Journal of Affective Disorders 313 (2022) 1–7

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns

Adam Horwitz [a,*], Ewa Czyz [a], Nadia Al-Dajani [a], Walter Dempsey [b], Zhuo Zhao [c], Inbal Nahum-Shani [b], Srijan Sen [a,c]

[a] Department of Psychiatry, University of Michigan, USA
[b] Institute for Social Research, University of Michigan, USA
[c] Molecular and Behavioral Neuroscience Institute, University of Michigan, USA

ARTICLE INFO

Keywords:
Depression
Suicidal ideation
Daily diary
Wearable sensors
Mobile health

ABSTRACT

Background: Intensive longitudinal methods (ILMs) for collecting self-report (e.g., daily diaries, ecological momentary assessment) and passive data from smartphones and wearable sensors provide promising avenues for improved prediction of depression and suicidal ideation (SI). However, few studies have utilized ILMs to predict outcomes for at-risk, non-clinical populations in real-world settings.

Methods: Medical interns (N = 2881; 57 % female; 58 % White) were recruited from over 300 US residency programs. Interns completed a pre-internship assessment of depression, were given Fitbit wearable devices, and provided daily mood ratings (scale: 1–10) via mobile application during the study period. Three-step hierarchical logistic regressions were used to predict depression and SI at the end of the first quarter utilizing pre-internship predictors in step 1, Fitbit sleep/step features in step 2, and daily diary mood features in step 3.

Results: Passively collected Fitbit features related to sleep and steps had negligible predictive validity for depression, and no incremental predictive validity for SI. However, mean-level and variability in mood scores derived from daily diaries were significant independent predictors of depression and SI, and significantly improved model accuracy.

Limitations: Work schedules for interns may result in sleep and activity patterns that differ from typical associations with depression or SI. The SI measure did not capture intent or severity.

Conclusions: Mobile self-reporting of daily mood improved the prediction of depression and SI during a meaningful at-risk period under naturalistic conditions. Additional research is needed to guide the development of adaptive interventions among vulnerable populations.

1. Introduction

Depression is the leading cause of disease-associated disability worldwide (World Health Organization, 2017) and, in the United States (US), has an annual economic burden of over 200 billion dollars (Greenberg et al., 2015). The prevalence rate for Major Depressive Disorder (MDD) in the US has increased over the past 25 years, and is highest among young adults (Hasin et al., 2018). Depression is frequently present among individuals who die by suicide (Hawton et al., 2013). In parallel, trends for suicide in the US have closely mirrored those of depression, with suicide rates increasing by 35 % over the past 20 years (Hedegaard et al., 2020). Methodological limitations have hindered our ability to predict suicide-related outcomes in a way that meaningfully informs prevention efforts. For example, a meta-analysis of studies spanning over 50 years found that methods used for predicting suicidal thoughts and behaviors often rely on single measures and predict outcomes over long and unspecific periods of time (Franklin et al., 2017), versus shorter-term periods necessary to inform timely interventions.

Intensive longitudinal methods (ILMs) for collecting self-report [e.g., ecological momentary assessments (EMAs), daily diaries] and passive data from smartphones and wearable devices (e.g., smartwatches) provide significant advantages and opportunities to improve the prediction of mental health outcomes. Responding to prompts on a mobile phone daily or multiple times per day significantly reduces recall bias compared to assessments recounting the previous week (e.g., Gratch

* Corresponding author at: 4250 Plymouth Rd., Ann Arbor, MI, 48105, USA.
E-mail address: ahor@umich.edu (A. Horwitz).

https://doi.org/10.1016/j.jad.2022.06.064
Received 12 November 2021; Received in revised form 9 May 2022; Accepted 22 June 2022
Available online 25 June 2022

et al., 2021) and is particularly important in the context of assessing mood and suicidal thoughts, as these often fluctuate over time (e.g., Czyz et al., 2019; Kleiman et al., 2017; Targum et al., 2021). Another advantage of such repeated assessments is the ability to develop dynamic features, such as variability, that improve prediction of depression and suicidal ideation (e.g., Czyz et al., 2020; Peters et al., 2020; Schoevers et al., 2020). While studies utilizing intensive longitudinal assessments to predict suicidal ideation are promising, there are limitations regarding generalizability and implementation outside of research settings. Most studies to date have used small samples ($n < 100$) of clinical or inpatient populations, required multiple EMA responses per day (i.e., high participant burden), and provided monetary incentives for each completed survey (e.g., Rabasco and Sheehan, 2021; Sedano-Capdevila et al., 2021). Few studies have examined the potential for intensive longitudinal assessments to make predictions regarding risk for negative mental health outcomes during sensitive periods for less severe populations, under naturalistic conditions.

Prospective studies have established that too much, or too little, sleep can predispose, precipitate, and perpetuate depression (e.g., Zhai et al., 2015). Likewise, a lack of physical activity increases the risk of depression (e.g., Schuch et al., 2018). Passive sensing from smartphones and wearable devices, collected over weeks to months, have demonstrated an ability to predict mood and depressive symptoms from these (i.e., sleep, physical activity) and other (e.g., heart rate variability, speech patterns) features (e.g., Ben-Zeev et al., 2015; Or et al., 2017; Pedrelli et al., 2020; Tazawa et al., 2020; Yim et al., 2020), and provides the distinct advantage of low participant burden. While the evidence is strong for these features to aid in the detection and prediction of depression, empirical support for utilizing passive mobile or sensor data to predict suicidal ideation has been limited. A study by Haines-Delmont et al. (2020) used machine learning methods to analyze features from sensor and mobile data to predict suicidal ideation in the week following discharge from psychiatric hospitalization, but overall predictive accuracy was fairly weak (Area Under the Curve = 0.65). Further, among adults reporting past-month suicidal ideation and sleeping difficulties, sensor-based estimates of total sleep time predicted next-day suicidal ideation, but the effect was similar to subjective reports of sleep via EMA, with subjective sleep quality having the strongest predictive effect for suicidal ideation (Littlewood et al., 2018). Additional research is needed to clarify whether specific features from mobile or wearable data can be used to meaningfully aid in the prediction of suicidal ideation.

