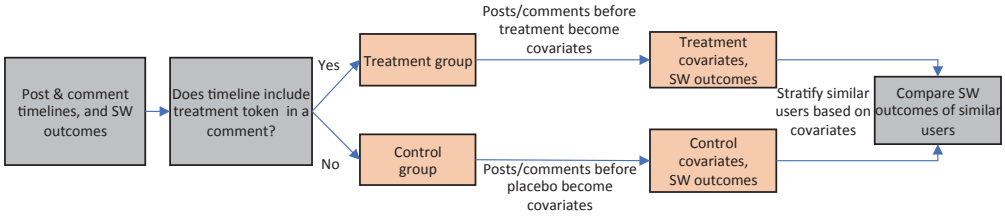


Figure 2: Schematic diagram of data processing and propensity score analysis procedure.

ized controlled trials are not always feasible, due to ethical or practical limitations. As an alternative, we work within the potential outcomes framework of causal analysis (Rubin 2011; Imbens and Rubin 2015) commonly used for observational studies. Specifically, we apply a high-dimensional stratified propensity score method (Rosenbaum and Rubin 1983), conditioning our analysis of the effects of comments on the content of the earlier posts and comments shared and received by users in our dataset (see Figure 2 for a schematic description). Note that, conditioned analyses generally account for observed confounding factors but, unlike randomized controlled studies, cannot account for unobserved confounding factors.

Terminology and Data Preparation. Our unit of analysis is an individual Reddit user, whose experiences are characterized by their posts to the MH communities as well as other users' comments on these posts. We featurize these posts and comments as a per-user sequence of timestamped linguistic *tokens*, or n-grams ($n = 2$); lower-cased and stop-word eliminated. In our analysis, every comment-token appearing in an user's timeline is a *treatment* on that user. In other words, a user is said to have received a given treatment if they author a post that receives a comment containing the corresponding token. All post and comment tokens in an user's timeline that occur before a given treatment are considered to be *covariates*, representing possible factors that might confound our analysis of the treatment's effect on a user's measured outcome (i.e., whether or not a user posts in MH→SW in the future). Correspondingly, a *treatment group* is the set of users who have received a given treatment, and a *control group* is the set of users who have not received that treatment. For a given treatment, the result of our analysis is the difference in measured outcomes between the treatment and control groups, conditioned on covariates.

Stratified Propensity Score Analysis. Stratified propensity score matching attempts to isolate the effects of a particular treatment from the effects of covariates by dividing the treatment and control groups into strata where the covariates of the treatment subgroup within a strata are statistically identical to the covariates of the control subgroup within a strata. Each strata is thus, in essence, artificially approximating a randomized controlled trial where the "assignment" of a treatment is statistically uncorrelated with covariates, allowing us to better distinguish the possible causal effects of a treatment on users' future participation in MH→SW.

The stratification of users is based on their estimated propensity to receive a given treatment. The estimated propensity is a machine-learned function of a user's likelihood of receiving a treatment based on the user's covariates (all prior post- and comment-tokens). Once our method has stratified users, we analyze strata that have common support (Caliendo and Kopeinig 2008)[1]. Within each such strata, the treatment effect is the difference between the measured outcomes of the treatment group and the control group. In our case, this is the difference between the percentage of treated users who eventually participate in MH→SW and the percentage of control users who do so. The population-weighted combination of these strata-level effects is the final local average treatment effect, where local refers to the fact that we are only estimating over strata with common support. We repeat this entire procedure, including the propensity score estimation and stratification, for each of our target treatment tokens.

Implementation. In our implementation, the propensity score function is estimated using the averaged perceptron learning algorithm (Freund and Schapire 1999). Estimation is conducted based on a binary vector representation of users' timelines, $H = h_1, ..., h_n$ where h_i is 1 for a given user if the token i appears in the user's timeline of posts and comments shared and received prior to the treatment token, and 0 otherwise. To distinguish between the effects of words written by users themselves and the effects of words written by others in comments, we treat an n-gram in a comment as being a distinct token from the same n-gram appearing in a post. Given a learned propensity score function, we divide our dataset into 10 strata. In addition to the local average treatment effect, we report the z-score and χ^2 tests of statistical significance. We perform this analysis for all target n-gram tokens that occur in the timelines of more than 10 individuals in the MH communities (11,278 tokens).

