Predicting Adolescent Suicide Risk From Cellphone Usage Data and Self-Report Assessments

Maya Stemmer
The Data Science Institute,
Reichman University
maya.stemmer@post.runi.ac.il

Shira Barzilay
Department of Community Mental Health, University of Haifa
shirabarzilay@univ.haifa.ac.il

Itamar Efrati
The Data Science Institute,
Reichman University
itamar.efr@gmail.com

Talia Friedman
The Data Science Institute,
Reichman University
talia.friedmann@gmail.com

Lior Carmi
The Data Science Institute,
Reichman University
li.carmi@gmail.com

Mishael Zohar
The Data Science Institute,
Reichman University
dr.mishaelzohar@gmail.com

Anat Brunstein Klomek
School of Psychology,
Reichman University
bkanat@runi.ac.il

Alan Apter
Schneider Children's Medical Center of Israel, Tel Aviv University
eapter@clalit.org.il

Shai Fine
The Data Science Institute,
Reichman University
shai.fine@runi.ac.il

Abstract

As suicide is a leading cause of adolescent death, innovative evaluation of imminent suicide risk factors is needed. This study followed high-risk adolescents who presented with recent suicidal thoughts and behaviors (STB) for six months. They were digitally monitored and periodically observed during in-clinic visits. We aimed to classify their STB levels and identify severe cases based on two types of digital monitoring: (1) weekly self-reported questionnaires by patients and (2) continuously collected cellphone usage data. We present a novel approach for utilizing the immense amounts of unlabeled cellular logs in a supervised classification problem. Satisfying prediction results from both data types showed the feasibility of using digital monitoring for STB prediction. Such a capability may enrich periodic clinical assessments with frequent digital follow-ups and raise awareness whenever necessary.

Keywords: Suicide Prediction, Abnormal Behavior Detection, Machine Learning, Digital Monitoring.

1. Introduction

1.1. Background

Suicide is the second cause of death globally among adolescents. Adolescence is a critically vulnerable developmental period for suicidal thoughts and behaviors (STB). Youth Emergency Department (ED) visits for STB increased in many countries, particularly during the COVID-19 pandemic (Barzilay and Apter, 2022). Suicidal crises in adolescents are often brief and episodic, with considerable potential for recurrence (Miller and Prinstein, 2019). Adolescents are susceptible to interpersonal belonging, rejection, and conflict, possibly predisposing imminent STB (Cheek et al., 2020). They have difficulty regulating emotions and employing cognitive control to inhibit problem behaviors, including suicidal acts when facing acute stress (Miller and Prinstein, 2019). Therefore, the rapid-fluctuating suicidal states in adolescents warrant an improved proximal prediction of imminent STB.

Identifying proximal predictors for STB is critical to better understanding *when* individuals are most at risk (Glenn and Nock, 2014). The fundamental problem is to predict whether an individual with well-known pre-existing risk factors is at *imminent suicide risk* (i.e., within the next hours, days, or weeks) (Galynker, 2023). Two pre-suicidal mental state-specific diagnoses have suggested to lead to imminent suicidal thoughts or behavior. The transdiagnostic approach they present not only enables a more accurate and objective assessment of imminent suicide risk but could also facilitate research in developing digital markers of suicide risk (Joiner et al., 2018; Voros et al., 2021).

1.2. Related work

Intensive longitudinal methods (i.e., repeatedly assessing individuals over time) provide the unique

URI: https://hdl.handle.net/10125/106824
978-0-9981331-7-1

opportunity to study proximal predictors of STB within individuals. Recent studies have begun using real-time mobile and wearable devices to predict STB, and the feasibility and acceptability of intensive monitoring in adolescents have been established (Sedano-Capdevila et al., 2021; Rabasco and Sheehan, 2022). Existing efforts mainly focused on using intensive self-report assessments such as ecological momentary and daily diaries. These studies provided valuable information about the temporal dynamics of suicidal ideation (E. K. Czyz et al., 2022; E. M. Kleiman et al., 2017) and significant proximal predictors (Glenn et al., 2022; Al-Dajani and Czyz, 2022). However, the substantial reliance on self-reports requires insight and compliance, and the associated assessment burden limits study durations and applicability to real-world settings.

Passive sensing of mobile data has been proposed as a promising direction for future research (Sedano-Capdevila et al., 2021; E. M. Kleiman et al., 2021). It allows naturalistic and continuous data collection with minimal burden, including social, communication, activity, and sleep patterns, in which change may indicate imminent STB (Morshed et al., 2019). Recognizing objective online and offline social behavior patterns preceding STB may particularly benefit adolescents. Pioneer studies have demonstrated that passive mobile sensing by actigraphy may predict STB (Horwitz et al., 2022; E. Kleiman et al., 2019). Ghandeharioun et al. (2017) showed the usefulness of predicting symptoms in patients with a major depressive disorder through passive data collection from built-in sensors in phones and a wearable device. Liu et al. (2020) demonstrated the feasibility of predicting patients' moods using machine learning (ML) algorithms with privacy-preserving techniques through multimodal data collection from patients' mobile devices. However, to our knowledge, no studies have investigated proximal predictors of youth STB via real-time mobile passive sensing, including clinical and self-report assessments over six months.

