Early detection of high-risk users in a digital peer-to-peer mental health support platform

Sourav Das

s.das@uni-kassel.de

University of Kassel

John Naslund

Harvard University

Ben Locke

Togetherall

Article

Keywords:

Posted Date: May 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9553769/v1

Additional Declarations: No competing interests reported.

Early detection of high-risk users in a digital peer-to-peer mental health support platform

Sourav Das (Corresponding author)
University of Kassel, Kassel, Germany
s.das@uni-kassel.de

Ben Locke
Togetherall, London, United Kingdom

John A. Naslund
Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, United States

Abstract

Digital peer support platforms have emerged as scalable solutions to mental health problems, but they face critical challenges in identifying users at risk of escalation, a situation that requires urgent clinical or administrative intervention. This study investigates whether user-level characteristics, rather than textual content alone, can be leveraged for early detection of high-risk cases.

We analyzed a dataset with N=178,612 users from Togetherall, a clinically moderated, anonymous online peer-support platform, covering demographics, health background, and engagement patterns (189 predictors). Escalation events (occurring in 1% of users) were modeled as a binary classification outcome. Using machine learning methods such as Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost we compared predictive performance under different feature strategies, including the full feature set, LASSO-based selection (aggressive and moderate), and factor analysis.

Results showed that ensemble models, particularly XGBoost (F1 ≈ 0.863) and Random Forest (F1 ≈ 0.855), achieved the highest predictive accuracy with the full feature set. Moderate LASSO preserved predictive performance (XGBoost F1 ≈ 0.862; Random Forest F1 ≈ 0.863), whereas aggressive LASSO substantially reduced accuracy. Factor analysis offered computational efficiency but with some loss in predictive power (XGBoost and Random Forest F1 ≈ 0.811). Most importantly, even with a highly parsimonious feature set derived from aggressive LASSO (time on site, page views, and age) the models retained reasonably strong predictive power (XGBoost F1 ≈ 0.762; Random Forest F1 ≈ 0.769). This underscores the central role of basic engagement metrics and demographics in identifying escalation risk.

Our findings demonstrate, for the first time, that user-level sociodemographic, health, and engagement characteristics provide strong predictive signals for escalation risk in online peer-support platforms. In particular, time on site, page views, and age emerged as the most predictive factors. Integrating such models into platform moderation systems could enable earlier detection of risks, enhance user safety, and improve cost-effectiveness in delivering digital peer-to-peer mental health support.

Short abstract (150 words)

Digital peer-support platforms provide scalable mental health support, yet timely identification of users at risk of escalation remains challenging. Most existing approaches rely on textual content analysis, which can be resource-intensive and difficult to deploy in real-time moderation systems. We examined whether user-level characteristics alone can enable early detection of escalation risk. We analyzed data from 178,612 users of Togetherall, a clinically moderated online peer-support platform, using 189 user-level characteristics. Escalation events (1% of users) were modeled with machine learning methods, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost, under different feature-reduction strategies. Ensemble models achieved the strongest performance, with XGBoost and Random Forest reaching F1 scores of approximately 0.86 using the full feature set. Moderate LASSO preserved accuracy, whereas aggressive feature reduction reduced performance but still produced meaningful predictive power (F1≈0.76). In particular, time on site, page views, and age emerged as the most predictive factors.

1. Introduction:

Mental health problems represent a major public health burden worldwide, particularly among young people. The World Health Organization (WHO) reported that one in seven adolescents experience a mental disorder (WHO, 2021). At the same time, barriers such as stigma, high treatment costs, and limited service availability mean that large segments of the population do not receive adequate support (Patel et al., 2018). Against this backdrop, digital mental health platforms have emerged as a promising way to extend access to support at scale (Torous et al., 2021; Torous et al., 2025). These platforms include apps, online counseling services, and peer-support communities.

Despite their growing popularity, digital mental health platforms face important challenges. One such challenge is that because of their widespread accessibility, these platforms will be used by people experiencing the full range of mental health concerns, including serious mental illness and/or risk/distress that requires intervention. Prior research on online peer-support communities (Huh et al., 2013; Milne et al., 2019) showed that while most interactions are supportive and beneficial, the risk of crises that are not detected early remains. The inability to detect users who are at risk or experiencing serious distress can result in a failure to safeguard users in need and place strain on the platform's clinical and moderation resources. Consequently, there is a pressing need to understand the predictors of high-risk cases and to develop early detection strategies that enable timely intervention. The existing literature pertaining to online peer support communities has focused largely on posts and comments by the platform users and relevant textual analysis to predict the risks (Huh et al., 2013; Milne et al., 2019; Coppersmith et al., 2018; Kessler et al., 2017). To our knowledge, no empirical study has attempted to predict escalation risk using user-level characteristics. This is the central focus of the current study, which seeks to address the question of whether user characteristics (demographic, health, and platform usage) can be used for early detection of high-risk members?

To answer this question, this study gathered data from a digital peer-to-peer support platform called Togetherall. This online platform has been providing clinically moderated, anonymous peer support since 2007 (Togetherall, 2022). The platform has served over 600,000 members, primarily across the United Kingdom, United States, Canada, and New Zealand, where its largest user bases are located (Togetherall, 2022). The users are largely drawn from three sectors: college/university students, employees, and public health. Users must be at least 16 years old to join.

Togetherall provides clinically moderated peer-support in an online environment in which people join the platform (members) and interact with the community anonymously (peer support) and use the available tools on the platform (direct messaging, journal, assessments, courses) to improve their mental health. The platform is moderated 24/7 by licensed/registered mental health professionals, known as Wall Guides. In addition to the general tasks of facilitating a large online community, a key responsibility of clinical staff is reviewing "flags". Flags are raised either automatically by the platform's algorithms or manually by users themselves. Flags indicate that there is some content or behavior that requires clinical review and potential moderation. Once a flag is raised, Wall Guides assess the situation, including the content of concern, the context, and the member's overall behavior on the platform. After assessing the situation they make a decision on appropriate actions, if needed. Moderation actions can consist of offering immediate support (via the community or directly to the member), requesting information to clarify the situation, removing/editing content, providing referrals to professional or crisis support, or generally intervening to support the member and community. Some of these cases require risk management attention beyond standard community moderation and are referred to as an "escalation" as the case is escalated through layers of clinical expertise and oversight. Across the Togetherall platform, less than 1% of all members are escalated; however, such cases tend to be resource intensive due to the involvement of multiple clinical staff and, in some instances, require intensive off-platform case-management such as activating emergency services or coordinating with other support systems. As such, the early identification of members who are likely to escalate would enable the platform and staff to intervene earlier while minimizing the added risks and resource requirements associated with an escalation.

