Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention

Prerna Chikersal¹, Danielle Belgrave², Gavin Doherty³, Angel Enrique⁴, Jorge E. Palacios⁴, Derek Richards⁴ and Anja Thieme²

¹Carnegie Mellon University, prerna@cmu.edu
²Microsoft Research Cambridge, (dabelgra, anthie)@microsoft.com
³Trinity College Dublin, gavin.doherty@tcd.ie
⁴SilverCloud Health & Trinity College Dublin, (angel.enrique, jorge.palacios, derek.richards)@silvercloudhealth.com

ABSTRACT
Online mental health interventions are increasingly important in providing access to, and supporting the effectiveness of, mental health treatment. While these technologies are effective, user attrition and early disengagement are key challenges. Evidence suggests that integrating a human supporter into such services mitigates these challenges, however, it remains under-studied how supporter involvement benefits client outcomes, and how to maximize such effects. We present our analysis of 234,735 supporter messages to discover how different support strategies correlate with clinical outcomes. We describe our machine learning methods for: (i) clustering supporters based on client outcomes; (ii) extracting and analyzing linguistic features from supporter messages; and (iii) identifying context-specific patterns of support. Our findings indicate that concrete, positive and supportive feedback from supporters that reference social behaviors are strongly associated with better outcomes; and show how their importance varies dependent on different client situations. We discuss design implications for personalized support and supporter interfaces.

Author Keywords
Mental health; digital behavioral intervention; CBT; support; machine learning; AI; unsupervised learning; data mining.

CCS Concepts
•Human-centered computing → Human computer interaction; •Computing methodologies → Machine learning;

INTRODUCTION
Mental illness is increasing in occurrence [39]. It presents the largest cause of disability worldwide and is the strongest predictor of suicide [64, 99]. This makes the prevention and treatment of mental health disorders a public health priority [102] and has led to explorations of how the field of HCI, and the development of technology more broadly, can support access to, and increase the effectiveness of, mental health treatment [8, 76, 95, 86]. Over the last decade, this has brought forward developments of mobile apps [19, 29, 55], and computerized psycho-educational and psycho-therapeutic interventions [5, 17, 81, 103], or chat-based [27, 48, 88] programs to complement, and expand access to, psychotherapy.

Most existing digital mental health services are based on Cognitive Behavioral Therapy (CBT); the most widely applied and most extensively empirically tested psychotherapy in Western Healthcare [11]. CBT is solution-focused, teaches the person to attend to the relationships between their thoughts, feelings and behaviors, and is frequently used in treating depression, anxiety or post-traumatic stress. Its highly structured format makes it well suited for support by digital technology [18]. Further, extensive research has evidenced the clinical effectiveness of internet-delivered CBT (iCBT) with sustainable results comparable to face-to-face therapy [4, 5, 100, 103].

Despite these benefits, a key challenge for digital behavioral interventions like iCBT is sustaining the users’ engagement with treatment [26, 43], where early disengagement and dropout from the therapy program can mean users may not get the desired benefits. Thus, approaches to design an engaging online experience to sustain use and ensure beneficial health outcomes have become a deliberate focus for (HCI) research and development. This often means increasing opportunities for: (i) interactivity; (ii) personalized experiences; and (iii) social support through a community of (peer) moderators, trained supporters, or remote therapists [21, 50, 70, 79].

To aid engagement with online therapy, the involvement of a human supporter (e.g., via text messages) has especially been shown to lead to more effective outcomes than unsupported interventions [40, 89, 103]. However, existing research on the effectiveness of supported-interventions has primarily assessed the impact of support duration and frequency [32, 45, 61, 96]; and to a lesser extent, different types of supporter behaviors [36, 69, 87]. Thus, it is less well understood how supporter behaviors impact program use, and therapeutic outcomes; and how this may differ between clients. Having a
more nuanced understanding of the impact of supporter behaviors on clients however could help to: better maximize the effect and outcomes of supporter involvement, assist in supporter training, and thus, increase quality of care for clients.

Simultaneously, the rise in the uptake of internet-delivered therapies and increase in the scale of automatically collected usage data from these treatment programs enables new methodological possibilities for improving our understanding of human behaviors and optimizing health outcomes (e.g., [6, 42, 56]). Specifically, the fields of data mining and machine learning (ML) provide advanced computational methods to construct robust systems that can automatically learn from behavioral data [91]. These techniques have been successfully used in gaming and for recommender systems; and show great potential for advancing our understanding of health data [28] and to assist in the context of mental health (e.g. [25, 34, 74]).

Our work presents the first application of unsupervised machine learning, and statistical and data mining methods for analyzing complex, large-scale supporter-client interaction data to identify supporter behaviors that correlate with better clinical outcomes. Our analysis is based on a fully anonymized dataset of 234,735 supporter messages to clients (sent by 3,481 supporters to 54,104 clients) from an established iCBT program for depression and anxiety, on the SilverCloud platform (www.silvercloudhealth.com), that delivers treatment with regular feedback messages sent by a human supporter. More specifically, our work makes the following contributions:

1. We describe our approach and the challenges involved in developing computational methods for this analysis. This includes: (i) clustering supporters based on how the support messages they sent to clients correlate with client outcomes; (ii) extracting linguistic features in support messages indicative of supporter behaviors that correlate with “high” outcomes across clients in different contexts or situations (e.g. different levels of usage); and (iii) taking into account co-occurrent patterns of different context variables and individual support strategies, we leverage data mining to identify salient context-specific patterns of support.

2. Our work indicates that concrete, positive, and supportive messages from supporters that reference social behaviors are strongly associated with better outcomes; and that the importance of support strategies can vary dependent on a clients’ specific context (e.g. their mental health, platform use). Based on these findings, we discuss: (i) design implications for personalized support in iCBT interventions; (ii) the need for human-centeredness in health data science; and (iii) ethical considerations for secondary data analysis.