Unlike previous research that considered individual risk factors, an ML approach processing passive mobile sensing can examine multiple risk factors and their combination to build superior prediction models (E. Czyz et al., 2021; Lejeune et al., 2022). Combining passive mobile sensing with an ML approach may identify complex real-life patterns and relationships and improve the near-term identification of risk and timely interventions. Mullick et al. (2022) used passively sensed mobile data to predict depression levels in adolescents. Even though their goal and setting were similar to ours, they used aggregative methods to transform phone data into feature vectors instead of the unsupervised clustering method we implemented.

Additional work in naturalistic settings is needed to unlock the potential of active and passive sensing and ML prediction models for suicide prevention (Schafer et al., 2021). Further research is required to develop prediction models optimized for implementation in clinical settings (Wang et al., 2022). In this study, we address these challenges and utilize intensive digital monitoring of self-reports and behavioral data in a real-time and real-world outpatient clinical setting.

1.3. Objectives and contribution

Our research addresses the knowledge gap in real-time proximal predictors of youth STB. It utilizes active and passive digital monitoring in a longitudinal prospective study among high-risk adolescents in a real-world clinical setting. It provides a proof of concept for digital markers of imminent STB to classify adolescents with high-risk STB vs. lower risk, enabling a more accurate and objective assessment of imminent suicide risk. The study integrates multiple digitally derived data sources, including behavioral, interpersonal, and frequent subjective self-reports using comprehensive and non-intrusive techniques.

The paper is divided into two parts describing the main aims of the study. The first aim was to demonstrate the clinical utility of digital weekly self-reports in identifying adolescents at risk for STB in a real-world high-risk clinical setting. The second was to develop an integrated prediction model using passive mobile sensing data for STB prediction. We addressed both tasks as supervised binary classification problems and used ML techniques to determine whether a particular adolescent is at risk. Clinician suicide risk assessments were considered as the gold standard. We wished to predict them from (1) self-reported weekly questionnaires and (2) passively analyzed cellular phone use. Each in-person visit provided a target label, and data from the preceding week was used to obtain its predicted risk level.

Our data were small and highly unbalanced. On the one hand, we had limited numbers of patients and labels from in-person visits overall. On the other hand, we had weekly self-reported questionnaires and immense amounts of unlabeled data describing cellular phone use. Moreover, as typical in suicide prevention research, low-risk STB scores were prevalent in our dataset compared to the less frequent high-risk scores. Extremely high STB scores indicating suicidal behavior beyond thoughts were even more rare. Therefore, we integrated different techniques to deal with our dataset's unique nature and utilize its many unlabeled data.

In the first part of the study, we derived classification features from the questionnaires and used standard classification algorithms to predict the assessment of the proximate visit. In the second part of the study, we designed a novel two-step procedure for detecting abnormal behavior to predict risk levels. We started with an unsupervised learning approach to cluster all patients' cellular use and obtain normal behavior. Then, high-risk events were considered rare, and we predicted risk levels using outlier detection. The predictions were evaluated as a supervised classification problem at this stage, as in the first part. Satisfactory results from both parts showed the potential of using self-reported and passive monitoring approaches as complementary predictors of STB.

The rest of the paper is organized as follows: in section 2, we summarize the experiment's protocol and describe the data that was gathered from the weekly questionnaires and cellular usage. In section 3, we explain the methods used for predicting STB from the weekly questionnaires and in section 4, we explain the methods used for predicting STB from the cellular usage data. Section 5 presents the results of the two tasks, each in a different subsection. In section 6, we discuss the implications of the results, note the limitations of the study, and suggest future research.

2. Experiment and data collection

In this section, we explain the experiment, list the inclusion and exclusion criteria for the study (2.1), describe the two types of data we used - from weekly self-reports (2.2) and passive cell phone logs (2.3), and provide descriptive statistics on the participants (2.4). More details are available in the complete research protocol (Barzilay et al., in press).

2.1. Recruitment and participant breakdown

This study included high-risk adolescents who presented to a pediatric hospital emergency department (ED) with recent STB and were recruited at the Depression and Self-Harm Clinic at Schneider Children's Medical Center of Israel. When eligible potential recruits presented at the clinic, senior research staff were notified and would then meet with the patients to explain the study and obtain signed parental consent and the child's consent (if over 16 years).

Inclusion Criteria were adolescents between the ages of 11-18 presenting with recent suicidal ideation or suicidal behavior, as defined by the Columbia Suicide Severity Rating Scale (C-SSRS). The C-SSRS is a brief, semi-structured interview designed to screen for the presence and intensity of STB (Posner et al., 2011). In research and clinical settings, the C-SSRS is used to determine the level of suicide risk and to inform safety planning. This scale is considered globally the gold-standard in suicide risk assessment in research and clinical practice, and used to determine suicide risk management and care. We measured STB severity via a composite STB score of 0-10 derived from C-SSRS categories by the maximum category present (Nilsson et al., 2013). Another inclusion criterion was possessing an Android mobile phone, as Apple mobiles are not supported by the iFeel data collection app used in this study. Exclusion criteria were acute medical conditions, mental retardation, cognitive impairment, or linguistic limitations that preclude understanding research questions.