Medical internship (i.e., first year of residency physician training) is associated with elevated rates of depression (Mata et al., 2015), with the prevalence of depression increasing from 3.9 % immediately before internship to 27.1 % by the end of the first quarter of the internship year (Sen et al., 2010). Suicidal ideation also increases during internship and physicians die by suicide at rates greater than the general population (Goldman et al., 2015). Taken together, the first quarter of internship presents a unique situation whereby the development of depression and suicidal ideation can be expected among a relatively healthy, non-clinical sample of mostly young adults. The present study seeks to build upon the existing literature of ILMs predicting mental health outcomes by using a large, naturalistic sample of medical interns to examine aggregated Fitbit and daily diary mood data, over a clinically meaningful time frame (i.e., the first quarter of internship), as prospective predictors of depression and suicidal ideation at the end of the quarter. To gain a better understanding of the predictive capacities for different data sources, we will examine traditional survey measures, passive data, and daily mood data in incremental steps. Results from this study may have implications for understanding the added value of ILMs to traditional assessments and offer insights into the development of interventions that can adaptively respond to survey assessments or passively collected data detecting risk for adverse outcomes during a period of significant vulnerability.

2. Methods

2.1. Participants

The Intern Health Study is a prospective cohort study assessing stress and depression during the first year of residency training in the United States (Sen et al., 2010). Participants included in this analysis were 2881 medical doctors from over 300 residency institutions across the Uniterrd States who enrolled in the study during the 2018–2019 and 2019–2020 academic years. The sample was 56.8 % female, and the mean age was 27.6 years ($SD = 2.6$). Racial/ethnic distribution was as follows: 58.0 % White, 22.7 % Asian, 4.7 % Black, 3.8 % Hispanic/Latinx, 8.4 % Multiracial, and 2.4 % Other race.

2.2. Measures

2.2.1. Baseline and quarterly survey measures

The baseline survey was administered via mobile application prior to the start of internship and assessed basic demographic information (e.g., age, race) and clinical symptoms. The follow-up survey was administered via mobile application during the third month of internship.

2.2.1.1. Depression. The Patient Health Questionnare-9 (PHQ-9; Kroenke et al., 2001) assesses the nine DSM-5 depressive symptoms. Items are rated on a 4-point Likert scale (0–27 full scale range) for frequency of being bothered by a particular symptom in the past two weeks, ranging from "not at all" to "nearly every day." The PHQ-9 was administered at baseline and at the follow-up survey, and had an internal consistency of $\alpha = 0.78$ in this sample.

2.2.1.2. Suicidal ideation. The final item from the PHQ-9 was used to assess frequency of suicidal ideation, "thoughts that you would be better off dead or hurting yourself in some way." This item was dichotomized based on presence or absence of any suicidal ideation in the past two weeks for the baseline and follow-up assessments.

2.2.2. Daily measures

2.2.2.1. Sleep and steps. Sleep and step data were collected passively via Fitbit. Averages for daily sleep minutes and daily step counts were aggregated from responses during the study period. Features were also generated from variability in these daily scores, changes in daily scores over time, and adherence to wearing their Fitbit devices. Previous studies have demonstrated the reliability and validity of using data collected by Fitbits to measure these constructs (for a review, see Evenson et al., 2015).

2.2.2.2. Mood. Mood valence was assessed with a single item using a mobile application, "On a scale of 1 (lowest) to 10 (highest), how was your mood today?" Single-item mood measures are clinically useful proxies for larger scales, such as the PHQ-9 (Aguilera et al., 2015). Participants were prompted through the mobile application notification to complete this measure daily at a user-specified time between 5 pm and 10 pm. Features derived from aggregated mood scores included daily average mood, variability in mood, changes in mood over time, and adherence daily mood prompts.

2.3. Procedures

The study was approved by the University of Michigan Institutional Review Board, and we obtained informed consent from all study participants. Two-to-three months prior to the start of internship, medical doctors were invited to participate in the study. Those consenting to participate received instructions for downloading the study mobile application, completed the baseline assessment, and provided a mailing

address for the Fitbit Charge 2 device, which was sent shortly thereafter. Participants were instructed to sync Fitbits to their phone and wear them regularly throughout the intern year (including when they sleep) to collect objective data on their daily activity such as steps, sleep, and heartrate. During internship, the mobile application was utilized to conduct assessments of daily mood and the follow-up survey, aggregate and visualize data, and deliver push notifications (for additional details, see NeCamp et al., 2020). For the purposes of this study, variables using aggregated daily mood, sleep, and step data only included data from the first day of internship until the follow-up survey at the end of Quarter 1. Participants were compensated with $25 at the baseline and the follow-up assessment, but importantly, they were not compensated for the completion of daily mood surveys or adherence to wearing their Fitbit. Out of the 3814 interns participating in the study, this study's analytic sample was restricted to 2881 interns (75.6 %) who had completed the baseline survey and the 1st quarter follow-up survey. Compared to interns in the analytic sample, interns who did not complete the 1st quarter follow-up survey had significantly higher mean baseline scores of depression (3.22 vs. 2.66), were less likely to be female (52.4 % vs. 56.8 %), and were more likely to report suicidal ideation at baseline (5.6 % vs. 3.0 %).