**RELATED WORK**

We begin with an overview of relevant literature that describes existing research on: (i) human support in digital mental health interventions; and (ii) data mining and ML approaches for understanding large-scale mental health behavior data.

**Human Support in Online Mental Health Therapy**

In online mental health interventions, the role of supporters, who can be trainees or therapists, often differs from the responsibilities of a therapist in more traditional face-to-face therapy. While supporters encourage and facilitate client use of an iCBT program, the clients themselves have to learn the necessary self-management knowledge and skills that are collectively the active ingredients of the intervention [81, 89].

**Modalities & Benefit of Human Support in iCBT**

Human support in digital mental health interventions can take various forms, ranging from different communication modes (e.g. email, text, phone, or video chat [16, 51]), to variations in support frequency and duration [32, 45, 61, 96], and supporter behaviors [36, 69, 87]. Most studies on the effects of human supported iCBT apps, programs or platforms, assess the therapeutic or working alliance —a bond that facilitates collaborative engagement in therapy [9]—between supporters and clients. The research suggests that such an alliance can be formed within iCBT environments [65] with consistent evidence of the benefits of support in those interventions [40, 82, 89, 103]. For example, Richards & Timulak [82] studied what clients identified as helpful or hindering in their treatment, and found that clients rated the helpfulness of supporter behaviors equal to the core content of the intervention. The literature however is less conclusive on how differences in communication medium [51], frequency [16] and duration [96] of support impact outcomes. For example, Titov [96] found no difference between interventions with low-intensity support (<3 hours) and high-intensity support (>3 hours).

**Human Support Behaviors & their Effectiveness in iCBT**

To date, only a small number of works have explicitly studied iCBT support behaviors and their impact on client outcomes. This includes quantitative and qualitative content analysis of therapist emails to clients receiving iCBT for depression [36], anxiety [69] or bulimia [87]. Here, Sanchez-Ortiz et al. [87] analysed 712 emails and found that 95.4% of all support emails included supportive comments, but little cognitive or behavioral guidance (<15%). Paxling et al. [69] studied 490 emails and found four support behaviors to positively correlate with module completion: task reinforcement —making positive references to what a client has already done or achieved in the program; self-efficacy shaping —prompting clients to engage in learned health promoting behaviors; task prompting —encouraging clients to complete the activities of the CBT program; and empathetic utterances —conveying an understanding of the person’s suffering or life situation. Task reinforcement was further correlated with better client outcomes; whilst deadline flexibility (e.g. therapists postponing tasks) correlated negatively. Similar to task reinforcement, Holländare et al. [36] found (analysing 664 emails) affirming and encouraging behaviors (e.g., validating and praising what the client did) most associated with immediate or longer-term improvements in outcomes, alongside therapist self-disclosure.

While previous research showed that the presence of a supporter correlates with better therapy outcomes, studies on the effectiveness of supporter behaviors remain sparse (cf. [36, 89]). Thus, there is a need for a deeper understanding of how supporter behaviors —as manifest in their online communications with clients—contribute to beneficial clinical outcomes.
Understanding Large-Scale Mental Health Behavior Data

Recent years have seen a growth in research exploring ML for mental health behavior data as captured by wearable health trackers [29]; mobile phones [10, 20, 22, 101, 104, 23]; social media [12, 44, 62, 68, 85]; or electronic healthcare records (EHR) [2, 98]. For the wealth of data that can be collected by these technologies, the fields of ML and data mining provide computational methods that can help improve our understanding of human behaviors and predicting or optimizing clinical outcomes [42]. Frequently applied methods that are particularly relevant to the approach taken in this paper are: clustering, text-mining and association rule mining (ARM).

Clustering & Text Mining in (Large-Scale) Mental Health Data

Clustering is an unsupervised ML technique to identify behavioral patterns in data through commonalities in each data piece; it is often used to identify features in unstructured data (e.g. [2, 14, 68]). In our research, two types of work are particularly relevant. Firstly, to better understand the behaviors of therapists engaged in text-based (SMS) counseling, Althoff et al. [3] clustered therapists, based on client outcomes, into groups of ‘more’ and ‘less’ successful counselors; and then compared how their communications differed using linguistic analysis. We followed a similar approach. Secondly, to identify support behaviors in thousands of anonymous supporter messages, we employ text-mining, which uses natural language processing (NLP) techniques to extract linguistic features or topics from large-scale text. In mental health, text-mining has been used to better understand discussion topics in online mental health communities [62, 68]; to study mental expressions [85], the receipt of social support [90], or cognitive change [74]. Few works seek to specifically aid moderators of online support communities in their work practices, e.g. by identifying helpful and unhelpful comments [12, 44]. Outside of social media, text mining is used to predict suicide risks from SMS messages [63], suicide notes [71], and EHRs [2] to aid care provision. Extending this body of work, we seek to better understand mental health support through a linguistic analysis of supporter messages as part of an iCBT intervention.

Using Association Rules for Behavioral Pattern Mining

Similar to clustering, association rule mining (ARM) is a common data mining technique for extracting behavioral patterns in data (e.g. [13, 23, 66, 104]). Here, the focus is on discovering interesting relations between variables in large data sets such as how patterns of certain data characteristics (e.g. client opinions, types of symptoms, demographics) relate to desirable outcomes (e.g. help-seeking behaviors, clinical score) [31, 66, 106]. We will show how we adapted an ARM algorithm to extract patterns of context-specific best practices of support.