This study explores whether user-level factors such as sociodemographic characteristics, health characteristics, and users' platform engagement are associated with escalation events on the platform and whether these characteristics can help inform early detection strategies.

2. Methods

2.1. Study design and ethics

Togetherall offers a moderated, community-based environment where users can seek support for mental health concerns anonymously. The platform systematically reviews content created by users, including member posts, comments, and direct-messages and flags content of concern for clinical review. In situations where a member is experiencing elevated distress, potential risk, or is not following house rules an escalation case is created for advanced clinical review/intervention. This study used anonymized data provided by Togetherall, covering all registered members who joined the platform between November 1, 2021, and February 3, 2025. The dataset included users from all countries where Togetherall operates, namely the United Kingdom, Ireland, the United States, Canada, Australia, and New Zealand. In total, the sample comprised 178,612 members. For each member, 189 predictors were available, spanning sociodemographic characteristics, health characteristics, and users' platform engagement. A detailed description of these variables is presented in Tables 1a–1c and described in Section 3.

The primary goal of this research was to develop and validate predictive models capable of identifying potential escalation cases earlier, thereby supporting proactive interventions and platform improvements. Notably, escalation events are rare, occurring in fewer than 1% of Togetherall users.

The processing and use of data for this research were conducted strictly in accordance with applicable data protection and privacy laws. All data were fully anonymized by Togetherall prior to transfer to the research team for analysis, ensuring that no personally identifiable information was accessible at any stage of the analysis. No attempt was made to re-identify individuals. This study was reviewed and approved by the Harvard Medical School Institutional Review Board.

2.2. Variables

2.2.1. Predictor Variables

Our predictor variables were grouped into three categories:

- Sociodemographic characteristics: This category covered personal attributes and social circumstances of the users, including gender, region of residence, living status (e.g., living alone or with others), ethnicity, work sector (e.g., education, healthcare, services), employment status, and age.

- Health characteristics: This category encompassed a wide range of indicators reflecting users' physical and mental health status, experiences, and sources of support. Specifically, this category included: (i) responses to statements such as "I have no health concerns," "I have a long-term health concern," "I have a disability," or "I am pregnant/recently gave birth"; (ii) self-reported stressors, such as loneliness or isolation, academic pressure (studying or exams), work-related stress, family issues, and other personal challenges; (iii) experiences with mental health difficulties, including thoughts of ending one's life, thoughts of self-harm, low mood or depression, and high levels of stress; and other (iv) self-reported sources of social support, whether primarily from individuals in users' immediate lives (family, friends, colleagues) or from broader networks and external sources.
- Users' platform engagement: This category captured multiple dimensions of user activity on the platform. It included: (i) posting behavior across different channels such as group posts, community posts, and private posts; (ii) commenting activity across the same channels; (iii) viewing behavior, including community dashboard visits, article and self-assessment views, and general page views; (iv) activity within topic-specific categories such as anxiety, depression, stress, relationships, loneliness, and others; and (v) other engagement features, such as setting goals, writing journal entries, updating moods, number of logins, and overall time spent on the platform.

2.2.2. Dependent Variable: Escalation Events

The primary outcome of interest was whether a user experienced an escalation event. Escalations are defined as instances where a member is experiencing elevated/complex distress, appears to be at risk, or is breaking house rules and the situation is not easily de-escalated/clarified by routine intervention (e.g., clarifying the meaning of a post through direct-messaging). In the event that a situation cannot be de-escalated/clarified, a case is opened by senior clinical staff who will then seek to work with the member to either clarify/de-escalate or confirm hand-off to the next best resource. As such, escalations can be thought of as "intensive off-platform case management" and may take the form of crisis-evaluation, referrals to emergency or other professional services or local authorities.

2.3. Analytical Framework

Our analytic approach was guided by a risk prediction and early-warning framework developed in the public health literature, with an emphasis on identifying risks at the earliest possible stage.

For example, Simon et al. (2018) and Coppersmith et al. (2018) applied machine learning models to predict suicide attempts, while Divya et al. (2021) used a similar approach for the early detection of Alzheimer's disease. Building on this framework, we conceptualized escalation events as a rare and risky outcome. Accordingly, we framed the task as a binary classification problem, coded as 1 when an escalation occurred and 0 otherwise, and implemented the following machine learning models described in the sections below.

2.3.1. Logistic Regression

The Logistic Regression model estimates the probability of the positive class as a function of a set of feature variables. Following Agresti (2018), Logistic Regression model estimates the conditional probability of the positive class as:

$$\tau(x) = \frac{exp(\beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_p x_p)}$$

Here, $x_1, x_2 \ldots\ldots x_p$ represent the vector of input features. β_0, $\beta_1 \ldots\ldots \beta_p$ are the model parameters (coefficients). $P(y = 1|x)$ denotes the probability that the outcome y equals 1 given the features. The predicted probability $\tau(x)$ takes values in the range [0,1]. The coefficients β are estimated via maximum likelihood. Logistic Regression is a widely used method for binary classification problems.

2.3.2. Support Vector Machine (SVM)

The SVM aims to find an optimal separating hyperplane that maximizes the margin between two classes in the feature space (Cortes & Vapnik, 1995). The distance from the hyperplane to the closest training instance is known as the margin. The hyperplane that maximizes the margin is the maximum-margin. That is why SVM minimizes the following function:

$$\frac{1}{2}\omega^2 + C\left(\sum_{i=1}^{l} \xi_i\right)$$

Subject to, $y_i(\omega x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$

Here, ω is the weight vector defining the orientation of the hyperplane, ξ_i are slack variables, and C is the regularization parameter.