Validating Comparability

When assessing the effect of a comment on a user's risk of belonging to MH→SW or MH in the future, our stratified analysis is comparing a treatment group of users who received a particular comment to a control group of people who did not. Having indistinguishable distributions of covariates in these two groups is very important to ensure that any dif-

[1]Common support requires that a strata have sufficient numbers of both treatment and control users. Strata without common support are generally very high- and very low-propensity strata. The semantic interpretation is that, in strata without common support, we cannot distinguish the effects of the treatment from the effects of the prior confounding factors that seem to have predetermined users' treatment status.

Negative treatment effect			
a reasonable	lucky	gently	shit you
heart and	problem but	enjoys	even think
Positive treatment effect			
pain and	stay strong	do well	not easy
advice but	struggled	hating	to respond

Table 2: Sample of comment tokens selected for estimating balance between treatment and control groups.

ference in their likelihood of future posting in SW can be attributed to the fact that one group received the comment and the other group did not. However, unlike a randomized controlled trial, we have no assurance that treatment is independent of *un*observed covariates (or independent of covariates incorrectly represented, from a machine learning perspective, in our covariate set). This can lead to confounded results. Therefore, ensuring that the stratified groups are correctly balanced is critical to correct estimation of outcomes i.e., the likelihood of a user being in MH→SW or MH.

Towards this goal, in this section, we describe an approach to first assess, qualitatively, the comparability (or *balance*) between the treatment and the control groups using human judges. Human judgments of balance are valid in our analysis because it is human commenters who are replying to posts, and thus deciding whether a user gets a treatment or not. We then quantitatively measure potential differences in sociolingustic measures and language models between the two groups. Together, this approach allows us to identify those subpopulations of treatment and control users, who are more likely to be affected by comments on their posts.

Qualitative Analysis of Balance

In this subsection, we present our qualitative approach of assessing balance between treatment and control groups belonging to different propensity strata. We first implemented our proposed propensity score matching technique among the MH posts and comments of the 880 users in our dataset. We identified tokens that had statistically significant treatment effects; negative effect would imply that receiving the comment token decreased chances of being in MH→SW, and vice versa. We randomly sampled 150 tokens with the most positive (or negative) z-scores for our ensuing qualitative assessment. Table 2 shows a sample of these tokens.

Thereafter, corresponding to each of the selected comment tokens, and spanning different propensity strata, we constructed post pairs, belonging to the treatment and control user groups. For each strata (per comment token), we randomly select 10 users from the treatment and 10 users from the control group. For each of these users, we pick their most recent post and all comments they received on that post up until the point of receiving the treatment token (or 'placebo' in the case of the control group). Table 3 gives some paraphrased examples of post pairs thus constructed.

Next we employed two raters, an expert in social media data analysis for mental health and a mental health professional, to qualitatively estimate balance in the post pairs generated above. In other words, the raters' task was to identify

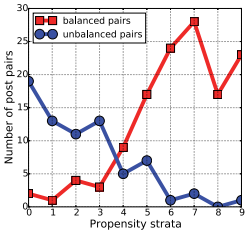


Figure 3: Distribution of post pairs rated to be balanced and unbalanced.

risk markers of suicidal ideation in the post pairs, that may not be observable to the propensity score analysis, however aligned with known observations of behaviors of individuals at risk of suicidal ideation. In a post pair, if the markers aligned, we would infer the treatment and control groups for that particular comment token and strata to be *balanced*. If not, we would assume that our initial propensity score matching analysis needs further tuning to identify more accurately balanced treatment and control groups.

To do so, the raters first utilized a codebook of suicidal ideation risk markers developed in (De Choudhury et al. 2016) to independently rate a random sample of 100 post pairs constructed above to be balanced (very similar: rating of 1), or imbalanced (very dissimilar: rating of 0). The risk markers in the codebook were validated in (De Choudhury et al. 2016) to align with the cognitive psychological integrative model of suicide (Dieserud et al. 2001). They include: mentions of *hopelessness* ("I feel so abandoned"), *anxiety* ("I am feeling panicked"), *impulsiveness* ("right now there's only two ways to ending it"), *lack of self-esteem* ("I am too ugly to make friends"), and expressions of *loneliness* ("i have no one and i never felt such pain"). Following the initial rating exercise followed by resolution of differences, the raters rated a larger sample of 200 post pairs for balance and imbalance estimation. The final agreement was found to be high (Cohen's $\kappa = .81$).

An interesting pattern emerged out of the qualitative coding of the post pairs. Figure 3 gives a distribution of the balance and imbalance ratings over different propensity strata. We observe that the distributions are nearly mirror images of each other. That is, the balance ratings tend to be more frequent among post pairs belonging to higher propensity strata, while the imbalance ratings peak for pairs in the lower propensity strata. The distributions cross each other between stratas 3 and 4. We adopt the lower value 3 as a threshold to identify a set of strata (and corresponding subpopulation of users) for which the treatment and control groups are balanced (propensity strata > 3). We also identify another set where the groups (and corresponding user subpopulations) are not balanced (strata ≤ 3).