THE I-CBT INTERVENTION

SilverCloud is an established iCBT platform for the treatment of depression, anxiety, and functional impairments. Its development builds on both HCI [21] and clinical research, including randomized controlled trials that evidence the clinical effectiveness of offered treatment programs [83]. In this paper, we focus on one of its most frequently used programs: treatment for depression and anxiety. Accessed online or via mobile, the program presents a self-guided intervention of seven core psycho-educational and psycho-therapeutic modules that are delivered using textual, video and audio contents as well as interactive activities, tools, quizzes and personal stories. Clients work through the program content at their own pace and time, with the recommendation to complete one module each week.

To encourage engagement and continued use, clients receive support from a trained supporter in the form of weekly reviews throughout their treatment journey. The supporters are graduate psychologists with further training in low-intensity interventions that are CBT based, including iCBT. Their support involves writing feedback messages to the client on their work, which usually takes 10-15 minutes to complete. Finally, to assess and monitor clients mental health throughout treatment, clients also complete clinical questionnaires each week, including the PHQ-9 for depression [47] and GAD-7 for anxiety [53]. Overall, the service aims to increase reach and provide effective care for those living with mental illness.

Frequency & Format of Supporter Interactions

Supporters typically review clients’ work on a weekly basis over the 6-8 week treatment period. This serves to manage client expectations of support being restricted to certain times as opposed to immediate 24/7 availability. To this end, supporters can only see clients’ activities and messages on, or after, the agreed review day. To review clients’ progress, supporters have access to information about clients via usage metrics. These show: completed clinical scores of PHQ-9 and GAD-7; any client messages to the supporter; and how many content pages the client viewed, tools they used, and times they logged into the system. For each of these usage metric items, supporters can expand the view to retrieve more details about the content the client reviewed, and their responses in completing interactive tools. Clients have full transparency on what information is shared with their supporter [21].

In response to this client information, supporters compose a personalized feedback message. To facilitate this, they can select and adapt a messaging template from a drop-down menu within a specific supporter interface. These templates tend to be written by the supporters in their own words, and are then specifically tailored to each clients’ situation. During training, supporters learn to personalize messages. Following prior research and guidelines this involves: referencing the clients name and things they did or said with a specific focus on activities of task reinforcement and task prompting [69]; encouragement [36, 82]; guidance and advice [82]; and effective communication using simple language and explanations [75]. As a final step, supporters can bookmark existing, and unlock additional therapy contents on the platform; and they select a subsequent review date. Once their message is submitted via the system, clients receive a notification email. They can view their supporter message when next accessing the program, at which time they will be automatically prompted to complete a new set of clinical questionnaires.

Dataset Description

Our dataset consists of information about: the supporter feedback messages; and the number and types of interactions that
a client had with the platform (e.g., how many CBT content pages they viewed); as well as the number and types of information that clients shared with their supporters (e.g., number of journal entries, tool use data) in the time before and after each feedback message. Across the review period, we also have a subset of clinical scores indicative of the symptoms of the clients’ depression (PHQ-9) and anxiety (GAD-7) before and after a supporter message. For our sample, initial PHQ-9 scores indicated that 32% of clients had minimal-to-mild symptoms of depression, 30% were moderate, 23% moderately-severe, and 15% severe. For GAD-7, initial scores showed that 36% of clients had minimal-to-mild symptoms of anxiety, 31% were moderate, and 33% severe. Typically, each client is assigned only 1 supporter, but if that supporter becomes unavailable, they may be assigned a new one. Table 1 contains basic dataset statistics.

### Table 1: Overview of basic dataset statistics.

<table>
<thead>
<tr>
<th>Dataset statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Supporters</td>
<td>3,481</td>
</tr>
<tr>
<td>Clients</td>
<td>54,104</td>
</tr>
<tr>
<td>Clients with &gt;1 supporters</td>
<td>5,967</td>
</tr>
<tr>
<td>Messages</td>
<td>234,735</td>
</tr>
<tr>
<td>Messages with Pre &amp; Post PHQ-9 &amp; GAD-7</td>
<td>77,006</td>
</tr>
<tr>
<td>Average message length (in words)</td>
<td>191</td>
</tr>
<tr>
<td>Average message length (in sentences)</td>
<td>9.5</td>
</tr>
</tbody>
</table>

To protect full anonymity of both clients and supporters, only non-person identifiable, high-level interaction data was used for the analysis. For clients this included numbers, types and frequencies of interaction events, and aggregates of clinical scores. For supporters this meant the number, frequency and length of each feedback message. Features extracted from message texts were restricted to coarse-grained linguistic structures to preserve anonymity. This matched the terms and conditions of the service, and user consent, which permits the analysis of anonymous data for research purposes, and to improve the effectiveness and service tools of the treatment platform.

Our research employs ML and data mining methods to better understand what support strategies (e.g., use of encouraging words) characterize supporter messages that are correlated with better clinical outcomes for clients. As a first step, this requires us to identify what constitutes ‘successful support messages’ based on clinical outcomes. To this end, we next describe: (i) how we defined clinical outcomes as change and improvement rates in clients over time; and (ii) then use these measures in clustering to achieve three clusters of supporters whose messages correlate with either ‘high’, ‘medium’, or ‘low’ success rates in improving clients’ clinical scores; we use the ‘high’ and ‘low’ clusters in further analysis.

### IDENTIFYING CLUSTERS OF SUCCESSFUL SUPPORT

To better understand how different support strategies in support messages correlate with better clinical outcomes for clients, we need to identify what constitutes an appropriate ‘outcome measure’ in relation to supporter messages. While clients are typically asked to complete PHQ-9 and GAD-7 questionnaires immediately following the receipt of a new supporter message, clients can complete the questionnaires at any time before their next scheduled supporter review. Further, for some supporter messages, an impact on client outcomes may not be apparent immediately (week-on-week); and instead requires a focus on how outcomes develop over time.