2.3.3. Random Forest

Random Forest classifier generates T bootstrap samples by sampling with replacement. For each bootstrap sample, the algorithm grows a decision tree by splitting the data at different points (Breiman, 2001). But instead of looking at all available features to decide the best split at each node, it randomly selects a smaller number m of features (where $m < p$, and p is the total number of features). Each tree makes a prediction for an observation x. The individual prediction from the t-th tree is:

$$h_t(x) \in \{0,1\}$$

The final prediction is obtained by taking the average predictions of all these individual trees:

$$\hat{p}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x)$$

2.3.4. XGBoost

XGBoost constructs an ensemble of decision trees in a sequential manner, where each tree attempts to correct the residual errors of the previous trees (Chen & Guestrin, 2016). The main objective of XGBoost is to minimize the following function:

$$L(\phi) = \sum_{i=1}^{n} l(y_1, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k)$$

Where, $l(y_1, \hat{y}_i)$ is the loss function, and $\Omega(f_k)$ is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$$

Where, T is the number of leaves in the tree, ω represents the vector of leaf weights, γ is a penalty parameter for the number of leaves, and λ is the regularization parameter.

2.4. Performance matrix

After running the predictive models described above, the following performance metrics were calculated. Higher values of the performance metrics indicate better model performance.

Precision (Prec): Precision measures the proportion of correctly predicted positive cases among all instances predicted as positive.

Precision = TP/ (TP + FP)

where TP denotes true positives and FP denotes false positives.

True Positive Rate (TPR): TPR also called recall, measures the proportion of actual positive cases.

TPR = TP / (TP + FN)

Where FN denotes false negatives.

True Negative Rate (TNR): TNR, or specificity, measures the proportion of actual negative cases.

$$TNR = TN / (TN + FP)$$

Where TN denotes true negatives.

F1 Score: F1 score is the harmonic mean of precision and recall, balancing the trade-off between the two.

$$F1 = 2 \times Precision \times TPR / (Precision + TPR)$$

Accuracy (Acc): Accuracy measures the overall proportion.

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

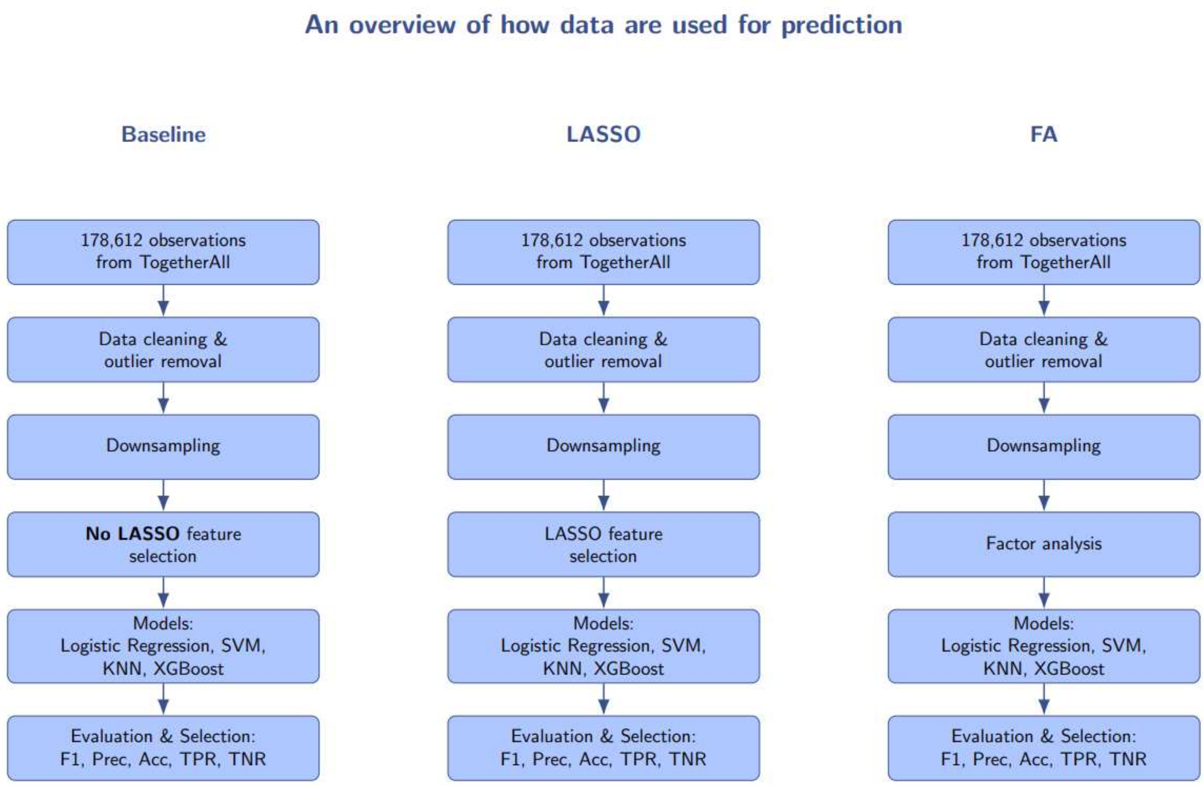
2.5. Downsampling, feature selection and dimensionality reduction

One important consideration with the escalation outcome is that it is highly imbalanced: only about 1% of the cases experienced escalations, while the remaining 99% are non-escalations. Such imbalance poses a challenge for building a reliable predictive model. To address this, we applied a downsampling approach (Susan & Kumar, 2020), reducing the majority class (non-escalations) to create a more balanced dataset.

In the next step, we proceeded with feature selection and dimensionality reduction. High-dimensional feature spaces (in our case, 189 features, see the descriptive section for details) can lead to overfitting. LASSO (Least Absolute Shrinkage and Selection Operator) can shrink the coefficients of less relevant features toward zero, effectively retaining only the most informative features. As an alternative, factor analysis reduces the original variables to a smaller set of latent factors, without giving a reduced set of most informative features. After implementing these downsampling, feature selection/dimensionality reduction we proceeded with the above stated predictive models.

2.6. Overview

A brief overview of the workflow is given below:



All data management, preprocessing, descriptive analysis were conducted using STATA (version 15). Predictive modeling was conducted using Python (version 3.12).

3. Descriptive statistics

The overall sample consisted of 178,612 user-level observations collected from the Togetherall platform. Descriptive statistics for this sample are presented in Tables 1a through 1c. A summary of these statistics is provided below.