A total of 71 adolescents and their parents consented to the study and were guided to download the iFeel app. Once downloaded, the app initiated collecting relevant passively sensed data from the mobile phone and weekly questionnaires. Research assistants trained the participant on feeling the questionnaires, emphasizing the importance of adherence and providing instructions on the use of the subjective rating scale. They also contacted participants during the follow-up period in case of missing active or passive data collection, and assisted in technical or adherence issues.

Participants were invited for extended follow-up evaluations at one, three, and six months following the initial assessment. These sessions included in-person or remote video clinical assessments and self-report assessment scales administered via secured online surveys. In practice, there was variability in the follow-up frequency, as some participants rescheduled or skipped visits altogether without indicating an intent of dropping out of the study. As such, the timing of the four visits was adhered to in most, but not all of the cases, and in most cases the patient had even more than four follow-up clinical visits.

The evaluations performed at the clinic during the patient's follow-up visits were recorded by clinicians both as clinical notes and as a score on the STB scale. This scale ranged from 0 to 9, with 9 being the most severe: 0-no STB, 1-passive suicidal ideation (SI), 2-active SI, 3-SI with methods, 4-SI with intent, 5-SI with a plan, 6-preparatory acts, 7-aborted attempt, 8-interrupted attempt, 9-actual attempt. The STB score calculated from the clinician's assessment at each follow-up meeting was subsequently used as the prediction model's ground truth (label).

Of the 71 recruited participants, 65% (46/71) completed the study. 35% (25/71) dropped out due to various reasons, including switching from Android to iPhone (incompatible with the iFeel app), undue burden of participation, or any other type of parental or participant reluctance to continue in the study.

2.2. iFeel app: weekly self-assessment data

iFeel is an innovative digital health research platform enabling passive and active digital monitoring and providing continuous objective measurements for any disorder. A brief self-report questionnaire was generated by the iFeel app using push notifications once a week. Three items were selected to assess the main risk factors for STB based on the pre-suicidal mental state-specific diagnoses. These include general mood ("In general, how was your mood this week"?), entrapment ("Did you feel like there was no way out"?), and belongingness ("Did you feel lonely?"), rated on a visual analog scale between 1 to 5. The STB items were adapted from the C-SSRS and included assessing suicidal ideation intensity (rated on a visual analog scale between 1 to 5), intent, plan, behavior, attempt, and non-suicidal self-injury. If a participant indicated yes to a suicidal plan, current intent, or behavior, a safety protocol was initiated. Crisis and community services were made available to the adolescent, and a parent was contacted.

The questionnaire was only available for filling out on Tuesdays. However, on that day, participants could fill out the questionnaire multiple times, and all versions of the weekly questionnaire were saved. The questions presented to the user followed a predefined flow. While all participants were asked the first three questions, their answers determined the rest of the questionnaire based on the C-SSRS administration manual. Patients who provided serve answers were presented with follow-up questions exploring the intent and possible immediate danger they might be in. Specific control questions were displayed if the user indicated a clear intention to pursue suicide. These alerted the clinical staff of a possible high-risk situation that required immediate intervention. The diagram in Figure 1 shows the inherent flow of the questionnaire and its questions.

2.3. iFeel app: real-time mobile data

iFeel harnesses mobile phone data to find associations between digital markers and STB ratings obtained via weekly self-reports and clinical assessments during study visits. The passively sensed data is collected continuously, 24/7.

For this study, the data consisted of communication patterns (number of phone calls, duration of incoming and outgoing calls, communication app usage duration), device usage (power on/off, doze mode in/out, number of device screen opens and locks, Wi-Fi connections, Bluetooth connections, battery usage, network mobile on/off, airplane mode on/off), and app usage (name of app and seconds of active app usage). All the collected information had non-identifiable information and fully complied with the General Data Protection Regulation (GDRP). No app-specific data, content, texts, or other sensitive data was collected. Per the study's protocol, data were collected over six months from the intake date.

2.4. Descriptive statistics and train-test split

The adolescent sample included participants aged 11-18, with an average age of 14.5. Female participants accounted for 60% of the participants (43/71), males accounted for 34% (24/71), and transgender/other accounted for 6% (4/71). Hence, females were over-represented in our data set. Females were also more compliant with clinical follow-up visits and digital self-assessment questionnaire submissions. In addition, females had a more meaningful variation in evaluations and self-assessments (males were more likely to self-present as low-risk). Since we lacked sufficient and balanced samples for males and non-binary/transgender users, we focused our analysis on females only.

The positive class of interest in our study was suicidal behavior, which was the minority group. Its scarcity in the data led to an imbalanced dataset. As such, stratification was performed when splitting into train and test groups. Each participant had multiple visits (labels), all assigned to either the train or the test group, to avoid data leakage. Due to the small sample size, avoiding any bias between the train/test groups was essential. Therefore, we aligned the distribution of several parameters, such as age and compliance rate, between the two groups.

Only participants who completed the entire course of the study and follow-up visits were included in the test group. Ten of the 43 female participants were removed for failing to meet the inclusion criteria. The remaining 33 females were included in the sample – 23 for training and 10 for testing. As a result, the training dataset contained 174 labels, and the test set had 71 labels.