2.4. Data analytic plan

Data were analyzed with R version 4.02. Features were derived from daily diary and Fitbit data aggregated during the first two-to-three months of internship (up until an individual's completion of the quarterly survey). This interval was chosen to capture the period of internship in which depressive symptoms have the greatest increase (Sen et al., 2010) and allow for an examination of changes over time. We developed four features for each of the mood, sleep, and step domains: mean scores, standard deviations (variability), completion percentage (adherence), and slope (change in mean score during this period). Our outcome variables of interest, assessed at the quarter #1 survey, were depression (scores of 10 or higher on the PHQ-9, indicating at least moderate depression) and suicidal ideation (any non-zero score for PHQ-9 item #9). *t*-tests and chi-squares examined between-groups differences in demographics and clinical characteristics for those who provided daily diary and Fitbit data and those who did not provide this data during the study period. At least two entries were required for each of the three domains (i.e., sleep, steps, mood) to have been considered adherent. Likewise, *t*-tests and chi-squares examined between-group differences in demographic, clinical, and daily-level data based on whether moderate depressive symptoms and suicidal ideation were reported at follow-up.

To assess the incremental predictive validity of Fitbit and daily diary features beyond the base model incorporating baseline depression and suicidal ideation, we conducted three-step hierarchical logistic regressions that included baseline and demographic variables in Step 1, passively collected Fitbit features in Step 2, and daily diary mood assessment features in Step 3. Individuals who did not provide daily diary or Fitbit data (n = 643) were not included in the regression analyses. We included multiple metrics to demonstrate model fit, and changes in steps, including Bayesian Information Criterion (BIC), Area Under the Curve (AUC), and Nagelkerke pseudo-R^2. Age, completion rate, and slope features were not significant in any models and were removed for parsimony. Tests of multicollinearity indicated that collinearity assumptions were met (VIF < 2.5) for all variables in the final regression models. Some of the interns in the study received push notifications containing tips and personalized feedback related to mood, sleep, and steps during the study period (NeCamp et al., 2020). This intervention variable was initially included as a covariate, and then removed for non-significance.

3. Results

3.1. Adherence to daily diary and Fitbit

Of the 2881 interns in the study, 2238 (77.7 %) provided at least two entries for each predictor: daily mood scores, sleep, and step count. Race was the only baseline variable that differed significantly between interns who did and did not adhere to daily diaries and wearing Fitbits. Specifically, interns who identified as Asian were overrepresented in the group of interns not completing daily diaries or wearing Fitbits, and White interns were overrepresented in the group of interns providing data in these domains. Follow-up mean scores of depression were slightly higher among those who did not adhere to daily diaries and Fitbits (See Table 1).

3.2. Univariate associations with depression and suicidal ideation at follow-up

Depression, defined as a PHQ-9 score of 10 or higher at the follow-up assessment was present in 500 (17.6 %) interns, and suicidal ideation in the past two weeks was endorsed by 217 (7.7 %) interns. Univariate prospective associations of sociodemographic, clinical, Fitbit, and daily diary variables with depression and suicidal ideation are presented in Table 2. As may be expected, pre-internship depression and suicidal ideation had significant associations with depression and suicidal ideation at follow-up. Mean scores and variability in daily mood reports were also associated with depression and suicidal ideation at follow-up. Greater variability in sleep and lower completion rates of daily mood

Table 1
Sample demographics, characteristics and adherence for daily diary and Fitbit data.

Variable	Total sample n = 2881 n(%)	No DD or Fitbit n = 643 (22.3 %) %	DD and Fitbit n = 2238 (77.7 %) %	Test statistic *t*(df) or χ^2(df)
Sex (% Female)	1636 (56.8 %)	55.1 %	57.3 %	$\chi^2(1) = 1.01$
Race/Ethnicity				$\chi^2(5) = 41.49$***
White	1671 (58.0 %)	**48.7 %**	**60.7 %**	
Black	134 (4.7 %)	4.7 %	4.6 %	
Hispanic/ Latinx	110 (3.8 %)	4.5 %	3.6 %	
Asian	654 (22.7 %)	**30.1 %**	**20.4 %**	
Multi-racial	243 (8.4 %)	7.9 %	8.6 %	
Other	69 (2.4 %)	3.6 %	2.1 %	
Age (M(SD))	27.58 (2.6)	27.61 (2.6)	27.57 (2.5)	t(2865) = 0.33
Pre PHQ-9 (M (SD))	2.66 (2.9)	2.80 (3.1)	2.61 (2.8)	t(2879) = 1.44
Pre SI (% Yes)	88 (3.1 %)	3.0 %	3.1 %	$\chi^2(1) = 0.01$
Q1 PHQ-9 (M (SD))	5.67 (4.2)	6.08 (4.3)	5.56 (4.1)	t(2839) = 2.70**
Q1 Dep (% Yes)	500 (17.6 %)	18.7 %	17.3 %	$\chi^2(1) = 0.69$
Q1 SI (% Yes)	217 (7.7 %)	6.1 %	8.1 %	$\chi^2(1) = 2.72$

Note. ** $p < .01$ *** $p < .001$. M(SD) = Mean(standard deviation). DD = Daily Diary. Pre = Pre-internship. PHQ-9 = Patient Health Questionnaire-9 score. Dep = PHQ-9 depression score ≥10, indicating moderate-to-severe symptoms. Q1 = Quarter 1 follow-up assessment. SI = Suicidal Ideation.
The bolding here indicates the groups that were signiifcantly different within the racial/ethnic groups based on adherence to the daily diary and FItbit data. (Demonstrating the more specific effects of the reported $\chi^2(5) = 41.49$***' omnibus test).

Table 2
Univariate associations of demographic, clinical, Fitbit, and daily diary variables with depression and suicidal ideation at Q1 follow-up.