Table 3 gives some example (paraphrased) tokens in our rated sample, example strata corresponding to them, and the post pair. As can be observed, the high propensity strata post pairs exhibit many of the markers of mental health challenges identified in our balance assessment codebook. These pairs are also semantically similar ("I have been alone", "there's

Token	Strata	Treatment post	Control post
High propensity strata			
not easy	6	a reason behind my depression is how small by body frame is. i've never cared much about muscle but it's obviously one of the reasons i've been alone (friendships and relationships) for my whole life.	i'm aware there's no way to avoid pain 100%, which is why i'm attempting to go for the least painful way. we've talked in detail about exactly why our issues are troubling for each of us, so he knows that already
advice but	6	i don't even know what all i feel. ashamed, angry, at myself and at the family that never did a thing to support me before. i'm seriously thinking about just pulling out i'm tired of trying, and failing, over and over again.	feeling like shit but noone to talk to, just need a friend who can cheer me up. noones online on facebook that i can talk to so just alone right now...
Low propensity strata			
seek	2	i realize that i'm having depression. i have not showered for a week now, unable to sleep and always thinking negative about myself	i noticed during the livestream, even though that he wasn't using their (i'm assuming) condenser microphone, i felt that his volume and the tones of his voice sounded much more "comfortable" with the headset.
slow down	1	an american football fan but i am intrigued by the world cup. i remember watching 4 years ago and was fascinated.	greetings people, i am a worthless nobody.i guess i want to take more of your time in the vain hopes that you'll somehow be able to make me feel better.

Table 3: Post pairs and associated comment tokens qualitatively assessed to correspond to balanced and imbalanced treatment and control groups. Text has been slightly paraphrased to protect the identities of the users.

no way to avoid pain 100%", "tired of trying, and failing", "feeling like shit but noone to talk to"). While for the lower propensity strata post pairs, only one of the posts shows these markers of risk to suicidal ideation. Summarily, the examples illustrate that our ratings that identified post pairs of these lower strata to have greater imbalance.

Quantitative Analysis of Subpopulation Differences

The above balance analysis showed that the effects of getting a token in a comment may not be homogeneous. Certain users may see little effect of getting a token (low propensity strata), while others see a large effect (higher propensity strata). Essentially, in what ways are the subpopulations of users who fall in the high and low propensity strata different? This is an important consideration in order to understand whether and how users with different characteristics might be less or more likely to be affected by specific tokens appearing in the comments they receive, in their likelihood of being in MH→SW in the future. To answer this and to give more credence to our qualitative balance assessment, we adopted the following quantitative approach.

Specifically, we adopt the use of the LIWC lexicon (Chung and Pennebaker 2007) to quantify the extent to which a variety of sociolinguistic measures are present in the posts of the two subpopulations in the high (strata >3) and low propensity strata (strata ≤ 3). Let us call them subpopulation H and subpopulation L respectively.

In Table 4 we present the results of our analysis of the subpopulations H and L using the LIWC measures – our choice of these measures is motivated from prior work where they were used to examine and understand different types of mental health content on social media (De Choudhury et al. 2013). We observe significant differences across the subpopulations, as given the "Diff" metric—it is the relative percentage difference between the value of a specific measure in the mental health posts of subpopulation H and that of subpopulation L.

The subpopulation H tends to express notably lower *anger*, *sadness*, and *NA* in their posts. Their posts also show lower *inhibition* and higher *cognitive processing*. As shown in the measures of *lexical density and awareness*, this subpopulation also presents more awareness of their context and environment (De Choudhury et al. 2016). Greater use of function words in their posts indicates greater logical coherency in their writing style (Chung and Pennebaker 2007). Interestingly, this subpopulation also tends to share more about their *health* and *work*, and discuss more about *social* and *family* related concerns. Finally, subpopulation H, compared to subpopulation L, tends to show lower self-focus or self-preoccupation as noted in the use of their *1st person singular pronouns*. Conversely, the language of their posts tends to show higher interpersonal focus through use of *2nd and 3rd person pronouns*.

Summarily, the above qualitative and quantitative analyses reveal significant differences between subpopulations H and L: those likely to be affected by commentary (H) manifest a type of underlying behavior and traits that put them in a position to derive greater benefit (or counter-benefit) from comments received on their mental health posts.