3. STB prediction from self-reports

We decided to predict visit labels based on questionnaires from the preceding week. Visit dates not preceded by a questionnaire in the previous week were removed from the analysis. Since patients were allowed to complete the questionnaire more than once on the same day, we had specific dates containing several self-reports in which patients' responses were not identical. As E. M. Kleiman et al. (2017) showed, suicidal ideation varies within a day. In such cases, we adopted the following mechanism for obtaining one

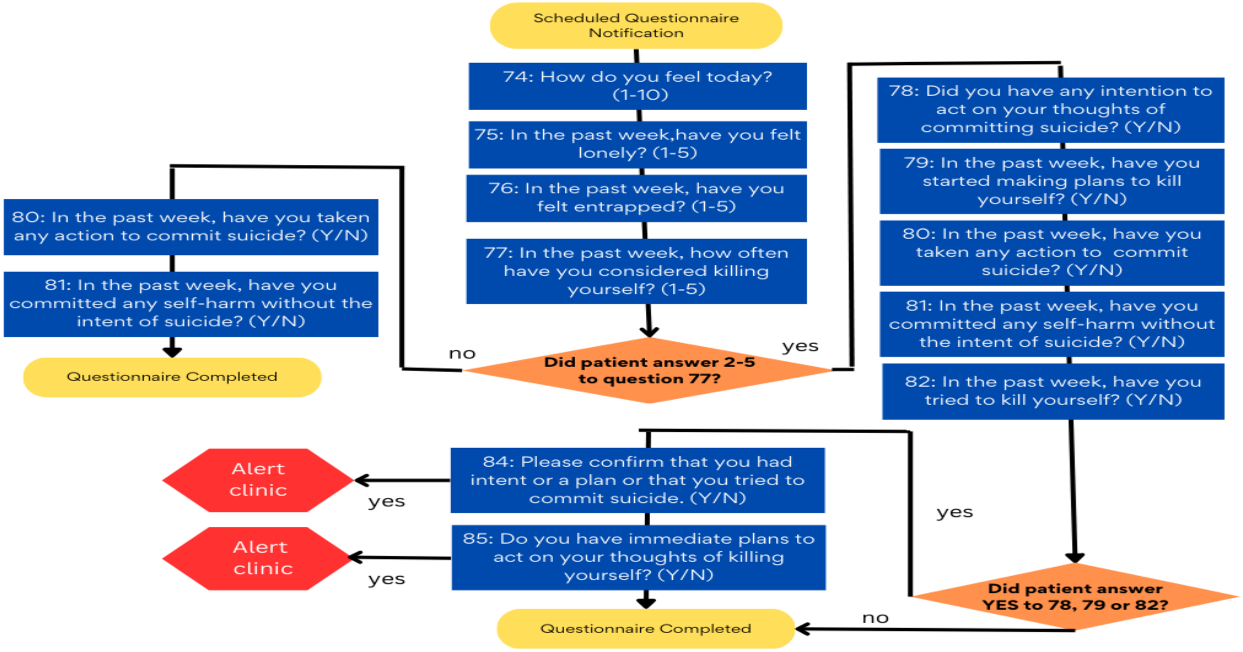
Scheduled Questionnaire Notification
74: How do you feel today? (1-10)
75: In the past week,have you felt lonely? (1-5)
76: In the past week, have you felt entrapped? (1-5)
77: In the past week, how often have you considered killing yourself? (1-5)
Did patient answer 2-5 to question 77?
no
yes
80: In the past week, have you taken any action to commit suicide? (Y/N)
81: In the past week, have you committed any self-harm without the intent of suicide? (Y/N)
Questionnaire Completed
78: Did you have any intention to act on your thoughts of committing suicide? (Y/N)
79: In the past week, have you started making plans to kill yourself? (Y/N)
80: In the past week, have you taken any action to commit suicide? (Y/N)
81: In the past week, have you committed any self-harm without the intent of suicide? (Y/N)
82: In the past week, have you tried to kill yourself? (Y/N)
Did patient answer YES to 78, 79 or 82?
yes
no
84: Please confirm that you had intent or a plan or that you tried to commit suicide. (Y/N)
85: Do you have immediate plans to act on your thoughts of killing yourself? (Y/N)
yes
yes
Alert clinic
Alert clinic
Questionnaire Completed

Figure 1. Questionnaire questions and flow.

questionnaire per date: first, we removed duplicate questionnaires. Then, we kept the questionnaires with the highest compliance, i.e., where patients responded to the highest number of questions. Finally, given two questionnaires with the exact compliance rate and different answers, we chose the more severe one to obtain the maximal severity of STB during that week. Eventually, we predicted each visit label based on one questionnaire – the one closest to its visit date. The intersection between visit dates and questionnaire adherence resulted in a dataset of 134 labeled samples.

The digital questionnaires were intended to capture active suicidal thoughts. To obtain clinically significant active suicidal ideation, we used an STB threshold of 1 to divide our data into low-risk (STB<=1) and high-risk (STB>1). This chosen threshold distinguished between passive, death-related thoughts and active suicidal thoughts (Nilsson et al., 2013). As a result, our training set contained 88 (65.7%) low-risk cases and 46 (34.3%) high-risk cases, and our test set contained 31 (62%) low-risk cases and 19 (38%) high-risk cases.