3.1. Sample characteristics

3.1.1. Sociodemographics

Table 1a presents the summary of users' sociodemographic characteristics. Most Togetherall members were female (68.15%) and the average age was 29.26 years (SD=11.98). Geographically, most members were based in United Kingdom (54.3%), followed by the United States (29.27%) and Canada (15.19%). Only a small proportion of users were from Australia (0.22%) and New Zealand (1.02%). In terms of living arrangements, members reported a variety of statuses: 16.11% live alone, 19.04% live with friends, 22.56% live with a partner, and 17.02% live with parents. The remaining members reported other living situations such as living with a partner and children, as a single parent, with extended family, experiencing homelessness, or other unspecified arrangements.

Regarding ethnicity, most members identified as White (62.62%), followed by Asian (14.98%) and Black/African/Caribbean (8.68%). Other ethnic identities such as Hispanic/Latino, Australian, Aboriginal Australian, Australian Asian, and Mixed or multiple ethnicities were reported in smaller proportions. With respect to employment sectors, 57.44% of members were affiliated with the higher education sector, 18.83% with health services, 8.81% with private organizations, 6.62% with further education, and 4.44% with local authorities. Smaller proportions work in the charity sector (2.28%) or the armed forces (1.58%). In terms of work status (with multiple responses permitted), nearly half of the members identified as students (46.05%). An additional 28.94% were employed full-time, 24.68% part-time, and 11.68% were unemployed. A minority of members identified as caregivers (2.49%), retired (1.14%), or unable to work due to health or other reasons (4.40%).

3.1.2. Health characteristics

Health characteristics were categorized into five distinct domains: (1) overall health status, (2) sources of stress, (3) mental health-related experiences, (4) psychological difficulties, and (5) perceived support networks. Table 1b presents the descriptive statistics of these five distinct domains.

Regarding overall health status (with multiple responses permitted), 36.49% of members reported no health concerns, 12.62% indicated having a long-term health condition, and 7.67% reported living with a disability. In terms of sources of stress (also allowing multiple responses), 32.64% identified loneliness or isolation as a key stressor, followed by studying or academic exams (31.92%), work-related stress (24.50%), family-related stress (23.24%), and financial concerns such as money or debt (23.00%). Additional stressors included experiencing a traumatic event (21.61%), grief or loss (16.38%), friendship-related difficulties (18.95%), and relationship challenges with a partner (18.07%). A minority of participants also reported other stressors such as relocation/moving, parenting responsibilities, discrimination, bullying, and pregnancy (all of which were endorsed by fewer than 10% of respondents).

Concerning mental health experiences, 28.40% of participants reported having thoughts of ending their life or self-harming, and 27.35% reported experiencing thoughts of self-harm specifically. Moreover, 19.92% reported high levels of stress, 13.30% reported experiencing panic attacks, and 13.18% indicated engaging in self-harming behaviors. Less commonly reported experiences (each endorsed by fewer than 10% of participants) included having attempted suicide, feeling at risk from someone else, contemplating harming another person, and engaging in harm toward others.

Participants also reported experiencing various psychological difficulties, with the most common psychological difficulties are feeling down or depressed (47.85%), stress (47.62%), feeling nervous or on edge (44.36%). A significant proportion also reported frequent unwanted thoughts (32.81%), and anxiety in social situations (32.62%). Additionally, issues related to attention span (24.33%) and memory (19.03%) were noted. Sleep difficulties affected 24.24% of participants, while 23.73% struggled with concerns related to food and body image. Panic attacks were experienced by 21.76%, and the same percentage reported upsetting dreams or memories. Depression or low mood was identified by 15.15% of participants as a major issue.

Finally, participants were asked about their sources of support (multiple responses permitted). Overall, 49.6% reported receiving support from their immediate social network, including family, friends, and coworkers, while 10.26% specifically relied on family support. A minority of respondents (each endorsed by fewer than 10% of participants) reported receiving support from various sources, including their college or university, a therapist or specialist mental health practitioner, a doctor or psychiatrist, a community mental health team, an employee assistance program, and, in some cases, specifically from a psychiatrist.

3.1.3. Users' platform engagement

Table 1c presents descriptive statistics on Togetherall members' engagement with the platform. Overall, users generated an average of 3.80 content creation activities (SD = 11.96).[1] On average, members recorded 64.06 general page views. We also included login frequency and time spent on the platform. On average, Togetherall members logged in 3.38 times (where 4,258, or 2.38 percent of members did not log into the site at all) and spent approximately 34.75 minutes on the site.

Finally, we report the top activity categories used by Togetherall members. Table 1c presents data on top 5 activity categories out of 38. Among these, anxiety and depression emerged as the most common categories, accounting for 14.77 percent and 14.35 percent of total posts, respectively. Other prominent categories include stress (11.11 percent), loneliness (9.36 percent), and relationships (8.82 percent).

3.1.4. Primary outcome: Escalation behavior

Escalation behavior was observed in 1% of cases (1,760 of 178,612), indicating that the vast majority of Togetherall members did not escalate. Among those who did, most escalated only once. Multiple escalation events were rare (only 7.90 percent escalation cases), with a maximum of five recorded in the dataset. Escalation is the primary outcome variable in our predictive analysis. For modeling purposes, escalation was treated as a binary variable, where a value of 1 indicates that a user experienced at least one escalation, and 0 indicates no escalation.

[1] The variable Content creation is constructed as the row-wise sum of all user-generated activities on the platform, including group posts, community posts, private posts, brick posts, group comments, community comments, private comments, brick comments, goals, journal entries, and mood updates.

4. Results

As outlined in the main workflow, we adopted three strategies for predicting escalation cases: (i) using the full set of 189 variables, (ii) applying LASSO-based feature selection, and (iii) employing a factor-analysis based dimensionality reduction technique instead of LASSO. For LASSO, we tested two configurations. First one is an aggressive approach and the second one is a more moderate approach.[2] Table 2 reports the performance of all tested models, ranked by F1 score within each feature strategy.