Results

Propensity Score Analysis

Once we identify the reliably balanced strata (>3) comprising users are likely to be affected by commentary on their posts, we adapt our stratified propensity score algorithm to ignore other strata, and then computing the final outcome effect over the population of treatment and control users in the remaining strata. Being calculated only over well-balanced strata, this local average treatment effect is thus a more reliable estimate of the effect of a particular comment token, though it is also more limited in its coverage over the population of users.

In Table 5 we report 40 comment tokens that give the most negative or positive z scores in distinguishing between MH→SW and MH users. Corresponding to each token, we also report the absolute number of users in our dataset (out of a total of 880) who received the token in one of their comments (treatment count), the proportion of users in our data who fell into an unclipped strata of the token (coverage), the percent increase in likelihood of belonging to MH→SW in the future

based on getting the token in a comment in the past (local average treatment effect), the z score of the token's likelihood of appearance between the two user classes and associated χ^2 statistic, and the pointwise mutual information (PMI) between the comment token and the outcome.

We find that controlling for historical use of different tokens in the posts of the users as well as the tokens received by them in the comments associated with these posts, getting comments tokens such as "gently" ($z = -2.18$), "is helpful" ($z = -1.45$), "fight the" ($z = -1.22$), "enjoyed it" ($z = -1.1$), "nice i" ($z = -1.06$), "really fun" ($z = -1.01$), "totally agree" ($z = -.93$), "be super' ($z = -.54$), "instructions" ($z = -.53$) significantly decrease a user's likelihood of being in MH→SW in the future. For "gently" this decrease is 31%, for "be helpful" it is 12%, for "fight the" it is 23%, while for "instructions" it is 17%.

Which are the comment tokens getting which *increases* the likelihood of being in MH→SW? We show comment tokens with the most positive treatment effects and high z-scores in Table 5. Getting tokens like "proud" ($z = 5.35$), "am sorry" ($z = 4.77$), "suicide" ($z = 4.67$), "medication" ($z = 4.51$), "depressed" ($z = 4.28$) and "pain and" ($z = 4.24$), "hating" ($z = 3.99$) results in increasing the likelihood of being in MH→SW by 28-58%.

Exploring Context of Use of Comment Tokens

Next we explore the context of usage of the above identified significant comment tokens. This analysis is meant to enable us to understand how different types of social support relate to the outcome of being in MH→SW or MH.

We randomly sampled a set of 100 comments that contained the 20 tokens identified in Table 5 to decrease the likelihood of being in MH→SW, as well as another 100 comments with the 20 tokens that were found to increase future risk to being in MH→SW. We employed qualitative inductive open coding on this sampled corpus to probe into characterizing social support expressed through the tokens.

To develop a codebook and a categorization scheme for the sampled comments, we followed an iterative, semi-open coding procedure. We adopted concepts and characterizations of social support given by the social support behavioral code framework (Cutrona and Suhr 1992), and in the literature on mental health, suicide, and social media (De Choudhury and De 2014). The raters included a mental health expert and a social media expert like our balance analysis. First the raters independently coded a random sub-sample of 50 comments, then discussed each comment together with assigned codes to establish a shared vocabulary. Next, they independently coded the remaining 150 comments based on the shared vocabulary and social support coding scheme thus developed. Interrater reliability Cohen's κ for this task was found to be high: .86.

The final set of social support codes that described the comments included emotional, esteem, informational, instrumental, and network support, as well as interpersonal acknowledgments, aligning with the types given in prior work (Cutrona and Suhr 1992). Table 6 gives example comment experts associated with these categories.

The comment excerpts in Table 6 help us understand, in what ways the tokens identified to have large effects on one's likelihood of being in MH→SW in the future, are used in social

Measures	Diff.	t	p	Measures	Diff.	t	p
Affective attributes				Temporal References			
NA	-7.18	-2.95	*	future_tense	18.28	4.89	***
anger	-12.46	-3.55	**	Social/Personal Concerns			
sadness	-9.40	-3.01	*	family	6.04	2.64	*
Cognition and Perception				friends	20.23	4.62	***
cog. mech	40.75	6.65	***	social	23.67	5.08	***
inhibition	-15.10	-4.45	***	health	12.55	3.58	**
hear	9.40	3.09	**	work	8.98	3.08	**
Lexical Density and Awareness				Interpersonal Focus			
verbs	31.24	6.94	***	1st p. sin.	-20.69	-4.51	***
aux verbs	9.46	3.00	**	2nd p.	15.68	3.27	**
article	25.51	6.24	***	3rd p.	-20.69	5.77	***
adverbs	-10.00	-3.14	**	indef p.	18.76	4.67	***

Table 4: Results of independent sample t-tests between the posts of subpopulations affected by and not affected by comment tokens. Significance is reported following Bonferroni correction: $^{***}: p < 10^{-5}; ^{**}: p < 10^{-4}; ^{*}: p < 10^{-3}$.