3.1. Classification features

Classification features were constructed from patients' responses, patients' engagement in the questionnaire throughout the experiment, and the inherent flow of the questionnaire. First, a classification feature was built for each question, with values corresponding to patients' responses. Null values were replaced by -1 to indicate invalid answers. Eventually, we decided to forfeit questions 74 (which differed in meaning and scaling from other questions) and the control questions 84 and 85 (which showed low feature importance due to correlation with other features).

Patient engagement was described by three features: a daily count of questionnaires capturing a patient's tendency to fill more than one questionnaire on a specific date; an aggregated user counter indicating how many times the patient responded to questionnaires throughout the entire experiment; and a feature recording the difference in days between consecutive questionnaires.

We used three binary features to capture the questionnaire sequence, each indicating if the patient continued to the next step: whether a high-risk response was recorded in question 77, whether at least one question out of 78, 79, and 82 triggered the control questions, and whether one of the control questions (84, 85) was positive, showing extra caution is in order. In addition, we consolidated the sequence into one categorical feature with three risk levels: low (if no risk was found according to question 77), medium (if a threat was found according to question 77 but not according to questions 78, 79, or 82, and high (if a risk was found according to questions 77 and 78, 79, 82).

3.2. Algorithms and model evaluation

We trained classification models using three algorithms: random forest, AdaBoost, and logistic regression. We performed hyperparameter optimization using a grid search over several parameters for each one. All models were trained and evaluated using Python's scikit-learn (sklearn) library.

To compare model performance, we implemented a leave-one-patient-out cross-validation (CV) method. During each run, a different patient was extracted from the training set and used as a testing set for model validation. Eventually, we obtained an aggregated classification matrix for the entire training set and computed standard classification metrics: accuracy, precision, recall, and f1-score. The recall score was fundamental in our setting since we aimed to identify high-risk patients. High recall scores meant seldom missing patients that should be further monitored.

The best model to perform on our training set was also applied to the test set for evaluation. In addition to calculating the standard classification metrics, we performed two statistical tests to confirm the significance of the classification. Fisher's exact and Barnard's tests were applied to the testing set confusion matrix to assess the plausibility of obtaining the resulting matrix by chance. These two tests have been used for the same purpose in previous studies with small datasets like ours (Vidyasagar, 2017; Diedrich and Niggemann, 2022).

4. STB prediction from cellular data

We aimed to predict clinical assessments from measurements passively collected by the iFeel app from patients' cell phones. These included app usage (app type and duration) and cellphone events (e.g., the screen on/off) and were collected continuously during the clinical study. We term this goal "Passive STB Prediction" to emphasize that the patient is not required to take any action (other than installing the iFeel app), nor is any intervention made. Thus, the iFeel app serves as a passive monitoring device for patients' behavior and routines. This setting poses a few inherent challenges:

- Data's nature – many different apps and events can be monitored. However, at each point, only a handful are active. Thus, the data is highly sparse. Also, we did not collect any specific app data or content to preserve privacy. Only indications for app usage were used.

- Dataset size – data was collected continuously, 24/7, for 6 months, but for a relatively small number of subjects. Thus, we had a reasonable amount of data per patient (intra-patient examples), but a relatively small amount of cross-patient data (inter-patient sequences). Also, there was a fair amount of missing data, due to connectivity issues, version updates, etc.

- Labels – the labels were based on in-clinic clinical assessments according to the C-SSRS index that ranges from 0 to 9. On average, approximately 5 clinical assessments per patient were conducted during the clinical study. The duration between each patient's visits to the clinic varied. The relatively small number of labels and their heterogeneous distribution led us to define a binary classification task with a clinically significant threshold: STB$>$3. It reflects active suicidal ideation with intent to act or even actions. This setting resulted in an unbalanced classification problem of approximately 88% low-risk patients vs. 12% patients at high risk.

- Data and label alignment – syncing between labels derived from the (higher level) clinical assessments and the (low level) cellphone features is also challenging: clinical assessment reflects patient condition "around the time of the visit." In contrast, the high-frequency cell phone data demonstrates the patient's instantaneous activities and state. Duplicating the same label over multiple adjacent feature vectors or labeling just one instance and utilizing semi-supervised methods is fundamentally wrong since the clinical assessment and the collected cellphone measurements act at different time scales and reflect different levels of information.

We addressed these challenges via a carefully designed preprocessing step, followed by a novel two-step classification method to maximize the utilization of labeled and unlabeled data. At the preprocessing step, we first mapped the various apps into predefined app types. Then, we aggregated the data on an hourly basis and constructed feature vectors where each feature recorded hourly usage (in seconds) of a specific app type. Applying these actions reduced the feature space dimension and sparsity.

To sync between the labels and the feature vectors, we grouped the feature vectors from the preceding week for each clinical visit (label). Thus, a labeled example was a set of 181 feature vectors with a single STB score, i.e., the set's label. It should be noted that for all patients, most weeks had no label since the patient did not visit the clinic. Only about 5 labeled sets could be extracted for each patient. Namely, we ended with many unlabeled feature vectors and a few labeled sets. Furthermore, the events that we tried to predict (active suicidal thoughts and attempts) were highly uncommon.