In the first approach, we used the complete set of 189 features to build the predictive models. With the full feature set, ensemble methods clearly outperformed the others. XGBoost achieved the highest F1 score (≈0.863), precision (≈0.843), and balanced sensitivity (TPR ≈ 0.883) and specificity (TNR ≈ 0.836), indicating strong capacity to detect both escalation and non-escalation cases. Random Forest performed comparably (F1 ≈ 0.855, TPR ≈ 0.880), while Logistic Regression (F1 ≈ 0.808) and SVM (F1 ≈ 0.797) offered competitive but slightly lower recall.

Given the large number of features and the possibility that some are correlated or have minimal predictive power, the next step considered the LASSO feature selection technique. We began with an aggressive LASSO approach (using elastic net), which retained only three features (time on site (minutes), page views, and age). After developing predictive models based on these three features, performance dropped sharply across all algorithms. The best-performing model, Random Forest, achieved F1 ≈ 0.769, with a 0.086 point drop compared to its full-feature counterpart. Other models, including XGBoost (F1 ≈ 0.762) and SVM (F1 ≈ 0.737), exhibited similar declines. Logistic Regression was particularly affected (F1 ≈ 0.679). Such decline in model performance is not surprising. The aggressive selection retained only three predictors, thereby excluding variables that could have contributed to the models' predictive power.

[2] By moderate LASSO, we refer to an elastic net specification that allows a relatively wide range of regularization strengths ($C \in [10^{-2}, 10^{3}]$) and moderate L1 penalties (l1_ratio $\in [0.2, 0.5]$). In contrast, aggressive LASSO denotes an elastic net specification that restricts regularization to substantially stronger penalties ($C \in [10^{-5}, 1]$) and lower L1 ratios (l1_ratio $\in [0.1, 0.3]$), thereby inducing greater coefficient shrinkage and sparser solutions.

In the next step, instead of applying aggressive LASSO, we used a moderate LASSO (elastic net) feature selection, which retained 173 of the 189 predictors.[3] This approach preserved nearly all predictive capability while modestly reducing dimensionality. Random Forest (F1 ≈ 0.863) and XGBoost (F1 ≈ 0.862) matched full-feature performance. Logistic Regression (F1 ≈ 0.809) and SVM (F1 ≈ 0.804) also performed better than in the aggressive LASSO case.

Finally, to address the large feature set, we also applied factor analysis to reduce the predictors to seven latent factors based on the scree plot (Figure 1). This approach resulted in moderate performance degradation for all models (XGBoost and Random Forest F1 ≈ 0.811) compared to full-feature counterpart. This means factor analysis substantially reduced dimensionality and computational cost, however, compressing complex, non-linear relationships into orthogonal factors resulted in some loss of predictive accuracy.

Overall, the predictive modeling results highlighted four key insights:

1. Ensemble tree-based methods (XGBoost, Random Forest) consistently outperform.
2. Aggressive LASSO harms performance by discarding too much relevant information.
3. Moderate LASSO preserves accuracy while slightly reducing dimensionality.
4. Factor analysis offers substantial simplification but at the cost of predictive power, particularly for non-linear models.

In addition to overall model performance, the feature selection strategies revealed interesting contrasts. Moderate LASSO retained the majority of predictors and therefore did not yield a sharply reduced set of variables. By contrast, the aggressive LASSO approach produced a distinct subset of key predictors, namely time on site (minutes), page views, and age. Although model performance under this specification was weaker than with other approaches, relying solely on these three predictors still resulted in reasonable predictive outcomes. Specifically, Random Forest achieved an F1 score of ≈0.769, XGBoost ≈0.762, SVM ≈0.737, and Logistic Regression ≈0.679.

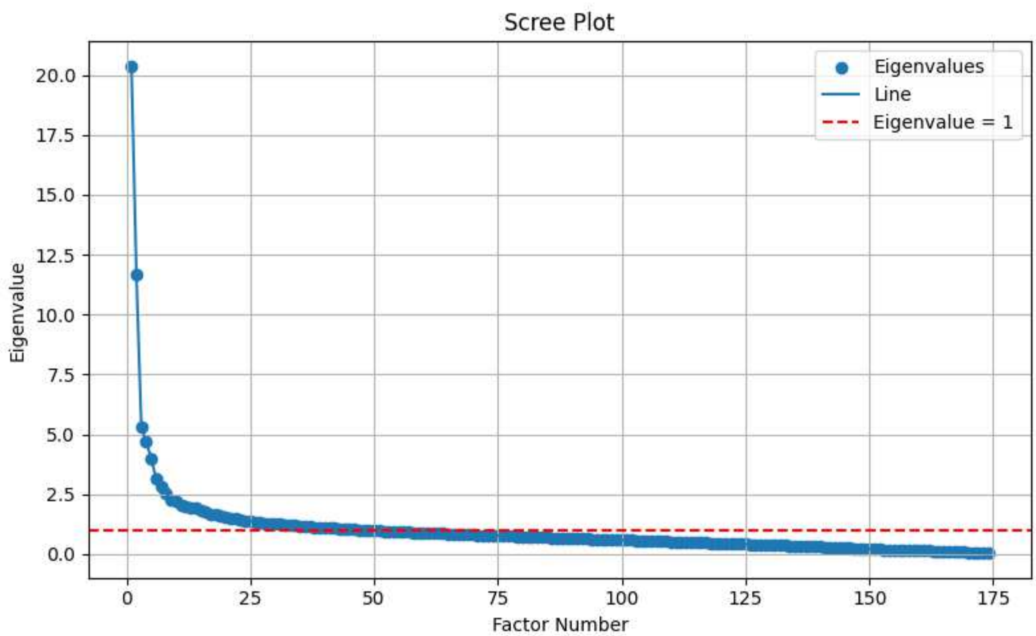
Table 2: Classification metrics for all models tested ranked by F1 score

Model	F1	Prec	Acc	TPR	TNR	Features included

[3] We experimented with different elastic net specifications by tuning the regularization parameters. However, the model consistently selected either a very sparse set of predictors (3 variables) or an almost full set (173 variables).