To measure clinical outcomes over a period of time, Randomized Controlled Trials (RCTs) tend to assess the difference between the clients’ initial and final clinical scores [97, 41]. While feasible in study set-ups where clients complete an ‘exit survey’ after a fixed period of time, this is more complicated for real-world observational data that can include client dropout. Further, client clinical outcomes can be highly dependent on their situation (e.g., loss of a loved one), symptomatology (e.g., seasonal or cyclic depression), personality or learning style, meaning that some clients will not improve despite the use of good support strategies in supporter messages. However, good strategies used consistently with a set of clients, should result in improvement for the majority of clients in that set. Hence, in our analysis, we focus on the ‘actors’ (supporters) who employ the support strategies in their messages, instead of the ‘receivers’ (clients). To this end, we propose computing message-level and client-level clinical outcomes for each supporter in our dataset; we describe this next.

### Change & Improvement Rates as Clinical Outcomes

We compute the following clinical outcomes by averaging post-message change in PHQ-9 and GAD-7 clinical scores across the messages sent by each supporter. Clients should complete these questionnaires in between messages. Messages with incomplete before or after questionnaires are excluded.

1. **Message-level Change (MC):** The clinical score after a message is highly dependent on the clinical score before the message, as clients that have more severe symptoms before the message also tend to improve more on an average after the message. Hence, we measure Message-level Change as the difference between actual change and the expected change given the client score before the message. That is, for each supporter $S$ with $NM$ messages, compute:

$$\frac{1}{NM} \sum_{r=1}^{NM} (actual\_change_m - expected\_change_m)$$

$$actual\_change_m = score\_before(m) - score\_after(m)$$

$$expected\_change_m = score\_before(m) - \mathbb{E}(score\_after(m)|score\_before(m))$$

2. **Message-level Improvement Rate (MR):** If Message-level Change $> 0$, then the client improved more than expected post-message, and we label the message as “improved (1)”. Otherwise, we label the message as “not improved (0)”. For each supporter $S$ with $NM$ messages, we average these labels across all messages to compute this outcome.

3. **Client-level Change (CC):** While MC captures changes in clinical scores across all messages by a supporter, CC normalizes these changes across all clients of the supporter. For each supporter $S$, we first compute the MC for each client of $S$ separately using the messages that $S$ sent to them. Then, we average the MCs per client across all clients of $S$ to get CC. E.g., if a supporter sends 6 messages to client A whose change is $+1$ after each messages and 4
messages to client B whose change is always 0, the MC will be $\frac{6}{6} = 1$ while the CC will be $\frac{6(6) + 0(4)}{2} = 1.5$. Thus, MC can be high even when only a few clients improve, whereas CC will only be high when these improvements are consistent across all/many clients. This makes CC more robust to a single client’s changing situations or symptoms.

4. **Client-level Improvement Rate (CR):** For each supporter $S$ with clients $NC$, we first compute the average Message-level Improvement Rate using messages $S$ sent to each client separately, and then sum these rates across all clients and divide by the total number of clients.

### Clustering Supporters Based on Support Outcomes

When computing the above 4 outcome measures for both PHQ-9 and GAD-7, we achieve 8 outcome measures per supporter. As a next step, we use these 8 measures as ‘features’ for clustering. We apply $K$-means clustering to obtain $K=3$ clusters of supporters whose messages are generally linked with either ‘high’, ‘medium’ or ‘low’ improvements in client outcomes. The number of k=3 clusters was determined by plotting the sum of squared distances of samples to their closest cluster center against the number of clusters, and visually inspecting the elbow obtained. We hypothesize that there are differences in the support messages sent by supporters in the ‘high’ versus ‘low’ outcome clusters; and that these differences will help us identify what may constitute more effective support strategies.

Table 2 reports the summary statistics of the three obtained clusters of supporters. Figure 1 shows the mean values of all 8 outcomes measures in these clusters, along with the 95% bootstrapped confidence intervals as error bars. Given the mean values of the outcomes and the narrow 95% confidence intervals, we can see that our clustering has reliably divided the supporters into 3 robust groups where these outcomes are typically: high, medium, and low. We did additional statistical analysis that confirmed the results in Figure 1 and showed how, independent of clients initial clinical scores, the differences in mean PhQ-9 and GAD-7 scores between the high and low clusters was significant ($p<0.05$).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>#Supporters</th>
<th>#Clients</th>
<th>#Messages</th>
<th>#Messages Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>438</td>
<td>11068</td>
<td>42734</td>
<td>14519</td>
</tr>
<tr>
<td>Medium</td>
<td>767</td>
<td>31789</td>
<td>123303</td>
<td>42740</td>
</tr>
<tr>
<td>Low</td>
<td>393</td>
<td>10828</td>
<td>47023</td>
<td>14266</td>
</tr>
</tbody>
</table>

Table 2: Statistics for the ‘high’, ‘medium’ and ‘low’ outcome clusters. Only supporters with >9 (Median value) labeled messages for PHQ-9 and GAD-7 were included in the clustering.

### IDENTIFYING SUCCESSFUL SUPPORT STRATEGIES

As a next step, we want to identify what semantic or linguistic strategies are frequently associated with ‘high’ client outcomes. For this purpose, we analyze the differences between the messages sent by supporters in the ‘high’ outcomes cluster and those in the ‘low’ outcomes cluster. We are interested in identifying support strategies that occur significantly more often in the messages of supporters in the ‘high’ outcome cluster. Further, this difference needs to be consistent across different client contexts; meaning that the result has to be independent from variations in individual client context variables such as: the extent to which a client engages with the iCBT program (ContentViews), shares content with their supporter (Shared), the sequence number of the message i.e. the number of messages received so far plus one: MessageNumber, or clients’ current mental health state (CurrentPHQ-9 and GAD-7).