	Depression Q1			Suicidal Ideation Q1		
	No (82.4 %) Mean (SD)	Yes (17.6 %) Mean (SD)	Test statistic t(df) or χ^2(df)	No (92.3 %) Mean (SD)	Yes (7.7 %) Mean (SD)	Test statistic t(df) or χ^2(df)
Sex (% Female)	55.7 %	61.6 %	$\chi^2(1) = 5.75^*$	57.1 %	53.5 %	$\chi^2(1) = 1.06$
R/E (% White)	58.4 %	58.0 %	$\chi^2(5) = 9.24$	58.7 %	55.3 %	$\chi^2(5) = 5.38$
Pre SI (% Yes)	1.5 %	10.1 %	$\chi^2(1) = 100.40^{***}$	1.5 %	21.8 %	$\chi^2(1) = 278.00^{***}$
Age	27.48 (2.9)	27.64 (3.0)	$t(2832) = 1.10$	27.50 (2.9)	27.78 (2.7)	$t(2805) = 1.38$
Pre PHQ-9	2.16 (2.3)	4.96 (3.8)	$t(579.4) = 15.73^{***}$	2.45 (2.6)	5.00 (4.1)	$t(231.5) = 9.02^{***}$
Sleep features						
Mean	6.36 (1.1)	6.28 (1.3)	$t(526.9) = 1.18$	6.34 (1.1)	6.40 (1.0)	$t(2328) = 0.66$
Variability (SD)	1.55 (0.6)	1.65 (0.6)	$t(2295) = 3.10^{**}$	1.56 (0.6)	1.63 (0.6)	$t(2277) = 1.55$
Slope	−0.358 (17.1)	−0.354 (19.9)	$t(2295) = 0.01$	−0.269 (17.7)	−0.545 (12.5)	$t(2277) = 0.21$
Completion %	54.62 (38.3)	51.48 (38.6)	$t(2839) = 1.67$	53.94 (38.3)	57.45 (37.7)	$t(2812) = 1.30$
Step features						
Mean	8623 (2689)	8373 (2696)	$t(2461) = 1.75$	8585 (2670)	8616 (2963)	$t(2438) = 0.15$
Variability (SD)	3662 (1259)	3552 (1126)	$t(2444) = 1.66$	3642 (1228)	3641 (1304)	$t(2421) = 0.01$
Slope	34.72 (507.8)	35.36 (135.6)	$t(2444) = 0.02$	36.87 (552.8)	28.31 (135.6)	$t(2421) = 0.21$
Completion %	68.68 (37.5)	66.50 (38.4)	$t(2839) = 1.72$	68.93 (37.8)	72.71 (35.9)	$t(257.6) = 1.49$
Mood features						
Mean	7.42 (1.0)	6.44 (1.3)	$t(632.1) = 15.51^{***}$	7.33 (1.1)	6.27 (1.4)	$t(237.8) = 10.66^{***}$
Variability (SD)	1.00 (0.4)	1.27 (0.5)	$t(651.0) = 10.56^{***}$	1.03 (0.5)	1.28 (0.5)	$t(241.4) = 6.56^{***}$
Slope	0.0015 (0.03)	0.0029(0.04)	$t(623.6) = 0.77$	0.0017 (0.04)	0.0033(0.05)	$t(2752) = 0.61$
Completion %	51.08 (26.1)	47.46 (26.9)	$t(2839) = 2.80^{**}$	50.50 (26.2)	50.35 (27.0)	$t(2812) = 0.08$

Note. $^*p < .05$ $^{**}p < .01$ $^{***}p < .001$. Sleep variables were scaled to hours. Mood was scored from 1 to 10, with lower scores indicating worse mood. Depression as an outcome was indicated by a score of 10 or higher on the PHQ-9. R/E = Race/Ethnicity. Pre = Pre-internship. Q1 = Quarter 1 follow-up assessment. SI = Suicidal Ideation. PHQ-9 = Patient Health Questionnaire-9 score. SD = Standard deviation.

diaries had significant associations with depression only. None of the other features derived from Fitbit sensors or daily diary data were independently associated with depression or suicidal ideation at follow up.

3.3. Hierarchical logistic regressions

The three-step hierarchical logistic regression examining baseline demographic and clinical variables in step 1, Fitbit features in step 2, and daily diary mood features in step 3 as longitudinal predictors of depression is presented in Table 3. Receiver operating characteristic (ROC) curves for each step are included in Fig. 1. Intern sex, pre-internship depression, and pre-internship suicidal ideation were significant independent predictors of depression in Step 1. The inclusion of sleep and step variables from Fitbit features provided a small improvement in model fit in step 2, whereas the inclusion of daily diary mood features in step 3 significantly improved overall model fit and reduced BIC. In the final model, adjusted odds for depression increased by 28 % for each additional point on the pre-internship PHQ-9, 75 % for each one-point decline in mean mood score, and 79 % for each one-point increase in mood variability.

The three-step hierarchical logistic regression examining baseline demographic and clinical variables in step 1, Fitbit features in step 2, and daily diary mood data in step 3 as longitudinal predictors of suicidal ideation is presented in Table 4 (ROC curves in Fig. 1). Pre-internship depression and suicidal ideation were significant predictors of suicidal ideation at follow-up in Step 1. Sleep and step features from Fitbit data did not significantly improve the overall model fit in Step 2. The inclusion of mean mood and variability in mood scores in step 3 significantly improved model fit and reduced BIC, and both variables were significant independent predictors of suicidal ideation. In the final model, adjusted odds for suicidal ideation increased by 12 % for each additional point on the PHQ-9, by a factor of 7.93 for those endorsing pre-internship suicidal ideation, 82 % for each one-point decline in mean mood score, and 53 % for each one-point increase in mood variability.[1]

4. Discussion

In a naturalistic study of medical interns, lower average mood ratings and greater mood variability from daily diaries during the first quarter of internship were significant independent predictors of depression and suicidal ideation at the end of the quarter, even after accounting for sociodemographic and pre-internship clinical characteristics. Contrary to expectation, passively collected Fitbit features (e.g., mean-levels, variability, changes over time in sleep and step counts) did not provide incremental improvement in the prediction of suicidal ideation, and provided only a negligible improvement in the prediction of depression. This was the first study to utilize a combination of intensive longitudinal methods and sensor data to predict depression and suicidal ideation in the context of a large, non-clinical sample. While the modest improvement in prediction from daily diary data provides support for use of intensive longitudinal methods to enhance prediction of depression and suicidal ideation during a period of heightened vulnerability to these outcomes, this line of inquiry is still relatively new. Additional research is needed to clarify the optimal assessment methods and the features from mobile and wearable sensors with the greatest potential for improving predictive accuracy.