Baseline: all features						
XGBoost	0.863	0.843	0.859	0.883	0.836	189
Random Forest	0.855	0.832	0.851	0.880	0.822	189
Logistic Regression	0.808	0.823	0.812	0.793	0.830	189
SVM	0.797	0.784	0.793	0.811	0.776	189
Applying LASSO (aggressive)						
Random Forest	0.769	0.778	0.769	0.760	0.778	3
XGBoost	0.762	0.759	0.758	0.764	0.752	3
SVM	0.737	0.796	0.752	0.686	0.820	3
Logistic Regression	0.679	0.819	0.723	0.580	0.868	3
Applying LASSO (moderate)						
Random Forest	0.863	0.852	0.859	0.873	0.845	173
XGBoost	0.862	0.854	0.859	0.871	0.847	173
Logistic Regression	0.809	0.805	0.806	0.813	0.799	173
SVM	0.804	0.831	0.808	0.779	0.838	173
Applying factor analysis						
Random Forest	0.811	0.812	0.804	0.810	0.797	7
XGBoost	0.811	0.804	0.802	0.818	0.785	7
SVM	0.797	0.807	0.792	0.786	0.797	7
Logistic Regression	0.772	0.811	0.774	0.737	0.815	7

Figure 1: Scree plot of eigenvalues from factor analysis



5. Discussion

This study examined whether user-level sociodemographic, health, and engagement characteristics can help in the early detection of escalation events (members who may be at risk) in a digital peer-to-peer mental health support platform. Our findings suggest that these characteristics can predict escalation events with high accuracy using machine learning binary classification methods. To the best of our knowledge, this is the first study to predict escalation events in online peer-to-peer mental health support platforms using user-level sociodemographic, health, and engagement characteristics.

Models using the full feature set (189 variables) achieved the strongest performance, with XGBoost and Random Forest identifying approximately 88% of all escalation cases while maintaining high specificity, implying that fewer than one in five would be false positives. Very interestingly, using LASSO feature selection (aggressive approach) this study identified a distinct subset of key predictors (using aggressive LASSO)—time spent on the platform (in minutes), number of page views, and age (the strongest predictors of escalation events). Despite relying on only three features, XGBoost and Random Forest models still identified approximately 76% of escalation cases with high specificity. This is particularly important for practical implementation as early detection of escalation risk can be achieved at minimal computational and data-collection cost.

There are, however, some related studies that provide useful points of comparison. Similar to our study, Milne et al. (2019) developed a predictive model for online support forums, but their focus was on identifying posts requiring moderator response through automated triage rather than predicting escalation events. Their approach relied on rule-based and algorithmic classification of message content, whereas our approach leverages user-level features across multiple domains. Along similar lines, Huh et al. (2013) applied text-based classification methods to an online diabetes community to detect posts that warranted moderator attention, again without distinguishing escalation events per se.

Work from clinical and crisis-intervention contexts provides additional parallels. For instance, using social media posts Coppersmith et al. (2018) built machine learning models to detect those who are at risk of suicide. Additionally, Kessler et al. (2017) demonstrated that ensemble methods outperform regression-based models in suicide risk prediction using large-scale health datasets, which completely aligns with our finding. Taken together, our findings contribute to this growing literature by demonstrating that not only textual or linguistic markers but also user-level sociodemographic, health, and engagement characteristics provide meaningful predictive signals for escalation events. This creates opportunities for targeted moderation and support, while preserving the low-cost, scalable nature of digital peer-to-peer support platforms.

Our study makes novel contribution to existing research on the Togetherall platform. Prior studies have primarily focused on platform engagement using conventional regression-based approaches. For example, Marinova et al. (2022) examine predictors of user engagement and symptom change among adolescents using regression models. Gordon et al. (2021) analyze predictors of the nature of user engagement using regression analysis. Thomson et al. (2024) advance the field conceptually by developing a stakeholder-informed Theory of Change that explains how digital peer-support platforms should function. Against this backdrop, our study shifts the focus from engagement and conceptual frameworks to predicting escalation events, using machine-learning models applied to a large-scale user dataset (N = 178,612). We provide actionable insights that can support early identification of escalation cases in digital peer-to-peer mental health platforms.

Our findings have three important implications for the use, effectiveness, and safety of online peer-to-peer support platforms. First, online peer-to-peer support platforms can integrate such predictive models into their existing moderation systems. This will allow them to detect individuals at risk of escalation earlier and take necessary measures to support prevention or early response. Importantly, this approach complements prior qualitative evidence showing that moderators play a central role in monitoring risk and facilitating supportive interactions (Deng et al., 2023), by equipping them with data-informed signals that help prioritize cases. Second, early detection of escalation has the potential to improve the timeliness of support and ultimately, reduce burden on individuals and contribute to better mental wellbeing. Evidence from related domains, such as early-warning systems for suicide risk, also suggests that predictive models can support earlier identification of high-risk individuals (Zheng et al., 2020). This is an essential prerequisite for timely intervention and care. Third, such data-informed moderation strategies can enhance the cost-effectiveness of maintaining such platforms (Buntrock, 2024).

This study has two important strengths. First, it adopts a comprehensive modeling strategy by comparing multiple machine learning algorithms across three distinct feature-selection and dimensionality-reduction approaches. This multi-strategy design enhances the robustness and credibility of the findings. Second, the study leverages rich user-level sociodemographic, health, and engagement data rather than relying solely on text-based or post-level content. In doing so, it demonstrates that escalation events can be detected with high accuracy even in the absence of text-based or post-level content analysis.

At the same time, two main limitations of our study should be acknowledged. First, our dataset is cross-sectional, meaning that each individual's activation (baseline user-level) characteristics are observed only once.. Since health and engagement characteristics are inherently dynamic, this design does not allow us to capture the temporal changes in users' behaviors on the platform. Second, our analysis relies solely on user-level characteristics and does not incorporate the content of users' posts or comments, which could provide additional insights into escalation dynamics. Future research should therefore focus on collecting longitudinal data to better capture the evolving nature of users' health and engagement, and on integrating textual and behavioral data from posts and comments to develop more comprehensive predictive models.