To meet these characteristics, we designed a dedicated two-step method for detecting abnormal behavior, as described in detail in subsections 4.1 and 4.2.

4.1. First step – clustering and representation

The purpose of this stage was to utilize the immense amounts of unlabeled data the cellphones provided. To this end, we clustered the (unlabeled) feature vectors using the k-means clustering algorithm. The outcome was a set of clusters, each reflecting an hourly behavioral pattern common in this patient population.

We mapped each feature vector to the closest cluster, as measured by Mahalanobis distance (De Maesschalck et al., 2000). The distances from each cluster were binned into nine quantiles, and each vector ascribed to the cluster belonged to one of them. Thus, each vector was essentially mapped to a specific distance quantile in its closest cluster. Unique vectors (representing uncommon behavior) were mapped to higher quantiles.

Then, we represented a set of feature vectors by a vector of counts by recording the number of vectors belonging to each bin (quantile) in each cluster. As a result, a labeled dense vector represented a sparse labeled set of feature vectors. It should be noted that even high-risk patients usually exhibit normal behavior patterns. However, we expected they would display a growing amount of abnormal behavior patterns, reflected by higher counts at the high quantiles entries.

4.2. Second step – classification

Common supervised learning methods and best practices were applied at this step. Those included further preprocessing steps (such as feature construction and selection) and classification methods. The outcome was a model designed to predict a patient's clinical state given behavioral patterns from the preceding week, as captured by their cellphone usage.

For model evaluation, we repeated the process described in subsection 3.2: leave-one-patient-out CV, classification metrics calculation, and statistical validation using Fisher and Barnard's test.

5. Results

This section presents results from the two parts of our research. Subsection 5.1 describes the results for predicting STB from the weekly self-reported questionnaires, and subsection 5.2 presents the results for predicting STB from passively monitored cellular usage.

Table 1. Test set classification matrix for predicting STB from self-reports

	Predicted positive	Predicted negative
True positive	15	4
True negative	7	24

Table 2. Classification metrics for predicting STB from self-reports

	Precision	Recall	F1
0	0.86	0.77	0.81
1	0.68	0.79	0.73
Accuracy			0.78
Macro avg	0.77	0.78	0.77
Weighted avg	0.79	0.78	0.78

5.1. STB prediction from self-reports

The random forest algorithm with the following hyperparameter setting: *class_weight='balanced', max_features=None, max_depth=3, n_estimators=10, criterion='gini', min_samples_leaf=3, ccp_alpha=0, bootstrap=None* achieved the best CV results. Applying the same model to the test set resulted in the classification matrix presented in Table 1. Table 2 summarizes the classification metrics and their values showing the high recall score.

Fisher and Barnard's tests were used to determine the likelihood of receiving the confusion matrix in Table 1 by chance. The P-value for the Fisher test was 0.000133, and the P-value for Barnard's test was 0.000097. Both tests showed P-values far less than 0.05, indicating highly statistically significant results. Therefore, it is unlikely to achieve this confusion matrix by randomly assigning positive and negative labels to samples.

Using the feature importance attribute of the random forest classifier, we explored which features contributed most to the classification. Our best classification feature was the aggregated user count of questionnaires, capturing patient participation and engagement with the study. The second best classification feature was question 77, influencing the questionnaire sequence.

5.2. STB prediction from cellular data

The raw cellphone data was mapped to 26-dimension feature vectors representing hourly app type usage (in seconds). After cleansing, we had an unlabeled dataset of 48981×26 vectors. We then applied unsupervised feature selection and reduced dimensionality to 16. The top five selected features were: *'VideoPlayers&Editors'*, *'Tools'*, *'Social'*, *'Communication'*, and *'Personalization'*.

For clustering (step 1), we used the k-means algorithm with 50 initial starts. The number of clusters was set to 24 using held-out data. We transformed the labeled example sets into 216-dimension feature vectors in the representation change phase. Using the F-test, which measures linear dependency, we applied supervised univariate feature selection to retrieve the best 16 features. We ended up with 125×16 and 51×16 labeled vectors in the training and test sets, respectively.

Due to the small training set, we only used traditional classification models (rather than deep learning) while carefully controlling model selection to avoid overfitting. We tested random forest, logistic regression, linear and nonlinear support vector machine (SVM), and a shallow neural network. We selected random forest as the classification method of choice.

The random forest algorithm with the following hyperparameter setting: *n_estimators=20, criterion='gini', max_depth=3, max_features='log_2', class_weight='balanced', bootstrap=True, min_weight_fraction_leaf=0.3, ccp_alpha=0* achieved the best CV results. Applying the same model to the test set resulted in the classification matrix presented in Table 3. Table 4 summarizes the classification metrics and their values showing the high recall score.

As described in subsection 5.1, Fisher and Barnard's tests were used to determine classification significance. Since both tests showed highly statistically significant results (P-value for Fisher test: 0.000447, P-value for Barnard's test: 0.000144), it is unlikely to receive the confusion matrix in Table 3 purely by chance.