Feature	Count	Coverage	Effect	z-val	χ^2	PMI
Negative treatment effect (increased likelihood of being in MH)						
gently	43	0.3	-0.31	-2.18	2.55	0.02
sure of	43	0.49	-0.22	-2.04	2.31	0.15
is helpful	37	0.49	-0.12	-1.45	1.86	0.15
be tough	39	0.51	-0.25	-1.44	0.82	0.01
fight the	34	0.51	-0.23	-1.22	1.53	0.16
enjoyed it	46	0.3	-0.18	-1.1	0.71	0
be ready	39	0.49	-0.04	-1.06	1.41	0.18
nice i	54	0.39	-0.06	-1.06	1.01	0.13
really fun	35	0.2	-0.06	-1.01	1.23	0.13
totally agree	37	0.49	-0.05	-0.93	0.98	0.17
completed	54	0.4	-0.09	-0.91	0.25	0
enjoys	32	0.51	-0.08	-0.85	1.23	0.18
defeat	46	0.4	-0.28	-0.83	0.55	0
to defend	44	0.2	-0.08	-0.79	1.18	0.14
was nice	37	0.49	-0.12	-0.77	0.97	0.1
really liked	40	0.51	-0.1	-0.61	0.88	0.08
be super	42	0.6	-0.03	-0.54	1.38	0.17
instructions	54	0.39	-0.17	-0.53	0.32	0
your home	33	0.49	-0.11	-0.45	0.55	0.08
kindness	42	0.4	-0.11	-0.37	3.42	0.19
Positive treatment effect (increased likelihood of being in MH→SW)						
proud	127	0.6	0.31	5.35	4.14	0.55
a hobby	35	0.49	0.53	4.87	4.57	0.76
am sorry	34	0.49	0.53	4.77	4.69	0.77
suicide	123	0.49	0.28	4.67	4.21	0.49
you wish	32	0.49	0.55	4.54	4	0.8
together with	32	0.49	0.51	4.54	4.16	0.72
medication	114	0.49	0.35	4.51	4.13	0.56
friend you	32	0.49	0.52	4.43	3.8	0.74
your opinion	33	0.4	0.5	4.35	4.44	0.69
to respond	40	0.49	0.49	4.34	3.41	0.69
i care	40	0.4	0.55	4.31	4.57	0.81
depressed	187	0.4	0.3	4.28	5.01	0.53
seek	132	0.39	0.27	4.26	4.41	0.47
pain and	51	0.3	0.58	4.24	6.05	0.87
do well	44	0.4	0.56	4.11	4.32	0.82
stay strong	48	0.4	0.52	4.09	4.16	0.74
medical	133	0.49	0.19	4.05	2.67	0.37
vent	84	0.49	0.32	4.02	3.59	0.54
hating	39	0.49	0.52	3.99	3.02	0.74
misery	34	0.49	0.5	3.96	3.13	0.7

Table 5: Comment tokens given by propensity score matching that contribute to increased or decreased change in likelihood of being in MH→SW or MH respectively.

Higher likelihood of being in MH	**Higher likelihood of being in** MH→SW
Emotional Support	
i *totally agree*. It is hard. I have been there and it is not easy to handle the financial stress, buying a house, girlfriend being eight months pregnant, car issues, job issues, family issues. (↓5%)	I've recently lost friends whom I've known for 10 years, due to me being 'insensitive'. So yes sadly it does happen, I get you and what you are doing through. you are *not alone* (↑16%)
Esteem Support	
cheers mate, *fight the* stigma, you can do it! (↓23%)	You have great potentials in self-actualizing your own situation and ending your *misery*. (↑50%)
Informational Support	
Ever thought of trying to find professional care? I suggest you do that. You need to give life a second chance it may surprise you a lot. I know it can *be tough*, but worth it (↓25%)	If your issue is with the taking of *medication*, talk to them about taking it, discuss your issues with it. Like the guy above said, it may help and could be worth a try, but it is good to discuss concerns about that sort of thing with the person prescribing it. (↑35%)
Instrumental Support	
Start by going for meditation. it can *gently* help you break habitual negative thought patterns, and might also help you get a little bit of that "distance" from yourself that you are looking for (↓31%)	Bro, eat healthy, run, keep your room clean, actively suppress negative thoughts, force yourself to do something productive, even if it's just pursuing *a hobby*. (↑53%)
Network Support	
There is no reason to be nervous and yet everyone here understands and have been precisely at the same place you are in your brave post. [...] i hope some of this discussion *is helpful* to you. (↓12%)	Thats not true at all. everyone in this community really wants to hear your story. They would want *to respond*. Everyones story is worth a listen don't you think? (↑49%)
Acknowledgments	
Exactly this. i would *be super* frustrated too. Anxiety is debilitating and very difficult to cope with (↓3%)	I understand you are *depressed*. Depression is the annihilation of motivation. So it's no wonder u quit the job (↑30%)