### Methodology: <Strategy> Across <Context> Bins

We are mindful that the actions of supporters, as manifested in their feedback messages to clients, present a direct response to what they know about their clients’ situation (e.g. symptom severity, level of platform use). To disentangle the clients’ context from, whilst understanding the role that context can play in the use of, specific support strategies, we decided to first divide the messages of supporters in the ‘high’ and ‘low’ outcome clusters into different ‘data bins’ for each of 5 client context variables. This allows us to compare the differences in strategies found in support messages of the ‘high’ and ‘low’ clusters separately for different client contexts; and thereby assess, if identified significant differences between the two groups are consistent across, and independent of, variations in client contexts. We found this approach to be more feasible than the use of context as a control variable in linear regression, which due to the large size of the data sample frequently resulted in statistically significant results, but whose effect sizes were difficult to interpret.

Next, we detail on: (i) the client context variables that we defined; (ii) semantic or linguistic features we extracted from supporter messages as strategy variables; and (iii) how we...
analyzed what semantic or linguistic support strategies were salient in the ‘high’ versus ‘low’ support clusters across various bins of each client context variable.

**Client <Context> Variables**

For each message i of every client, we extracted 5 context variables that describe the client’s situation just before the message i.e. between the (i−1)th and ith messages:

1. **ContentViews**: Number of content views i.e. number of times the client clicked on a module or topic in the iCBT program (10 Bins1: 0 views, 8 bins of 1-80 views in increments of 10 views, and >=81 views).

2. **Shared**: Number of data instances shared with the supporter i.e. the total number of tools and journal entries shared, and messages sent to the supporter (4 Bins: 0 shared instances, 3 bins of 1-15 in increments of 5. We never had >= 16).

3. **MessageNumber**: clients are expected to receive up to 8 reviews. For the message number i we have 9 bins (8 bins for i = 1 to 8, and 1 bin for i >= 9).

4-5. **CurrentPHQ−9** and **CurrentGAD−7**: Clinical scores measuring the client’s depressive and anxious symptoms, as seen by the supporter before the ith message. Higher scores indicate more severe symptoms (7 Bins1: score <= 9, 5 bins for scores 10-19 in increments of 2, and score >= 20).

**Semantic or Linguistic <Strategy> Variables in Messages**

Similar to text-mining approaches for online mental health content analysis [12, 44, 62, 68, 85], a lexicon-based approach was used to extract comprehensive, high-level language characteristics from the supporter messages without risking to identify any text content. For each supporter message i, NLP techniques were used to extract 23 features indicative of a support strategy (e.g., use of encouraging phrases). The features were achieved by mapping each word in the message i to a word category of a lexicon. We defined 23 strategy variables based on the literature and available lexicons:

1-3. **Sentiment**: To capture the overall sentiment of a message, we used the NRC Emotion Lexicon [57] to extract the percentages of positive (Pos) and negative words (Neg), and the difference between them (Pos − Neg). Word percentage were used instead of absolute word counts, to better compare messages of different lengths [92].

4-11. **Emotions**: As an indicator of the emotional tone of a message, we extracted the percentages of words related to the 8 emotion categories of the same lexicon [57]: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust.

12-13. **Pronouns**: We extracted the percentage of first person plural pronouns (e.g. We, us, our), assuming that supporters will use these to convey a supportive alliance [9]. Assuming second person pronouns (e.g. you) are used more often to direct clients to engage in specific program activities (e.g., task prompting behaviors [69]), we calculated the difference between the percentages of first person plural and second person pronouns (We − You) as indicator of supportive alliance.

1 Increments of bins were heuristically chosen by rounding up half the standard deviation, after excluding the first and last bins.

(14) **Encouraging Phrases**: Based on a series of support message examples used in supporter training, we derived a list of 15 commonly used encouraging phrases (e.g. ‘good job’, ‘well done’). As an indicator of the motivational tone of a message, we calculated the ratio of the number of encouraging phrases to the number of sentences overall (Encouragement : Sentences).

(15-22) **Mental Processes and Behaviors**: We used the Regressive Imagery Dictionary [54], an extensively used lexicon in mental health research (e.g. to analyze psychotherapy dialogue [80, 37, 93]), for analyzing different categories of conceptual thought in text. Specifically, the analysis includes the seven categories of secondary cognition that are logical, reality-based, and focused on problem solving. This includes the percentages of words related to mental processes of: Abstraction (e.g., know, may, thought), InstrumentalBehavior (e.g., make, find, work), MoralImperative (e.g., should, right, virtue), Order (e.g., simple, measure, array), Restraint (e.g. must, stop, bind), SocialBehavior (e.g., say, tell, call), and TemporalReferences (e.g., when, now, then). As additional behavioral cues, the percentage of ActionVerbs in the text were also extracted, as specified by the Moby Project [73].

(23) **Quantity**: As a measure of quantity of support, we assess the length of the text messages via number of words (WordCount) and report number of sentences where relevant.

**Analysis of Strategies Across Contexts**

We identify a semantic or linguistic support strategy as successful, if: (i) it occurs significantly more often in messages by supporters in the ‘high’ outcome cluster (compared to the ‘low’ outcome cluster); and (ii) this difference is consistent across different client contexts. We operationalize this definition by considering each <context, strategy> pair separately in our analysis. Since we have extracted 5 context variables and 23 strategy variables, we analyze 115 <context, strategy> pairs. For each pair, we do the following:

1. We divide the messages from the ‘high’ and ‘low’ supporter clusters into multiple bins according to the client’s context before the supporter composed a message.

2. For each bin, we then compute the mean of the strategy variable for composed messages in the two clusters separately along with their 95% bootstrapped confidence intervals.

