

Building personalized machine learning models using real-time monitoring data to predict idiographic suicidal thoughts

Received: 10 May 2023

Accepted: 19 September 2024

Published online: 24 October 2024

 Check for updates

Shirley B. Wang  ¹✉, Ruben D. I. Van Genugten², Yaniv Yacoby  ³, Weiwei Pan⁴, Kate H. Bentley^{5,6,7}, Suzanne A. Bird⁷, Ralph J. Buonopane⁸, Alexis Christie⁵, Merryn Daniel¹⁵, Dylan DeMarco⁵, Adam Haim⁹, Lia Follet¹⁰, Rebecca G. Fortgang^{5,6,7}, Flynn Kelly-Brunyak¹¹, Evan M. Kleiman¹², Alexander J. Millner^{5,8}, Onyinye Obi-Obasi¹³, J. P. Onnella¹⁴, Narise Ramlal⁵, Jordyn R. Ricard  ¹, Jordan W. Smoller^{6,7}, Tida Tambedou⁵, Kelly L. Zuromski^{5,8} & Matthew K. Nock⁵

Suicide risk is highest immediately after psychiatric hospitalization, but the field lacks methods for identifying which patients are at greatest risk, and when. We built personalized models predicting suicidal thoughts after psychiatric hospital visits ($N = 89$ patients), using ecological momentary assessment (EMA; average EMA responses per participant = 311). We built several idiographic models, including baseline autoregressive and elastic net models (using single train/test split) and Gaussian process (GP) models (using an iterative rolling-forward prediction method). Simple GP models provided the best prediction of suicidal urges ($R^2_{\text{average}} = 0.17$), outperforming baseline autoregressive ($R^2_{\text{average}} = 0.10$) and elastic net ($R^2_{\text{average}} = 0.06$) models. Similarly, simple GP models provided the best prediction of suicidal intent ($R^2_{\text{average}} = 0.12$) compared to autoregressive ($R^2_{\text{average}} = 0.08$) and elastic net ($R^2_{\text{average}} = 0.04$). Here we show that idiographic prediction of suicidal thoughts is possible, although the accuracy is currently modest. Building GP models that iteratively update and learn symptom dynamics over time could provide important information to inform the development of just-in-time adaptive interventions.

Suicide is a leading cause of death worldwide. Over 700,000 people die by suicide each year, and suicidal thoughts are even more common, with a cross-national lifetime prevalence estimate of ~9.2% (refs. 1–3). Unfortunately, until recently, our ability to prospectively predict suicidal thoughts and behaviors has remained at near-chance levels⁴. This limited predictive power is due—at least in part—to a historical reliance on retrospective data (typically interviews or self-reports at one or two single time points in hospital or laboratory settings), which cannot provide information on the dynamic nature of suicide risk as it unfolds in people's natural environments⁴. Fortunately, recent

advances in ecological momentary assessment (EMA) smartphone monitoring methods have shown promise in gathering rich and granular data on suicidal thoughts and behaviors in the real world, in real time. This work has demonstrated that dynamic fluctuations and variability in suicidal thoughts over time are important for predicting the risk of future suicide attempts, and that real-time suicidal thoughts (and their putative risk factors) are characterized by substantial within- and between-person heterogeneity^{5–7}.

Most EMA studies have used between-subjects methods to study the relationships between risk factors and suicidal thoughts and

behaviors on average, when pooling data across participants. Of note, this is the case even for multilevel (that is, mixed-effects) models, which are often employed to analyze time-series data, including EMA data. However, these are not truly idiographic models, as they only estimate individual-level effects as they differ from group-level averages^{8,9}. This reflects larger trends in empirical psychopathology research, which tends to rely on between-person analysis methods, despite most theories focusing on individual-level processes¹⁰. Group-level prediction models (built with out-of-sample predictions, and especially those that are externally validated) are useful for answering the question of who might be at risk for suicide in any given time period. Ultimately, however, we are also interested in identifying when a specific person may be at risk, to determine when and how to optimally deliver interventions for individual patients. Importantly, unless a process is ergodic (that is, invariant in processes across all individuals in a group and across time), findings detected at the group level cannot be readily assumed to apply to the level of the individual¹¹. Unfortunately, most processes of interest in psychopathology rarely meet these requirements, with notable between-person differences in intra-individual mechanisms. For instance, across six EMA datasets, Fisher and colleagues demonstrated a striking lack of group-to-individual generalizability in processes including negative affect, positive affect, depressed mood and anhedonia¹². In such cases, when the underlying data-generating process that produces psychopathology symptoms may differ meaningfully across participants, building idiographic prediction models may be crucial for anticipating and intervening upon risk (for example, for suicide) in the real world, in real time.