[1] In order to examine potential influence of temporal overlap, models were also examined with the 14 days prior to the follow-up assessment excluded from the calculation of daily mood and Fitbit features. The AUCs for five of the six models were within 0.001. The step 3 model for depression had a difference of 0.010 (AUC of 0.789 vs 0.799)

Table 3
Three step hierarchical model predicting depression at Q1 follow-up.

	Step 1 (Baseline data)		Step 2 (Passive Fitbit data)		Step 3 (DD mood data)	
	z-score	AOR (95 % CI)	z-score	AOR (95 % CI)	z-score	AOR (95 % CI)
Pre SI	2.08*	1.84 (1.03, 3.28)	2.07*	1.84 (1.03, 3.30)	1.34	1.53 (0.82, 2.86)
Sex	2.01*	1.28 (1.01, 1.64)	1.97*	1.29 (1.00, 1.66)	1.53	1.23 (0.94, 1.60)
Pre PHQ-9	13.41***	1.33 (1.28, 1.39)	13.44***	1.33 (1.28, 1.39)	10.99***	1.28 (1.22, 1.34)
Sleep mean	–		−2.03*	0.89 (0.79, 0.99)	−1.95	0.89 (0.79, 1.00)
Sleep variability	–		1.90	1.21 (0.99, 1.47)	0.99	1.11 (0.90, 1.36)
Steps mean	–		−0.57	0.98 (0.93, 1.04)	−0.28	0.98 (0.93, 1.05)
Steps variability	–		−1.58	0.90 (0.80, 1.02)	−1.50	0.91 (0.79, 1.03)
Mood mean	–		–		−8.60***	0.57 (0.50, 0.65)
Mood variability	–		–		4.07***	1.79 (1.35, 2.38)
Model summary						
Overall χ^2	χ^2 (3) = 267.7, $p < .001$		χ^2 (7) = 281.2, $p < .001$		χ^2 (9) = 421.3, $p < .001$	
Step Δ	–		χ^2 (4) = 13.5, $p = .009$		χ^2 (2) = 140.1, $p < .001$	
Nagelkerke R^2	0.187		0.196		0.285	
Overall AUC	0.738		0.751		0.799	
BIC	1824.3		1841.6		1717.0	

Note. *$p < .05$ **$p < .01$ ***$p < .001$. Depression as an outcome was indicated by a score of 10 or higher on the PHQ-9. Sleep variables were scaled to one hour, step variables were scaled to 1000 steps. DD = daily diary. Pre = Pre-internship. Q1 = Quarter 1 follow-up assessment. SI = Suicidal Ideation. PHQ-9 = Patient Health Questionnaire-9 score. AUC = Area Under the Curve. BIC = Bayesian Information Criterion.

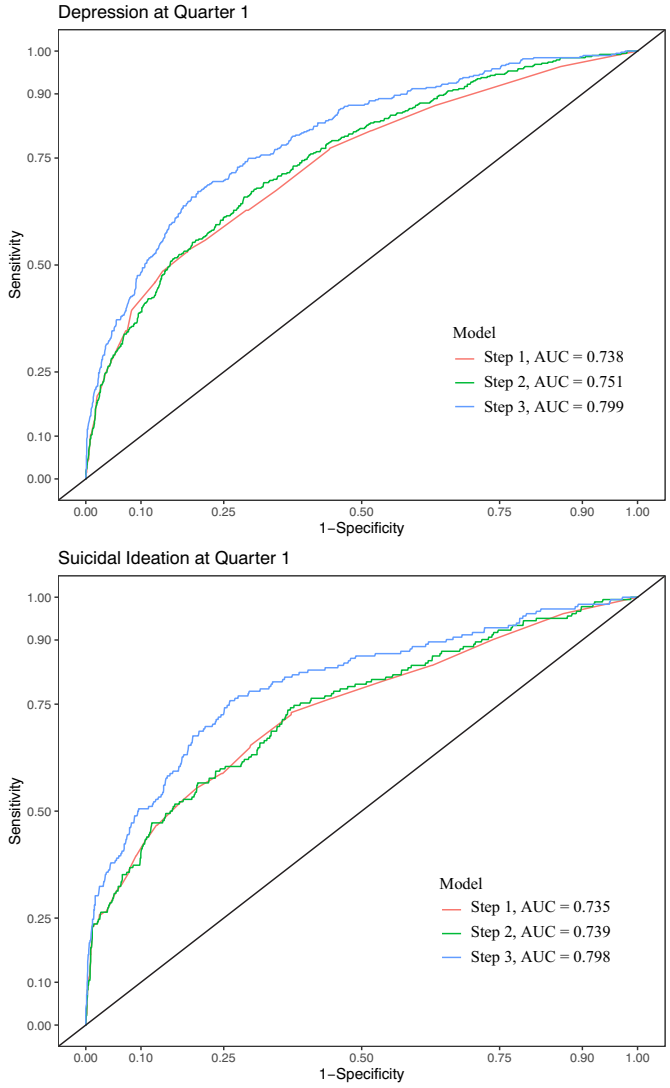

Fig. 1. ROC Curves predicting depression and suicidal ideation.