3. For each bin, we also compare the means of the strategy variable across the two clusters using a bootstrapped resampling test.

4. Relevant support strategies should have statistically significant differences in means between the two clusters (α = 0.05) and 95% confidence intervals that rarely overlap, across most bins.

Since the messages in each bin can belong to the same client or supporter, they are not independent. To address this, we use hierarchical bootstrapping for clustered data by randomly sampling supporters with replacement, and their clients and reviews without replacement 10000 times [78]. For comparing the means of the messages in each bin across the two clusters, we reject $H_0$: Difference in means = 0 (two-tailed) at level
We also found that if and only if a valid confidence interval with coverage probability $1 - \alpha$ does not cover 0 ($\alpha=0.05$ in our analysis).

While only messages with pre and post PHQ-9 and GAD-7 scores were used in the initial clustering, we used all messages by supporters in the ‘high’ and ‘low’ clusters for this analysis, so as to not miss out on clients with low levels of usage.

**Results: Strategies Used in Successful Messages**

What follows are the main findings of significant differences in the linguistic strategies that were used in supporter messages associated with ‘high’ versus ‘low’ outcomes (improvements in client clinical scores) across different client contexts.

**Sentiment and Emotion in Support Messages**

Across all supporter messages in our dataset, the sentiment analysis showed that positive words were generally used more frequently than negative words ($Mean_{pos} = 6.2\%, SD_{pos} = 2.8$; $Mean_{neg} = 1.7\%, SD_{neg} = 3.4$). $Mean_{pos-neg} = 4.5\%, SD_{pos-neg} = 1.3$). Further, more successful supporter messages consistently used more positive and less negative words. This effect remained consistent (i.e. $p < 0.05$ for all 3 sentiment strategy variables across all bins of the 5 context variables).

We also found that more successful messages had less occurrences of negative emotions conveying sadness and fear than less successful messages (i.e. $p < 0.05$ for these two emotion-related strategy variables across all bins of the 5 context variables). In addition, more successful messages used more frequently words that expressed joy, yet the difference in means was non-significant for several bins. Since anger and disgust rarely featured in the messages, we did not include them in further analysis, and there were no statistically significant results for any of the other emotions.

**Pronouns: Sense of Supportive Alliance**

Our results show that, across all messages, second person pronouns (e.g. you) were used more frequently than first person plural pronouns (e.g. we, us, our). However, we found that more successful messages consistently employed first person plural pronouns more frequently than less successful messages, and had greater differences between uses of first person plural and second person pronouns (i.e., $p < 0.05$ for both pronoun strategy variables across all bins of the 5 context variables). Consistent with previous work, first person plural pronoun use may reflect a sense of supportive affiliation [30] or increased presence in computer-mediated communication [46].

**Encouraging Phrases: Motivational Tone**

To assess the motivational tone of a message, we calculated the ratio of number of common encouraging phrases to sentences (Encouragement : Sentences). Across all dataset messages, the mean ratio is 0.04 ($SD = 0.09$). Given the average length of a support message of 9.5 sentences, we can estimate that an encouraging phrase is used, on average, once in every 2.6 messages. We found that more successful messages consistently contained significantly more encouraging phrases compared to less successful messages ($p < 0.05$ for this strategy variable across all bins of the 5 context variables except the 61-70 bin for ContentViews).

**Mental Processes & Behaviors: Action Orientation**

For our mental process variables, we found that more successful messages consistently employed more words associated with social behavior and less words associated with abstraction, when compared to less successful messages ($p < 0.05$ for these strategy variables across all bins of the 5 context variables). In order to better interpret the results of our analysis, we visualize each of these <context, strategy> pairs. Figure 2 (left) shows the percentage of words associated with SocialBehavior that are used in ‘more’ and ‘less’ successful support messages. More specifically it plots the mean of this strategy variable for all messages in each bin for the ‘high’ and ‘low’ outcome clusters on the Y-axis, and for each bins of the ContentViews context variable on the X-axis. The error bars are 95% bootstrapped confidence intervals of the means. Some of the most frequently used social behavior words were: help, call, discuss, and share.

Figure 2 (right) further shows how ‘more’ successful messages contained on average less words associated with the strategy variable Abstraction than ‘less’ successful messages; and this finding was consistent independent, e.g., of the clients’ depressive symptoms prior to the supporter review (cf. the bins of the CurrentPHQ — 9 variable). Frequently used abstraction words were: think/thought, know, understand, and learn.

**Quantity: Length of Supporter Messages**

Finally, we found that messages were on average 9.5 sentences long ($SD = 6.7$) and included 191 words ($SD = 147$). We found that more successful messages were consistently
shorter \(\text{Mean words} = 177, \text{Mean Sentences} = 9\) than less successful ones \(\text{Mean Words} = 214, \text{Mean Sentences} = 10.4\); with \(p < 0.05\) for Word Count across all bins of the 5 context variables with very few exceptions). This surprisingly contradicts previous findings by [3], and thus, requires further research.

CONTEXT-SPECIFIC PATTERNS OF SUPPORT SUCCESS
In the previous section, we identified support strategies that were consistently associated with better clinical outcomes across 5 types of contexts, which we each had treated in isolation. In this final analysis, we want to better understand the more complex relationship that likely exists between the use of support strategies and different client context variables. In other words, how may considering the combination of multiple context variables (rather than each client context variable by itself) shift how salient a specific support strategy is in messages associated with either ‘high’ or ‘low’ client outcomes. We believe that identifying such relationship patterns could enable a more effective tailoring of support strategies to specific client contexts. Next, we therefore describe our approach to identifying the salience of a rule \(A \rightarrow B\) is measured as the difference between percentage of strategy \(B\) occurring across the two clusters, when the client is in context \(A^*\), such that more salient rules are used more frequently by supporters in either the more or less successful cluster. We select a subset of rules to interpret by first choosing interesting contexts and strategies that are used in at least one rule with salience \(>= 0.20\), and then selecting all rules associated with them (giving us 1584 rules associated with 66 contexts and 8 strategy variables with 3 levels each\(^4\)), allowing us to interpret the most salient rules in context.