Table 6: Example (slightly paraphrased) comment excerpts containing one of the tokens identified to significantly decrease or increase likelihood of being in MH→SW or MH. We show the specific tokens in italics, and their treatment effects inside brackets.

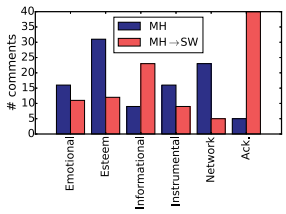


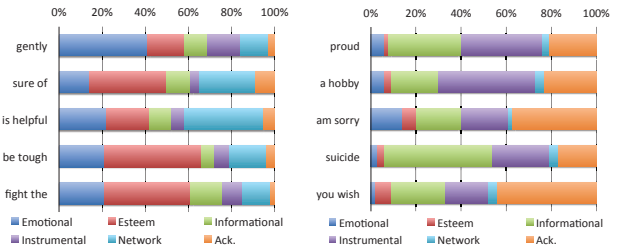
Figure 4: Distribution of different social support types.

support contexts. We observe that emotional support ranges from relating to specific challenging life situations ("totally agree"), to expressing empathy over how they impact one's life ("not alone"). Esteem support can include boosting one's morale to fight the stigma of mental illness ("fight the"), or encouraging hope and uplifting thoughts despite distressful experiences ("misery"). Through informational support, commenters provide advice and suggestions to seek professional mental health help ("be tough"), therapy and medication ("medication"). Commenters also provide various forms of instrumental support, including self-improvement activities like involvement in pastimes ("a hobby"). Next, network support comments express solidarity, connection, and social integration ("is helpful"). Finally, acknowledgments of social support include explicit recognition of the post author's feelings ("be super"), thoughts, or experiences ("depressed").

While the context of use of the specific tokens with significant treatment effects did not reveal notable differences between the likelihood of being in MH and MH→SW, the relative manifestation of the different social support categories in comments for the two user classes did show sharp contrast. This is observable from Figure 4, that gives the distributions of the different support categories over the 200 coded comments, associated with MH and MH→SW. A Kruskal-Wallis one-way analysis of variance test indicates that the difference between these two distributions of social support types is significant ($\chi^2(5; N = 200) = 5.92; p < 10^{-3}$).

For the coded comments associated with MH, we find greater expression of esteem (31%) and network support (23%), followed by emotional support (16%). Informational support (9%) and acknowledgments (5%) tend to be relatively lower for comments containing tokens that decrease likelihood of posting in SW. Overall, this distribution indicates the positive impact of esteem and network support in reducing one's future risk of suicidal ideation expression. However, the coded comments in MH→SW tend to involve largely acknowledgments (40%) and informational support (23%). Instrumental (9%) and network support (5%) constitute the smallest categories for these comments. It appears that receiving acknowledgments of one's feelings, or advice around mental health issues do not translate to reduced future expressions of suicidal ideation.

This analysis demonstrates how our propensity score analysis method can enable us understand the usefulness of various types of support. To probe further, we sought to directly examine how some of the comment tokens with large effects (negative or positive) are used in specific social support contexts. In Figure 5(a-b) we report these findings for five comment tokens with the largest negative and positive treatment effects (ref. Table 5). The relative distributions of support types reported in this figure were determined based on the 200 qualitatively coded comments above. Aligning with our above observations, we find that tokens associated with negative treatment effect appear in comments that our raters coded to be related to emotional, esteem, or network support. On the other hand, we observe that comment tokens that increase the likelihood of being in MH→SW (i.e., are counter-beneficial)



(a) Top five comment tokens with negative treatment effects.

(b) Top five comment tokens with positive treatment effects.

Figure 5: Relative distribution of support types for comment tokens with high (a) negative, and (b) positive treatment.

occur in comments spanning interpersonal acknowledgments and those that provide information or instrumental support.