6. Discussion

This study uses subjective self-reported weekly questionnaires and passively collected cellphone usage data to predict periodic clinical assessments of STB. It suggests a novel approach for utilizing many unlabeled cellular data in a limited-data classification problem via clustering. The high prediction levels provide additional validation for incorporating digital monitoring in the clinical field. It may equip clinicians with tools to assess patient health status, identify risks, and intervene earlier.

Nonetheless, translating predictions into prevention is challenging. Identifying a suicidal patient does not seamlessly translate into diminishing the intensity of suicidal thoughts or impeding their actualization. An evidence-based suicide prevention approach, capable of immediate and effective implementation, must be integrated to complement these alerts. Moreover, the management of false alarms at both individual and systemic levels necessitates consideration.

In this study, high suicide risk alerts were addressed, adhering to the clinical practice guidelines of the medical center. A licensed clinician promptly contacted the parents and conducted a thorough assessment of the participant on the same business day. The specific intervention plan was tailored to each participant's treatment status, encompassing communication with treating clinicians, reinforcement of the safety plan, and provision of subsequent referrals. The impact of these interventions, within this study and other cellphone monitoring studies of STB, remains a subject for exploration. Subsequent research may shed light on the effectiveness of mitigating suicide-related outcomes through cellphone-based risk detection.

Table 3. Test set confusion matrix for predicting STB from cellphone data

	Predicted positive	Predicted negative
True positive	5	2
True negative	3	41

Table 4. Classification metrics for predicting STB from cellphone data

	Precision	Recall	F1
0	0.95	0.93	0.94
1	0.62	0.71	0.67
Accuracy			0.90
Macro avg	0.79	0.82	0.80
Weighted avg	0.91	0.90	0.90

6.1. Limitations and future work

Due to insufficient samples for high-risk males and non-binary/transgender users in general, the study was based on a relatively small dataset of female participants and labeled data from clinical assessments. It posed inherent limitations in the study design while emphasizing the need to develop algorithms and models dedicated to a small data regime.

Labels of the same user were considered independent, even though assessments from consecutive visits may correlate. Feature vectors were also dependent due to the inherent sequential nature of the data. Both dependencies suggest room for stochastic personalized models that learn from patients' clinical assessments over time.

The iFeel data collection app used in this study is only supported by Android mobile phones. Hence, This study focused on Android mobile users, which limits the representativeness and generalizability of the findings. Investigating other data collection apps to include users of different platforms can provide more inclusive results.

The small amount of labeled data limited the ability to use advanced deep-learning methods. An encouraging outcome of this study is the correlation between the self-assessment and the clinical assessment. It opens the door for incorporating the weekly self-reported questionnaires as ground truth labels. Such a synergetic setting of active (self-reporting) and passive (cellphone) data collection may be further extended to probe the patient with pop-up questions based on the cellphone data analysis. The extended label set and the probing will allow the use of advanced deep architectures and active learning methods. From a scientific data standpoint, synergizing labels from different sources requires transformation to a unified scale while possibly adding a confidence indication. These directions are left for future study.

6.2. Conclusion

Cellphone monitoring holds the potential to enhance the clinical evaluation of imminent suicide risk substantially. Through this approach, the identification of the risk for suicidal behavior can achieve greater scalability and contribute to alleviating the present shortage of manpower resources within the realm of mental health. Data collection through passive means reduces the assessment burden (both for the patient and mental health services) and dependence on patients revealing their suicidal intentions. Integrating various self-reported and passive data sources into a validated predictive model can markedly enhance the identification of adolescents facing imminent risk of suicidal behavior.

Unlike self/clinician reports, which are subjective measures to detect suicide risk, cellphone monitoring provides objective indicators crucial for more accurate predictions. Objective measures will help the mental health assessments be more similar to the practices in medicine. Moreover, prediction by *passive* cellphone usage data, as used in this study, is less burdensome than self-report, which relies on compliance over time.

References

Al-Dajani, N., & Czyz, E. K. (2022). Suicidal desire in adolescents: An examination of the interpersonal psychological theory using daily diaries. *Journal of Clinical Child & Adolescent Psychology*, 1–15.

Barzilay, S., & Apter, A. (2022). Recent research advances in identification and prevention of youth suicide risk. *Current opinion in psychiatry*, *35*(6), 395–400.

Barzilay, S., Fine, S., Akhavan, S., Haruvi-Catalan, L., et al. (in press). Real-time real-world digital monitoring of adolescent suicide risk during six-month following emergency department discharge: Protocol for an intensive longitudinal study. *JMIR Res Protocol*.

Cheek, S. M., Reiter-Lavery, T., & Goldston, D. B. (2020). Social rejection, popularity, peer victimization, and self-injurious thoughts and behaviors among adolescents: A systematic review and meta-analysis. *Clinical psychology review*, *82*, 101936.

Czyz, E., Koo, H., Al-Dajani, N., King, C., & Nahum-Shani, I. (2021). Predicting short-term suicidal thoughts in adolescents using machine learning: Developing decision tools to identify daily level risk after hospitalization. *Psychological medicine*, 1–10.

Czyz, E. K., Koo, H. J., Al-Dajani, N., Kentopp, S. D., Jiang, A., & King, C. A. (2022). Temporal profiles of suicidal thoughts in daily life: Results from two mobile-based monitoring studies with high-risk adolescents. *Journal of psychiatric research*, *153*, 56–63.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1–18.