Results: Salient Context-Specific Support Strategies
Figure 3 shows a heat map of the salience of these rules. Next, we present a few examples of interesting results. Taking the third row from the bottom, read horizontally, we see the rules associated with low WordCount. The row is mostly green, apart from the first 6 context rules. The green indicates that the support strategy “WordCount = low” is more salient in more successful support messages, with a darker shade indicating higher salience. We see that for almost all contexts, a lower word count is more salient in more successful messages. However, this is flipped for the first 6 context rules that show no client engagement prior to a review \(\text{e.g. } \text{SharedContent} = 0, \text{ContentViews} = 0\). Thus, for disengaged clients, writing shorter messages is more strongly associated with less successful outcomes (shown as color pink).

Outside of ‘salience flips’ between the two groups (color change); interesting patterns can also be identified through variations in the shade of the same color. For example, for strategies “Fear = low” and “We = high,” we see high salience (dark green) in more successful messages for the same first 6 client contexts; and reduced effects thereafter. This means that writing messages with less words related to fear, and more first person plural pronouns are stronger associated with more successful support messages, and this effect is particularly salient in situations where clients are disengaged.

Aligned with our previous results, we further find for highly engaged clients that successful messages reference more social behavior (SocialBehavior\%=high) and less abstraction (Abstraction\%=low), while less successful messages reference less social behavior (SocialBehavior\%=low or medium) and more abstraction (Abstraction\%=high) in the same contexts. There remain more rules and patterns to unpack. Here, we see our analysis as contributing initial insights on how identified support strategies can be more or less salient or successful depending on a specific client context. So far, our results imply that for less engaged clients, writing longer, more positive and more supportive reviews is linked with

3Messages from the ‘high’, ‘low’, and ‘average’ clusters were used for discretization.
4Salience = \(\text{abs}(P(B|A=\text{high})-P(B|A=\text{low}))\)
5Rules containing ActionVerb\% were excluded, because action verbs occur too frequently and are not very informative

Extracting Multidimensional Context \(\rightarrow\) Strategy Patterns
To identify patterns, we needed to discover associations between client contexts containing two or more contextual variables, and each support strategy. To find associations among multiple items in datasets, we used the well-known frequent item set mining Apriori algorithm. It first generates a set of frequent items that occur together, and then extracts association rules that explain the relationship between those items. Given two sets of items \(A\) and \(B\), an association rule \(A \rightarrow B\) exists if \(A\) and \(B\) frequently appear together in transactions. In our case, “transactions” are supporter messages that occur in client context \(A\) and employ strategy \(B\). We consider each strategy separately in various client contexts, such that context \(A\) is multidimensional (it contains two or more of the 5 context variables) while strategy \(B\) will always be one singular strategy variable\(^2\).

We extracted association rules separately for both our ‘high’ and ‘low’ outcome clusters; and then compared the rules that occurred in both groups using two “interestingness” measures: Confidence and Coverage \([33, 49]\). Confidence is the percentage of transactions containing \(A\) that also contain \(B\) \(P(B|A)\), whereas coverage is the percentage of transactions containing \(A\) \(P(A)\), chosen as a measure of the applicability of a rule. For example, if a context \(A\) occurs very frequently (e.g. coverage is high), rules associated with it are more interesting as they can be applied more often when recommending support strategies personalized to the client’s context. We set minimum coverage to 0.01 and minimum confidence to 0 to extract a large number of rules, that we can then filter using a different metric that compares the salience of the rules.

For this analysis, we excluded the strategy variables: Anger, Disgust, Restraint, Order, and MorallImperative, as they rarely occurred in our data. The remaining 18 strategy variables were discretized to ‘high’, ‘medium’, and ‘low’, using equal frequency binning\(^3\). Then, Apriori was applied separately on messages from our ‘high’ and ‘low’ outcome clusters. We found 22599 rules in both clusters. The salience for each of these rules was calculated as the absolute confidence difference between the two clusters. That is, the salience of a rule \(A \rightarrow B\) is calculated as the absolute confidence difference between the two clusters. This surprisingly contradicts previous findings by [3], and thus, requires further research.

\(^2\)The arules R package was used to set up Apriori in this way [38].
greater outcomes; whilst more engaged clients appear to benefit more from messages with less negative words, less abstraction, and more references to social behaviors.

DISCUSSION
Our work presents the first application of unsupervised ML, and statistical and data mining methods to analyze complex, large-scale supporter-client interaction data of an established iCBT intervention for the treatment of depression and anxiety. We focused on developing a better understanding of how the behaviors of supporters, who assist clients’ engagement with this service, may correlate with better clinical outcomes for these clients; which presents a largely under-explored area. Below, we discuss the main implications of our work for future research that intersects the fields of HCI, ML and healthcare.

Identifying Effective Context-Specific Support Strategies
We described our approach to identifying ‘more’ and ‘less’ successful support behaviors that are manifest in communications to clients. Using semantic and linguistic feature extraction methods, our results indicate that supporter messages that typically achieve higher client outcomes contain more words that are positive, supportive, related to social behaviors, and less abstract; and those messages tend to be shorter than less successful messages. Largely, these findings align well with previous qualitative studies of iCBT support that emphasize the prevalence of supportive language [87], and importance of affirmations and encouragement [36, 69] for client outcomes. Extending this research in iCBT, our work presents novel findings of how the success of identified support strategies can vary dependent on a specific client context. Next, we discuss how having a better understanding of each persons’ context for support enables new opportunities for personalized care; which, in turn, can improve client engagement with iCBT interventions and benefit their mental health outcomes.