Consistent with previous studies utilizing intensive longitudinal assessments with clinically severe samples (e.g., Czyz et al., 2020; Peters et al., 2020), we found predictive effects for both mean-level and variability scores of self-report data. The effect of mean mood during the 60-day period over and above pre-internship depression and suicidal ideation may be explained by a closer temporal proximity to the assessment outcomes, and is consistent with previous studies demonstrating the advantages of daily ratings over retrospective recall (Ben-Zeev et al., 2009). Nevertheless, the z-scores in the suicidal ideation model suggested that mean mood score was just as strong of a predictor as baseline suicidal ideation. Variability also had significant main effects for predicting depression and suicidal ideation, suggesting that fluctuations in daily mood scores may be a particularly useful indicator for clinical interventions monitoring and responding to changes in mood. The performance of these features is particularly notable given that, in contrast to most intensive longitudinal studies that incentivize daily assessments to increase compliance (e.g., Rabasco and Sheehan, 2021; Sedano-Capdevila et al., 2021), our findings were informed by single, uncompensated daily mood assessments with a 50 % overall completion rate. This suggests that meaningful improvement in the prediction of suicidal ideation and depression for an at-risk group during a sensitive period can be made with relatively little additional burden or cost.

Daily monitoring of mood facilitated by mobile devices may have considerable implications for identifying early indicators of depression or suicide risk in non-clinical populations, which, in turn, could guide provision of interventions during transitional periods. Importantly, ILMs have significant potential to inform timely and tailored interventions, such as adaptive interventions that are concerned with adjusting the type, intensity, and timing of support based on early signs of risk to improve long-term outcomes while reducing burden (Murphy et al., 2007; Nahum-Shani et al., 2012). Additional research is needed to better understand the earliest periods in which risk can be accurately predicted for at-risk groups, and how to intervene to prevent the negative progression of symptoms.

While passively collected Fitbit data had limited utility in the prediction of depression or suicidal ideation, this is a particularly new area of inquiry, with emerging research illustrating the utility of passive sensing data in predicting changes in depression severity (Ben-Zeev et al., 2015; Pedrelli et al., 2020) and suicidal phenomena (Haines-Delmont et al., 2020; Littlewood et al., 2018). However, our results are consistent with prior work showing that passively assessed metrics do not necessarily outperform self-reported data (Littlewood et al., 2018). It is possible that more fine-grained measurement of passive sensing

Table 4
Three step hierarchical model predicting suicidal ideation at Q1 follow-up.

	Step 1 (Baseline data)		Step 2 (Passive Fitbit data)		Step 3 (DD mood data)	
	z-score	AOR (95 % CI)	z-score	AOR (95 % CI)	z-score	AOR (95 % CI)
Pre SI	7.16***	8.26 (4.64, 14.81)	7.11***	8.22 (4.61, 14.77)	6.46***	7.93 (4.24, 14.94)
Sex	−0.54	0.91 (0.66, 1.27)	−0.62	0.90 (0.64, 1.26)	−0.96	0.84 (0.59, 1.20)
Pre PHQ-9	6.80***	1.19 (1.13, 1.25)	6.79***	1.19 (1.13, 1.25)	4.16***	1.12 (1.06, 1.19)
Sleep mean	–		0.10	1.01 (0.86, 1.21)	0.32	1.03 (0.86, 1.25)
Sleep variability	–		0.49	1.07 (0.80, 1.42)	−0.25	0.96 (0.72, 1.28)
Steps mean	–		−0.01	1.00 (0.93, 1.08)	0.39	1.02 (0.94, 1.09)
Steps variability	–		−0.53	0.96 (0.81, 1.13)	−0.46	0.96 (0.81, 1.14)
Mood mean	–		–		−7.41***	0.55 (0.47, 0.64)
Mood variability	–		–		2.27*	1.53 (1.06, 2.20)
Model summary						
Overall χ^2	χ^2 (3) = 168.0, $p < .001$		χ^2 (7) = 168.7, $p < .001$		χ^2 (9) = 243.8, $p < .001$	
Step Δ	–		χ^2 (4) = 0.7, $p = .953$		χ^2 (2) = 84.9, $p < .001$	
Nagelkerke R^2	0.168		0.168		0.249	
Overall AUC	0.735		0.739		0.798	
BIC	1125.0		1155.2		1085.7	

Note. *$p < .05$ **$p < .01$ ***$p < .001$. Sleep variables were scaled to one hour, step variables were scaled to 1000 steps. DD = Daily diary. Pre = Pre-internship. Q1 = Quarter 1 follow-up assessment. SI = Suicidal Ideation. PHQ-9 = Patient Health Questionnaire-9 score. AUC = Area Under the Curve. BIC = Bayesian Information Criterion.

features beyond sleep duration and variability, such as changes in sleep and wake times (Fang et al., 2021; Wang et al., 2018), is needed to improve prediction of depression of suicidal ideation beyond self-reported mood ratings. Likewise, features for physical activity (e.g., physical activity minutes at varying intensity levels) from newer models of wearable sensors may be able to improve predictive accuracy. Importantly, these passive sensing studies have varied on length of time data is being aggregated, sources of passive data, and time to predicted outcome (e.g., next day, next week, next month). Further research is needed to examine the potential utility of these data and the contexts in which passive sensing may provide an advantage in risk detection and prediction.