Discussion

Theoretical and Practical Implications

In this paper we provided a principled, data-driven approach to reveal in what ways social support in the form of commentary can influence mental health outcomes of individuals, specifically their risk of posting about suicidal ideation in the future. Our work bears implications for research in mental health and suicidal ideation, by providing new, previously less understood insights into the mechanics of how (online) social support may impact future risk. Despite tremendous work attempting to establish the beneficial effects of support on well-being (Kaplan, Cassel, and Gore 1977), limited work has focused on *how* social support can improve health (LaRocco, House, and French Jr 1980; Cohen and Hoberman 1983). Literature also indicates that there is limited but under-investigated evidence that not all sources or types of social support are equally effective in reducing distress (Thoits 1982). Our findings speak to both these aspects. By focusing on the online behaviors of a large sample of individuals spanning several months, our work also opens up promising opportunities of employing an unobtrusive data source like social media to not only understand how and what attributes of social support may promote or hinder the well-being of individuals, but also to quantify *prospective* risk based on historically received support.

Beyond these theoretical implications, we believe our methods and these insights we gleaned may be used to create enabling tools and applications. Today, support related moderation practices in online health communities are largely non-algorithmic in nature. Moderators manually examine comments to assess whether they are could be potentially beneficial. In some social computing systems, Reddit inclusive, moderators may rely on community signals like upvotes or downvotes, or refer to comments flagged by community members to assess whether they are potentially helpful or harmful. However, since votes or user reports are accrued over a period of time, support moderation using these kind of approaches may not be prompt enough. For sensitive communities like the ones we study in this paper, judging the supportiveness of comments is time-critical, because unhelpful comments may exacerbate someone's vulnerability.

With our propensity score matching approach, interactive tools may be built for moderators, so that they are able to moderate comments in a more timely fashion. Using such tools, moderators can closely monitor (and discourage) the use of linguistic tokens in comments known to increase the likelihood of posting in SW in the future. On the other hand, the tokens that do include evidence of reducing suicidal ideation risk could be promoted dynamically, especially for those subpopulations, who our approach indicates to be at a greater likelihood of being affected by commentary.

Ethical, Social Challenges and Limitations

We now discuss some ethical and social challenges of this work. The (semi)-automated systems we presented above could allow moderators and support volunteers to make improved decisions and choices based on forecasted likelihood of risk. However, what are the obligations for the moderators or the volunteers when they discover an individual to be at a higher likelihood of suicidal ideation, that might be attributed to a specific instance of support? How can online communities reap the benefits of our method, gain from the design opportunities we outline above; at the same time, protect their ethical obligations? We also envision ethical questions regarding the use of certain type of language in comments. Does preventing a well-intentioned commenter from sharing something be considered to be a impediment to free speech? Taken together, collaborations between computing researchers, mental health experts, moderators, and ethicists can help develop protocols and guidelines that facilitate the use of our work in practical contexts in the future.

Discussing limitations, our method and findings do not reveal users' intent or motivation behind the sharing of specific linguistic cues in comments. We also cannot be sure why certain forms of social support tend to be associated with reduced likelihood of suicidal ideation in the future, or why certain others tend to show converse effects. We also presume self-selection biases in our data. Individuals who post on MH or SW, despite their expression of vulnerability, are after all individuals who are seeking help and advice. Therefore, it is perhaps not surprising that certain types of comments or forms of support are beneficial to them.

We acknowledge limitations to our propensity score analysis method as well. Our data does not meet the strong assumptions that are required to infer true causality: ignorability, and the stable unit treatment value assumption. There are also likely unobserved confounds, such as users' psychological traits, offline behaviors, and history that they may not be mentioned in Reddit posts. While these limitations prevent us from making strong causal claims, in practice we find the results of our analyses provide significant insight about the role of social support in suicidal ideation.

Conclusion

This paper made a methodological contribution in the analysis of online social support for mental well-being. We applied stratified propensity score matching to the content of com-

ments shared on mental health communities on Reddit. Incorporating human assessments into our stratification framework, we were able to identify subpopulations of individuals who were more likely to be affected by the comments. Finally, we qualitatively interpreted how specific linguistic cues of comments, that are associated with high or low likelihood of future suicidal ideation, were used in the context of different forms of social support. Our work bears implications for the design of tools that can improve moderation and support provisions in online support communities.

References

Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proc. CHI*, 3906–3918. ACM.

Billings, A. G., and Moos, R. H. 1984. Coping, stress, and social resources among adults with unipolar depression. *Journal of personality and social psychology* 46(4):877.

Burleson, B. R.; MacGeorge, E. L.; Knapp, M.; and Daly, J. 2002. Supportive communication. *Handbook of interpersonal communication* 3:374–424.