This study advances the field of digital mental health by demonstrating that escalation events in online peer-to-peer support platforms can be predicted with high accuracy using routinely collected user-level sociodemographic, health, and engagement data, even in the absence of text-based content analysis. From a practical perspective, these findings can improve the design and operation of online peer-support platforms by enabling the integration of escalation-identification models into platform workflows. Predictive models based on user-level characteristics can be integrated into existing moderation workflows to flag potential escalation cases earlier and more systematically, enabling timely support, triage, or referral while reducing reliance on intensive manual monitoring. Importantly, the identification of a small set of key predictors such as time spent on the platform, page views, and age—suggests that effective early-warning systems can be implemented using minimal data inputs, enhancing feasibility of identification. Future research should focus on validating these models across different platforms and populations, examining the real-time performance of these models in live settings.

Our findings are also relevant in the context of the rapidly expanding use of artificial intelligence (AI) in digital mental health. A key challenge for AI-based moderation systems in digital mental health services is the reliable detection and mitigation of risk (Wang et al., 2025). Our results can complement existing AI approaches that primarily rely on unstructured data, such as language or sentiment analysis. In particular, the models presented here achieve consistently strong predictive performance using user-level characteristics alone, suggesting their potential value as an additional layer of risk detection that is less sensitive to linguistic variation. When combined with AI-driven analyses of qualitative data, such approaches may contribute to more robust and reliable risk-detection systems in digital mental health platforms. This highlights important directions for future research, to evaluate how integrating user-level risk models with unstructured AI signals affects real-world detection accuracy and downstream safety outcomes in digital peer-to-peer support mental health platforms.

Overall, this study demonstrates that escalation events in digital peer-to-peer mental health support platforms can be predicted with high accuracy using user-level characteristics. These findings contribute novel evidence to the digital mental health literature and offer practical pathways to enhance platform safety and effectiveness.

References:

Agresti, A. (2018). *An Introduction to Categorical Data Analysis*. Wiley.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Buntrock, C. (2024). Cost-effectiveness of digital interventions for mental health: Current evidence, common misconceptions, and future directions. *Frontiers in Digital Health, 6,* 1486728. https://doi.org/10.3389/fdgth.2024.1486728

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights, 10*, 1178222618792860. https://doi.org/10.1177/1178222618792860

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20,* 273–297. https://doi.org/10.1007/BF00994018

Deng, D., Rogers, T., & Naslund, J. A. (2023). The role of moderators in facilitating and encouraging peer-to-peer support in an online mental health community: A qualitative exploratory study. *Journal of Technology in Behavioral Science, 8*(2), 128–139. https://doi.org/10.1007/s41347-023-00302-9

Divya, R., Shantha Selva Kumari, R., & the Alzheimer's Disease Neuroimaging Initiative. (2021). Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification. *Neural Computing and Applications, 33*(14), 8435–8444. https://doi.org/10.1007/s00521-020-05596-x

Gordon, D., Hensel, J., Bouck, Z., et al. (2021). Developing an explanatory theoretical model for engagement with a web-based mental health platform: Results of a mixed methods study. *BMC Psychiatry, 21,* 417. https://doi.org/10.1186/s12888-021-03391-z

Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics, 46*(6), 998–1005. https://doi.org/10.1016/j.jbi.2013.08.011

Kessler, R. C., Stein, M. B., Petukhova, M. V., Bliese, P., Bossarte, R. M., Bromet, E. J., Fullerton, C. S., Gilman, S. E., Ivany, C., Lewandowski-Romps, L., Millikan Bell, A., Naifeh, J. A., Nock, M. K., Reis, B. Y., Rosellini, A. J., Sampson, N. A., Zaslavsky, A. M., Ursano, R. J., & Army STARRS Collaborators. (2017). Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Molecular Psychiatry, 22*(4), 544–551. https://doi.org/10.1038/mp.2016.110

Marinova, N., Rogers, T., & MacBeth, A. (2022). Predictors of adolescent engagement and outcomes: A cross-sectional study using the Togetherall (formerly Big White Wall) digital mental health platform. *Journal of Affective Disorders, 311*, 284–293. https://doi.org/10.1016/j.jad.2022.05.058

Milne, D. N., McCabe, K. L., & Calvo, R. A. (2019). Improving moderator responsiveness in online peer support through automated triage. *Journal of Medical Internet Research, 21*(4), e11410. https://doi.org/10.2196/11410

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & Unützer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet, 392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X

Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., Beck, A., Waitzfelder, B., Ziebell, R., Penfold, R. B., & Shortreed, S. M. (2018). Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry, 175*(10), 951–960. https://doi.org/10.1176/appi.ajp.2018.17101167

Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art. *Engineering Reports, 3*, e12298. https://doi.org/10.1002/eng2.12298

Thomson, M., Henderson, G., Rogers, T., Locke, B., Vines, J., & MacBeth, A. (2024). Digital mental health and peer support: Building a theory of change informed by stakeholders' perspectives. *PLOS Digital Health, 3*(5), e0000522. https://doi.org/10.1371/journal.pdig.0000522

Togetherall. (2022). Togetherall Guardian Council report 2022. https://togetherall.com/en-ca/wp-content/uploads/sites/3/2022/11/2022-Togetherall-Guardian-Council-Report.pdf

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry, 20*(3), 318–335. https://doi.org/10.1002/wps.20883

Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A., & Firth, J. (2025). The evolving field of digital mental health: Current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry, 24*(2), 156–174. https://doi.org/10.1002/wps.21299

Wang, T., Bruckman, A. S., & Yang, D. (2025). The practice of online peer counseling and the potential for AI-powered support tools. *Proceedings of the ACM on Human-Computer Interaction, 9*(2), CSCW191. https://doi.org/10.1145/3711089

World Health Organization (WHO). (2021). *Adolescent mental health*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

Zheng, L., Wang, O., Hao, S., Ye, C., Liu, M., Xia, M., Sabo, A. N., Markovic, L., Stearns, F., Kanov, L., Sylvester, K. G., Widen, E., McElhinney, D. B., Zhang, W., Liao, J., Ling, X. B. (2020). Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Translational Psychiatry, 10*(1), 72. https://doi.org/10.1038/s41398-020-0684-2

Tables

Table 1a
Demographic characteristics (total sample size=178,612)