This study had several strengths, including a large sample size, use of a non-clinical population, and an intensive longitudinal design. However, findings should be interpreted in the context of the study's limitations. While medical interns in the study were participating from hundreds of programs across the United States, there is inherent heterogeneity as a function of education and occupation that limits generalizability to other populations. Furthermore, internship is characterized by long and irregular work hours, and the impact of these work responsibilities may have resulted in sleep and step patterns that are not as representative of the expected associations of sleep and steps with depression or suicidal ideation in other samples. Interns who did not complete the follow-up survey had higher scores of baseline depression and were more likely to endorse suicidal ideation. Thus, our analytic sample may represent a less clinically severe sample of interns. While the ninth item from the PHQ-9 has been widely used as an indicator of suicidal ideation (e.g., Rossom et al., 2017), it is somewhat nonspecific and does not capture important constructs such as suicidal intent or plan. Additionally, while Fitbit has undergone successful validation studies in the laboratory (e.g., de Zambotti et al., 2018), other studies have identified concerns related to accuracy (e.g., Feehan et al., 2018). Inconsistency in use, such as wearing a Fitbit for only part of the day, may have generated error that contributed to a non-significance in findings related to depression and suicidal ideation, and additional validation studies conducted in naturalistic settings are warranted. Finally, while previous ILM studies have typically focused on near-term (i.e., next-day) suicidal ideation, our study assessed suicidal ideation on a quarterly basis, preventing more near-term predictions. Nevertheless, this unique and large sample allowed us to examine the utility of intensive longitudinal data in a naturalistic setting with a sample at elevated risk for the development of depression and suicidal ideation.

In conclusion, our findings indicated that mobile monitoring of daily mood improves the prediction of depression and suicidal ideation during a meaningful at-risk period (first quarter of internship) under naturalistic conditions (e.g., without compensation), demonstrating potential for reach outside of research settings. Additional research is needed to address key questions that can further guide the development of adaptive interventions among less clinically severe, but nevertheless vulnerable, populations. For example, future research should identify the earliest period (e.g., first two or four weeks after start of internship) at which daily or sensor data demonstrate sufficiently strong prediction of adverse follow-up outcomes and examine different approaches to processing intensive longitudinal data (e.g., different features) to improve prediction. Taken together, mobile technology has the potential to capture time-sensitive markers of adverse outcomes and make an impact on facilitating risk detection and timely interventions in naturalistic settings, particularly among populations at risk for the development of symptoms.

CRediT authorship contribution statement

A. Horwitz led the conceptualization and design of the study, the writing of the manuscript, and the primary analyses. E. Czyz co-led the conceptualization of the study, the analytic design, and contributed to writing the introduction and discussion sections. N. Al-Dajani assisted with the primary analyses, generated study figures, and contributed to writing the introduction and discussion sections. W. Dempsey contributed to the analytic design and conceptualization of the study. Z. Zhao contributed to the data preparation and assisted with the primary analyses and interpretation. I. Nahum-Shani contributed to the conceptualization and design of the study. S. Sen led the conceptualization of the overarching research project and contributed to the conceptualization and design of the study. All authors provided a critical review of the manuscript prior to submission.

Conflict of Interest

Authors for this manuscript have no conflicts of interest to declare.

Role of the funding source

This work was supported by awards from the National Institute of Health (R01 MH101459 to Srijan Sen; K23 MH113776 to Ewa Czyz) and the National Center for Advancing Translational Sciences (KL2 TR002241 to Adam Horwitz). The content of this manuscript is solely

the responsibility of the authors and does not represent the views of the funders, who had no role in the data analysis, interpretation, or preparation of this manuscript.

Acknowledgements

Authors wish to thank the medical interns who participated in this study.