Diedrich, A., & Niggemann, O. (2022). On residual-based diagnosis of physical systems. *Engineering Applications of Artificial Intelligence*, *109*, 104636.

Galynker, I. (2023). *The suicidal crisis: Clinical guide to the assessment of imminent suicide risk*. Oxford University Press.

Ghandeharioun, A., Fedor, S., Sangermano, L., Ionescu, D., Alpert, J., Dale, C., Sontag, D., & Picard, R. (2017). Objective assessment of depressive symptoms with machine learning and wearable sensors data. *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, 325–332.

Glenn, C. R., Kleiman, E. M., Kandlur, R., Esposito, E. C., & Liu, R. T. (2022). Thwarted belongingness mediates interpersonal stress and suicidal thoughts: An intensive longitudinal study with high-risk adolescents. *Journal of Clinical Child & Adolescent Psychology*, *51*(3), 295–311.

Glenn, C. R., & Nock, M. K. (2014). Improving the short-term prediction of suicidal behavior.

American journal of preventive medicine, *47*(3), S176–S180.

Horwitz, A., Czyz, E., Al-Dajani, N., Dempsey, W., Zhao, Z., Nahum-Shani, I., & Sen, S. (2022). Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns. *Journal of Affective Disorders*, *313*, 1–7.

Joiner, T. E., Simpson, S., Rogers, M. L., Stanley, I. H., & Galynker, I. I. (2018). Whether called acute suicidal affective disturbance or suicide crisis syndrome, a suicide-specific diagnosis would enhance clinical care, increase patient safety, and mitigate clinician liability. *Journal of Psychiatric Practice®*, *24*(4), 274–278.

Kleiman, E., Millner, A. J., Joyce, V. W., Nash, C. C., Buonopane, R. J., Nock, M. K., et al. (2019). Using wearable physiological monitors with suicidal adolescent inpatients: Feasibility and acceptability study. *JMIR mHealth and uHealth*, *7*(9), e13725.

Kleiman, E. M., Bentley, K. H., Maimone, J. S., Lee, H.-I. S., Kilbury, E. N., Fortgang, R. G., Zuromski, K. L., Huffman, J. C., & Nock, M. K. (2021). Can passive measurement of physiological distress help better predict suicidal thinking? *Translational psychiatry*, *11*(1), 611.

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of abnormal psychology*, *126*(6), 726.

Lejeune, A., Le Glaz, A., Perron, P.-A., Sebti, J., Baca-Garcia, E., Walter, M., Lemey, C., & Berrouiguet, S. (2022). Artificial intelligence and suicide prevention: A systematic review. *European psychiatry*, *65*(1), e19.

Liu, T., Liang, P. P., Muszynski, M., Ishii, R., Brent, D., Auerbach, R., Allen, N., & Morency, L.-P. (2020). Multimodal privacy-preserving mood prediction from mobile data: A preliminary study. *arXiv preprint arXiv:2012.02359*.

Miller, A. B., & Prinstein, M. J. (2019). Adolescent suicide as a failure of acute stress-response systems. *Annual review of clinical psychology*, *15*, 425–450.

Morshed, M. B., Saha, K., Li, R., D'Mello, S. K., De Choudhury, M., Abowd, G. D., & Plötz, T. (2019). Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), 1–21.

Mullick, T., Radovic, A., Shaaban, S., Doryab, A., et al. (2022). Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning–based exploratory study. *JMIR Formative Research*, *6*(6), e35807.

Nilsson, M. E., Suryawanshi, S., Gassmann-Mayer, C., Dubrava, S., McSorley, P., & Jiang, K. (2013). Columbia–suicide severity rating scale scoring and data analysis guide. *CSSRS Scoring Version*, *2*, 1–13.

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., Currier, G. W., Melvin, G. A., Greenhill, L., Shen, S., et al. (2011). The columbia–suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, *168*(12), 1266–1277.

Rabasco, A., & Sheehan, K. (2022). The use of intensive longitudinal methods in research on suicidal thoughts and behaviors: A systematic review. *Archives of suicide research*, *26*(3), 1007–1021.

Schafer, K. M., Kennedy, G., Gallyer, A., & Resnik, P. (2021). A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis. *PloS one*, *16*(4), e0249833.

Sedano-Capdevila, A., Porras-Segovia, A., Bello, H. J., Baca-Garcia, E., & Barrigon, M. L. (2021). Use of ecological momentary assessment to study suicidal thoughts and behavior: A systematic review. *Current psychiatry reports*, *23*(7), 41.

Vidyasagar, M. (2017). Machine learning methods in computational cancer biology. *Annual Reviews in Control*, *43*, 107–127.

Voros, V., Tenyi, T., Nagy, A., Fekete, S., & Osvath, P. (2021). Crisis concept re-loaded?—the recently described suicide-specific syndromes may help to better understand suicidal behavior and assess imminent suicide risk more effectively. *Frontiers in psychiatry*, *12*, 598923.

Wang, S. B., Dempsey, W., & Nock, M. K. (2022). Machine learning for suicide prediction and prevention: Advances, challenges, and future directions. In *Youth suicide prevention and intervention: Best practices and policy implications* (pp. 21–28). Springer International Publishing Cham.