	Frequency	Percent
Gender		
Male	44,683	25.02
Female	121,715	68.15
Other*	12,199	6.83
Region		
UK	96,987	54.30
US	52,279	29.27
Canada	27,125	15.19
Australia	391	0.22
New Zealand	1,830	1.02
Living Status		
Living alone	28,774	16.11
Living with friends	34,014	19.04
Living with partners	40,287	22.56
Living with parents	30,399	17.02
Other*	45,126	25.27
Ethnicity		
White	111,849	62.62
Asian	26,763	14.98
Black African Caribbean	15,496	8.68
Hispanic Latino	8,112	4.54
Other*	16,392	9.18
Sector		
Higher education	102,596	57.44
Health service	33,635	18.83
Private organization	15,731	8.81
Further education	11,832	6.62
Local authority	7,923	4.44
Charity	4,070	2.28
Armed forces	2,825	1.58
Work Status (multiple response possible)		
Student	82,244	46.05
Full time	51,695	28.94
Part time	44,087	24.68
Unemployed	20,867	11.68
Caregiver	4,453	2.49
Retired	2,035	1.14
Unable to work	7,863	4.40
	Mean (standard deviation)	95% CI
Age	29.26 (11.98)	[29.20, 29.32]

*Note: In the gender category, "Other" includes Trans Man, Trans Woman, Gender Neutral, Gender Fluid, Non-Binary, Genderqueer/Non-Conforming, Trans (Unspecified), and other identities beyond the listed genders. It also includes individuals who did not disclose their gender or whose gender is unknown. In the ethnicity category, "Other" includes Australian, Australian aboriginal, Australian asian, Mixed or multiple etc.. In the living status category, "Other" includes single parents, individuals experiencing homelessness, those living with a partner and children, and other living arrangements beyond the stated categories. It also includes individuals who did not disclose their living status.

Table 1b
Health characteristics (total sample size=178,612)

	Frequency	Percent
Health (multiple response possible)		
I have no health concerns	65,172	36.49
I have a long term health concern	22,549	12.62
I have a disability	13,703	7.67
I am pregnant/recently gave birth	1,136	0.64
Stressor (multiple response possible)*		
Loneliness or isolation	58,294	32.64
Studying or exam	57,008	31.92
Work	43,757	24.50
Family	41,511	23.24
Money or debt	41,085	23.00
A traumatic event	38,596	21.61
Friendship	33,854	18.95
Relationship partner	32,280	18.07
Grief or loss	29,261	16.38
Experiences (multiple response possible)*		
Thought about ending my life	50,732	28.40
Thought about self-harming	48,856	27.35
Experienced stress	35,571	19.92
Experienced panic attacks	23,748	13.30
Self-harming	23,549	13.18
Experiencing problems (multiple response possible)*		
Feeling down or depressed	85,464	47.85
Stress	85,048	47.62
Feeling nervous or on edge	79,241	44.36
Frequent unwanted thoughts	58,597	32.81
Worrying in social situations	58,268	32.62
Coping regulating emotions	53,571	29.99
My attention span	43,449	24.33
Sleep	43,304	24.24
Food and body image	42,392	23.73
Panic attack	38,871	21.76
Upsetting dreams or memories	38,858	21.76
My memory	33,995	19.03
Depression or low mood	27,065	15.15
Supported by (multiple response possible)*		
People in my life eg. family friends coworkers etc.	88,637	49.63
Family	18,329	10.26

*(Note: apart from above stated **stressors**, a minority of respondents reported various stressors, including 9.01% moving, 6.32% parenting, 3.29% discrimination, 3.80% bullying, and 0.60% pregnancy. Apart from above stated **experiences**, a minority of respondents reported following experiences: 6.55% attempted to end their own life, 5.77% felt at risk from someone else, 3.45% contemplated harming another person, and 0.08% engaged in harming someone else. Apart from above stated **experiencing problems**, a minority of respondents reported various personal and psychological challenges: 8.53% exhibited autistic spectrum traits, 8.28% struggled with understanding learning, 7.60% experienced stress at home, 6.53% faced relationship problems, and 5.63% encountered stress at work. Additionally, 5.17% reported having a phobia, 5.16% cited issues related to sex, 4.81% struggled with managing change, and an equal percentage reported feeling excessively happy, high, or hyper. Alcohol-related concerns were reported by 4.53%, while 3.25% mentioned drug-related issues. Bereavement or loss was cited by 3.16%, eating disorders by 2.96%, and psychosis by 1.81%. Furthermore, 1.53% reported a lack of motivation, 1.44% expressed a general sense of uncertainty about their future, 1.30% experienced social isolation, and 0.60% reported issues related to gambling. Apart from above stated **supported by**, a minority of respondents reported receiving support from various sources: 9.53% from their college or university, 8.47% from a therapist or specialist mental health practitioner, 2.92% from a specialist, doctor or psychiatrist, 2.69% from a community mental health team, 1.62% through an employee assistance program, and 1.47% specifically from a psychiatrist.)

Table 1c
Users' platform engagement statistics (total sample size=178,612)

	Frequency	Percent
Activity Category (top 5 by percentage of total posts)		
Anxiety	78820	14.77
Depression	76596	14.35
Stress	59318	11.11
Loneliness	49975	9.36
Relationships	47057	8.82
Escalation		
Escalation=0	176,852	99.01
Escalation=1	1,621	0.91
Escalation=2	117	0.07
Escalation>=3	22	0.01
	Mean (standard deviation)	95% CI
Other member behavior		
Content creation	3.80 (11.96)	[3.75, 3.86]
Number of logins	3.38 (7.80)	[3.34, 3.42]
Page views	64.06 (1320.13)	[57.94, 70.19]
Time on site (minutes)	34.75 (139.41)	[34.11, 35.40]

Author contributions:

S.D. contributed to conceptualization, designed the study, analyzed the data, and wrote the manuscript. B.L. contributed to data acquisition, conceptualization, study design, interpretation of results, and manuscript revision. J.A.N. contributed to conceptualization, study design, supervision, and manuscript revision. All authors reviewed and approved the final manuscript.

Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability:

The data that support the findings of this study are not publicly available due to institutional restrictions, but are available from the corresponding author upon reasonable request and with permission from Togetherall.

Conflict of interest:

S.D., B.L., and J.A.N. declare that they have no financial or non-financial conflicts of interest related to this study.

Ethical statement

The study was formally determined to be "Not Human Subjects Research" by the Harvard Medical School IRB (Protocol # IRB24-1479).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CorrespondenceforIRB241479.pdf