References

Aguilera, A., Schueller, S.M., Leykin, Y., 2015. Daily mood ratings via text message as a proxy for clinic based depression assessment. J. Affect. Disord. 175, 471–474.
Ben-Zeev, D., Young, M.A., Madsen, J.W., 2009. Retrospective recall of affect in clinically depressed individuals and controls. Cognit. Emot. 23, 1021–1040.
Ben-Zeev, D., Scherer, E.A., Wang, R., Xie, H., Campbell, A.T., 2015. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 38, 218–226.
Czyz, E.K., Horwitz, A.G., Arango, A., King, C.A., 2019. Short-term change and prediction of suicidal ideation among adolescents: a daily diary study following psychiatric hospitalization. J. Child Psychol. Psychiatry Allied Discip. 60, 732–741.
Czyz, E.K., Yap, J.R., King, C.A., Nahum-Shani, I., 2020. Using intensive longitudinal data to identify early predictors of suicide-related outcomes in high-risk adolescents: practical and conceptual considerations. Assessment 1073191120939168.
de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I.M., Baker, F.C., 2018. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. Chronobiol. Int. 35, 465–476.
Evenson, K.R., Goto, M.M., Furberg, R.D., 2015. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int. J. Behav. Nutr. Phys. Act. 12, 1–22.
Fang, Y., Forger, D.B., Frank, E., Sen, S., Goldstein, C., 2021. Day-to-day variability in sleep parameters and depression risk: a prospective cohort study of training physicians. NPJ Digit. Med. 4, 1–9.
Feehan, L.M., Geldman, J., Sayre, E.C., Park, C., Ezzat, A.M., Yoo, J.Y., Hamilton, C.B., Li, L.C., 2018. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR mHealth and uHealth 6, e10527.
Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., Nock, M.K., 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol. Bull. 143, 187.
Goldman, M.L., Shah, R.N., Bernstein, C.A., 2015. Depression and suicide among physician trainees: recommendations for a national response. JAMA Psychiatry 72, 411–412.
Gratch, I., Choo, T.H., Galfalvy, H., Keilp, J.G., Itzhaky, L., Mann, J.J., Oquendo, M.A., Stanley, B., 2021. Detecting suicidal thoughts: the power of ecological momentary assessment. Depression Anxiety 38, 8–16.
Greenberg, P.E., Fournier, A.-A., Sisitsky, T., Pike, C.T., Kessler, R.C., 2015. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J. Clin. Psychiatry 76, 155–162.
Haines-Delmont, A., Chahal, G., Bruen, A.J., Wall, A., Khan, C.T., Sadashiv, R., Fearnley, D., 2020. Testing suicide risk prediction algorithms using phone measurements with patients in acute mental health settings: feasibility study. JMIR Mhealth Uhealth 8, e15901.
Hasin, D.S., Sarvet, A.L., Meyers, J.L., Saha, T.D., Ruan, W.J., Stohl, M., Grant, B.F., 2018. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. JAMA psychiatry 75, 336–346.
Hawton, K., i Comabella, C.C., Haw, C., Saunders, K., 2013. Risk factors for suicide in individuals with depression: a systematic review. J. Affect. Disord. 147, 17–28.
Hedegaard, H., Curtin, S., Warner, M., 2020. Increase in suicide mortality in the United States, 1999–2018. NCHS Data Brief 1–8.
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. J. Abnorm. Psychol. 126, 726.
Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613.
Littlewood, D.L., Kyle, S.D., Carter, L.-A., Peters, S., Pratt, D., Gooding, P., 2018. Short sleep duration and poor sleep quality predict next-day suicidal ideation: an ecological momentary assessment study. Psychol. Med. 49, 403–411.
Mata, D.A., Ramos, M.A., Bansal, N., Khan, R., Guille, C., Di Angelantonio, E., Sen, S., 2015. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. JAMA 314, 2373–2383.
Murphy, S.A., Collins, L.M., Rush, A.J., 2007. Customizing treatment to the patient: adaptive treatment strategies. Drug Alcohol Depend. 88, S1.
Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W.E., Gnagy, B., Fabiano, G.A., Waxmonsky, J.G., Yu, J., Murphy, S.A., 2012. Experimental design and primary data analysis methods for comparing adaptive interventions. Psychol. Methods 17, 457.
NeCamp, T., Sen, S., Frank, E., Walton, M.A., Ionides, E.L., Fang, Y., Tewari, A., Wu, Z., 2020. Assessing real-time moderation for developing adaptive Mobile health interventions for medical interns: micro-randomized trial. J. Med. Internet Res. 22, e15033.
Or, F., Torous, J., Onnela, J.-P., 2017. High potential but limited evidence: using voice data from smartphones to monitor and diagnose mood disorders. Psychiatric Rehabil. J. 40, 320–324.
Pedrelli, P., Fedor, S., Ghandeharioun, A., Howe, E., Ionescu, D.F., Bhathena, D., Fisher, L.B., Cusin, C., Nyer, M., Yeung, A., 2020. Monitoring changes in depression severity using wearable and mobile sensors. Front. Psychiatry 11, 1413.
Peters, E.M., Dong, L.Y., Thomas, T., Khalaj, S., Balbuena, L., Baetz, M., Osgood, N., Bowen, R., 2020. Instability of suicidal ideation in patients hospitalized for depression: an exploratory study using smartphone ecological momentary assessment. Arch. Suicide Res. 1–14.
Rabasco, A., Sheehan, K., 2021. The use of intensive longitudinal methods in research on suicidal thoughts and behaviors: a systematic review. Arch. Suicide Res. 1–15.
Rossom, R.C., Coleman, K.J., Ahmedani, B.K., Beck, A., Johnson, E., Oliver, M., Simon, G. E., 2017. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. J. Affect. Disord. 215, 77–84.
Schoevers, R., Van Borkulo, C., Lamers, F., Servaas, M., Bastiaansen, J., Beekman, A., Van Hemert, A., Smit, J., Penninx, B., Riese, H., 2020. Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. Psychol. Med. 1–10.
Schuch, F.B., Vancampfort, D., Firth, J., Rosenbaum, S., Ward, P.B., Silva, E.S., Hallgren, M., Ponce De Leon, A., Dunn, A.L., Deslandes, A.C., 2018. Physical activity and incident depression: a meta-analysis of prospective cohort studies. Am. J. Psychiatr. 175, 631–648.
Sedano-Capdevila, A., Porras-Segovia, A., Bello, H.J., Baca-García, E., Barrigon, M.L., 2021. Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Curr. Psychiatry Rep. 23, 41.
Sen, S., Kranzler, H.R., Krystal, J.H., Speller, H., Chan, G., Gelernter, J., Guille, C., 2010. A prospective cohort study investigating factors associated with depression during medical internship. Arch. Gen. Psychiatry 67, 557–565.
Targum, S.D., Sauder, C., Evans, M., Saber, J.N., Harvey, P.D., 2021. Ecological momentary assessment as a measurement tool in depression trials. J. Psychiatr. Res. 136, 256–264.
Tazawa, Y., Liang, K.-C., Yoshimura, M., Kitazawa, M., Kaise, Y., Takamiya, A., Kishi, A., Horigome, T., Mitsukura, Y., Mimura, M., 2020. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon 6, e03274.
Wang, R., Wang, W., DaSilva, A., Huckins, J.F., Kelley, W.M., Heatherton, T.F., Campbell, A.T., 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2, pp. 1–26.
World Health Organization, 2017. Depression and Other Common Mental Disorders; Global Health Estimates. Switzerland, Geneva.
Yim, S.J., Lui, L.M., Lee, Y., Rosenblat, J.D., Ragguett, R.-M., Park, C., Subramaniapillai, M., Cao, B., Zhou, A., Rong, C., 2020. The utility of smartphone-based, ecological momentary assessment for depressive symptoms. J. Affect. Disord. 274, 602–609.
Zhai, L., Zhang, H., Zhang, D., 2015. Sleep duration and depression among adults: a meta-analysis of prospective studies. Depress. Anxiety 32, 664–670.