Caliendo, M., and Kopeinig, S. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22(1):31–72.

Chung, C., and Pennebaker, J. W. 2007. The psychological functions of function words. *Social communication* 343–359.

Cohen, S., and Hoberman, H. M. 1983. Positive events and social supports as buffers of life change stress1. *Journal of applied social psychology* 13(2):99–125.

Cohen, S., and Wills, T. A. 1985. Stress, social support, and the buffering hypothesis. *Psychological bulletin* 98(2):310.

Cohen, S.; Underwood, L. G.; and Gottlieb, B. H. 2000. *Social support measurement and intervention: A guide for health and social scientists*. Oxford University Press.

Coulson, N. S. 2005. Receiving social support online: an analysis of a computer-mediated support group for individuals living with irritable bowel syndrome. *CyberPsych & Behavior* 8(6):580–584.

Cutrona, C. E., and Suhr, J. A. 1992. Controllability of stressful events and satisfaction with spouse support behaviors. *Communication Research* 19(2):154–174.

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proc. ICWSM*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Proc. ICWSM*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. CHI*, 2098–2110. ACM.

Dieserud, G.; Røysamb, E.; Ekeberg, Ø.; and Kraft, P. 2001. Toward an integrative model of suicide attempt: A cognitive psychological approach. *Suicide and Life-Threatening Behavior* 31(2):153–168.

Fox, S., and Jones, S. 2009. The social life of health information. *Washington, DC: Pew Internet & American Life Project* 2009–12.

Freund, Y., and Schapire, R. E. 1999. Large margin classification using the perceptron algorithm. *Machine learning* 37(3):277–296.

Greene, J. A.; Choudhry, N. K.; Kilabuk, E.; and Shrank, W. H. 2011. Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *Journal of general internal medicine* 26(3):287–292.

Huh, J., and Ackerman, M. S. 2012. Collaborative help in chronic disease management: supporting individualized problems. In *Proc. CSCW*, 853–862. ACM.

Imbens, G. W., and Rubin, D. B. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kaplan, B. H.; Cassel, J. C.; and Gore, S. 1977. Social support and health. *Medical care* 15(5):47–58.

LaRocco, J. M.; House, J. S.; and French Jr, J. R. 1980. Social support, occupational stress, and health. *Journal of health and Social Behavior* 202–218.

Leavy, R. L. 1983. Social support and psychological disorder: A review. *Journal of community psychology*.

Lieberman, M. A.; Golant, M.; Giese-Davis, J.; Winzlenberg, A.; Benjamin, H.; Humphreys, K.; Kronenwetter, C.; Russo, S.; and Spiegel, D. 2003. Electronic support groups for breast carcinoma. *Cancer* 97(4):920–925.

Munson, S. A.; Lauterbach, D.; Newman, M. W.; and Resnick, P. 2010. Happier together: integrating a wellness application into a social network site. In *International Conference on Persuasive Technology*, 27–39. Springer.

Newman, M. W.; Lauterbach, D.; Munson, S. A.; Resnick, P.; and Morris, M. E. 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 *Proc. CSCW*, 341–350. ACM.

Oh, H. J.; Lauckner, C.; Boehmer, J.; Fewins-Bliss, R.; and Li, K. 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):2072–2080.

Rains, S. A., and Young, V. 2009. A meta-analysis of research on formal computer-mediated support groups: Examining group characteristics and health outcomes. *Human communication research* 35(3):309–336.

Rappaport, J. 1993. Narrative studies, personal stories, and identity transformation in the mutual help context. *The Journal of Applied Behavioral Science* 29(2):239–256.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rubin, D. B. 2011. Causal inference using potential outcomes. *Journal of the American Statistical Association*.

Rudd, M. D. 1993. Social support and suicide. *Psychological reports* 72(1):201–202.

Stroebe, W., and Stroebe, M. 1996. The social psychology of social support.

Thoits, P. A. 1982. Conceptual, methodological, and theoretical problems in studying social support as a buffer against life stress. *Journal of Health and Social behavior* 145–159.

Wang, Y.-C.; Kraut, R.; and Levine, J. M. 2012. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. In *Proc. CSCW*, 833–842.

White, M., and Dorman, S. M. 2001. Receiving social support online: implications for health education. *Health education research* 16(6):693–707.

Winzelberg, A. J.; Classen, C.; Alpers, G. W.; Roberts, H.; Koopman, C.; Adams, R. E.; Ernst, H.; Dev, P.; and Taylor, C. B. 2003. Evaluation of an internet support group for women with primary breast cancer. *Cancer* 97(5):1164–1173.