Data-Enabled, Personalized Mental Health Support
So far, only few works have explored the design space and use of ML to personalize the treatment or delivery of (mental) health interventions (e.g. [24, 67, 60]). Most prominently,
Paredes et al. [67] applied reinforcement learning for recommending stress coping interventions tailored to the person. Other recent trends include the development of just-in-time adaptive interventions (JITAs) that utilize algorithms to provide the right type or amount of support, at the right time, in response to an individuals’ internal and contextual state [60].

**Design Implications for Personalized Human Support in iCBT**

For guided iCBT interventions, there are multiple ways in which gathered data insights about context-specific support strategies can inform supporter interface design. For example, as supporters review a client, they may be presented in their feedback interface with recommendations of what strategies specific to this clients’ situation may be strongly correlated with successful client outcomes, providing them additional input to their feedback practices. To help translate linguistically-derived support strategies more directly into feedback messages, strategy-related words (e.g. positive words, certain pronouns) could be highlighted in real-time as part of the message editor that supporters use. Especially for training purposes and to support skills acquisition of novice supporters, it may also be helpful to integrate examples of ‘support messages that were successful for similar clients’ as guidance.

**Human-Centered ML in (Mental) Healthcare**

The above design examples further illustrate our orientation to human-centred ML and the integration of data insights into healthcare practice. Rather than promoting the use of templates or automatizing client feedback away from the human supporter, we suggest designing interventions that seek to enhance supporter agency by enabling them to personalize their feedback more effectively for each person (cf. [94]), and to better understand how their actions make (ideally) a positive difference to their clients. Thus, while advanced data tools that can identify complex patterns are often seen to generate more accurate, objective and less biased insights (cf. [34, 98, 105]), it is important that we (i) do not take away from, but help foster, the important relationship and genuine human connection that is formed between supporter and client and crucial to their alliance and positive outcomes [9, 58]; and (ii) ensure supporters feel that their input and expertise is valued rather than made redundant or replaced in favor of data science.

**Understanding (Big) Data in Digital Health Interventions**

Next, we discuss identified challenges and opportunities for working with complex, large-scale observational data.

**Trade-Offs between Data Access and Use & Ethical Conduct**

Although we had unprecedented, privileged access to large-scale supporter-interaction data, we made necessary trade-offs as to what kind of analysis we could conduct to protect the full anonymity of both client and supporter data. This meant much of our analysis was restricted to coarse-grained linguistic features and high-level usage data. While our research captured individual word associations, the use of other linguistic features such as n-grams (e.g. [2, 44]) could expand the analysis to word pairings, or even short word sequences that could enable a richer contextual understanding of identified support behaviors. At the same time, however, such explorations need to be done with care so as to not risk making too much content, or the people who produced it, identifiable.

Caution is also required for secondary data analysis for which additional user consent is likely unfeasible to collect (for every subsequent analysis). As is common in ML approaches for mental health, user privacy should be carefully addressed to preserve anonymity (cf. phone data [10, 101], sensors [77], social media [68, 85, 105]), and analysis should occur in a context where there is a clear case for the prospective benefits that could arise, e.g., from improved healthcare provision (see public health ethics [15] and recent work on user acceptance of digital phenotyping [52, 84]). Going forward, we need to continue developing feasible, privacy-preserving data methods; and, as researchers, need to remain critical of, and sensitive to, the extent to which our data analysis is really ‘justifiable’ with regards to how it comes to benefit users and health services.

**Deriving Insight from Data: Interpretation & Future Directions**

Due to the necessary data use restrictions, we acknowledge that derived data insights — whilst novel — are limited to the definitions chosen, and require further research to validate. Future work may also explore multiple additional data mining avenues, including: (i) analysis of supporters’ use and adaptation of messaging templates (cf. [3]); (ii) studies into sequential routines of strategies (e.g. using Inverse Reinforcement Learning [7]); (iii) supporter clustering using ‘engagement’ as an outcome metric alongside clinical improvement (cf. [59]); or (iv) the combination of support strategies, supporter-features, and client engagement features to predict clinical outcomes.

Despite manifold possibilities for data mining, challenges remain for ensuring that derived insights are both human interpretable (e.g. [1, 35, 72]) and clinically useful. While we were deliberate in our choice of data tools and visualizing of their results to create representations that are comprehensive to laypeople, other research methods (e.g. qualitative studies [36, 69, 87]) are needed to help contextualize, validate and advance our understanding of support, or other data-derived health behaviors. More research is also needed to develop our understanding of the potential value that these new types of data insights could bring to actual healthcare practices.

**CONCLUSION**

Aiming to understand how the behaviors of iCBT supporters impact client outcomes, we presented our ML approach and the analysis of 234,735 supporter messages sent to an unprecedentedly large clinical sample of 54,104 clients with a mental health diagnosis. Using various computational methods we identified support behaviors associated with ‘more’ or ‘less’ improvements in clinical scores, and showed how their salience varied dependent on different client contexts. Our work enables a better understanding of best practices, and opens-up new opportunities for personalizing support provision within computer-delivered mental health services. We discussed: the implications of our findings for the design of iCBT supporter interfaces; the need for a human-centered approach for sensibly integrating data insights into healthcare practices; and ethical considerations for secondary data use.

**ACKNOWLEDGEMENTS**

We thank our broader team: James Bligh, David Carter, Dessie Keegan, Aditya Nori, Tim Regan, Sebastian Tschiatschek.
REFERENCES


[33] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. Data mining: concepts and techniques. Elsevier.