Psychological scientists have long called for idiographic research methods^{13–15}, with such calls receiving renewed enthusiasm as the field of clinical psychology moves towards personalized-medicine approaches^{8,16,17}. EMA methods allow for repeated measurements of within-person processes over time, and offer a unique opportunity to build personalized models. To date, one study has used EMA data to develop idiographic models of suicidal thoughts, finding evidence for high heterogeneity between risk factors (for example, negative affect, impulsivity) and suicidal thoughts across participants, in both the strength and direction of relationships¹⁸. The heterogeneity found in this explanatory approach (which asks ‘what factors are associated with suicidal thinking for this specific person?’) suggests that it may also be important to take an idiographic approach to suicide prediction (to ask ‘can we predict when this specific person is at risk?’). Indeed, idiographic machine learning models have recently shown promise in predicting individual-level alcohol use¹⁹, tobacco smoking²⁰ and loneliness²¹. If these methods can be applied to predict a person’s suicidal thoughts as they unfold in the real world, this could provide crucial information to use to deploy timely, targeted interventions to individuals when help is most needed (through just-in-time adaptive interventions)²².

This study leveraged EMA data to build personalized machine learning models predicting individual-level suicidal thoughts, including suicidal urges and suicidal intent. We explored two primary methods for doing so. First, following existing methods in idiographic psychopathology prediction^{19,20,23}, we split each individual patient’s time series into one training and one testing dataset. With this approach, we trained a machine learning model on the first 50% of a participant’s data, and tested on the held-out 50%. Second, we used an iterative rolling-forward approach to (1) train a machine learning model on an initial subset of EMA data provided by each individual patient, (2) predict their next (held-out) response, then (3) continuously update the model to condition on newly observed data points and predict each subsequent response. For both modeling approaches, we tested models with participants’ history of EMA suicidal thoughts only (that is, lagged suicidal thoughts), as well as both lagged suicidal thoughts and lagged affect (for example, angry, sad, hopeless, worried) as features. Finally, as a secondary analysis, we explored potential differences in

Table 1 | Participant demographics and history of suicidal thoughts and behaviors

Demographic/clinical characteristics	n (%)	M (s.d.)
Age		22.10 (10.02)
Race		
White	64 (80%)	
Black	10 (12.5%)	
Asian	9 (11.3%)	
American Indian/Alaskan Native	1 (1.3%)	
Native Hawaiian/Pacific Islander	1 (1.3%)	
Other/not listed	4 (5.0%)	
Ethnicity		
Hispanic or Latino/a	10 (12.5%)	
Gender		
Male	18 (22.5%)	
Female	47 (58.8%)	
Non-binary/gender-nonconforming	9 (11.3%)	
Transgender	5 (6.25%)	
Other	2 (2.5%)	
Suicidal thoughts (lifetime presence)	75 (92.6%)	
Past year frequency (weeks)		22.89 (27.91)
Past month frequency (days)		13.76 (11.15)
Past week frequency (days)		3.52 (2.63)
Suicide plan (lifetime presence)	61 (75.3%)	
Past year frequency (weeks)		10.61 (14.68)
Past month frequency (days)		7.68 (8.68)
Past week frequency (days)		2.04 (2.38)
Suicide attempt (lifetime presence)	50 (61.7%)	
Past year frequency		2.09 (2.59)
Past month frequency		0.86 (0.97)
Past week frequency		0.74 (0.93)

Some participants in the present sample (N=89) did not complete the baseline demographics and ‘self-injurious thoughts and behaviors interview’ (SITBI) questionnaires. Therefore, the demographics and SITBI data include data from 81 patients.

idiographic prediction accuracy by dynamics of suicidal thinking, to examine whether our models provided more accurate predictions for individuals with certain patterns of EMA suicidal thoughts over time.

Results

Demographic and clinical characteristics

Table 1 presents the demographic and clinical characteristics of the participants. During the study they completed an average of 311.2 EMA surveys (s.d. = 118.60) over an average of 71.4 days (s.d. = 18.58). Average EMA compliance was 72% (s.d. = 1.7%).

Model performance

Predicting suicidal urges. Model performance metrics for the baseline (using a single train/test split) and Gaussian process (GP; using an iterative rolling-forward approach) models predicting suicidal urges are shown in Fig. 1. Baseline autoregressive models (using only lagged suicidal urges) yielded a fair prediction overall, with an average R^2 of 0.10 ($n = 22$ participants above a medium effect size of $R^2 = 0.13$)²⁴, and wide variability in predictive accuracy across participants (s.d. = 0.13, range = <0.01–0.72). Interestingly, baseline elastic net models with additional lagged affect features did not improve prediction, with an

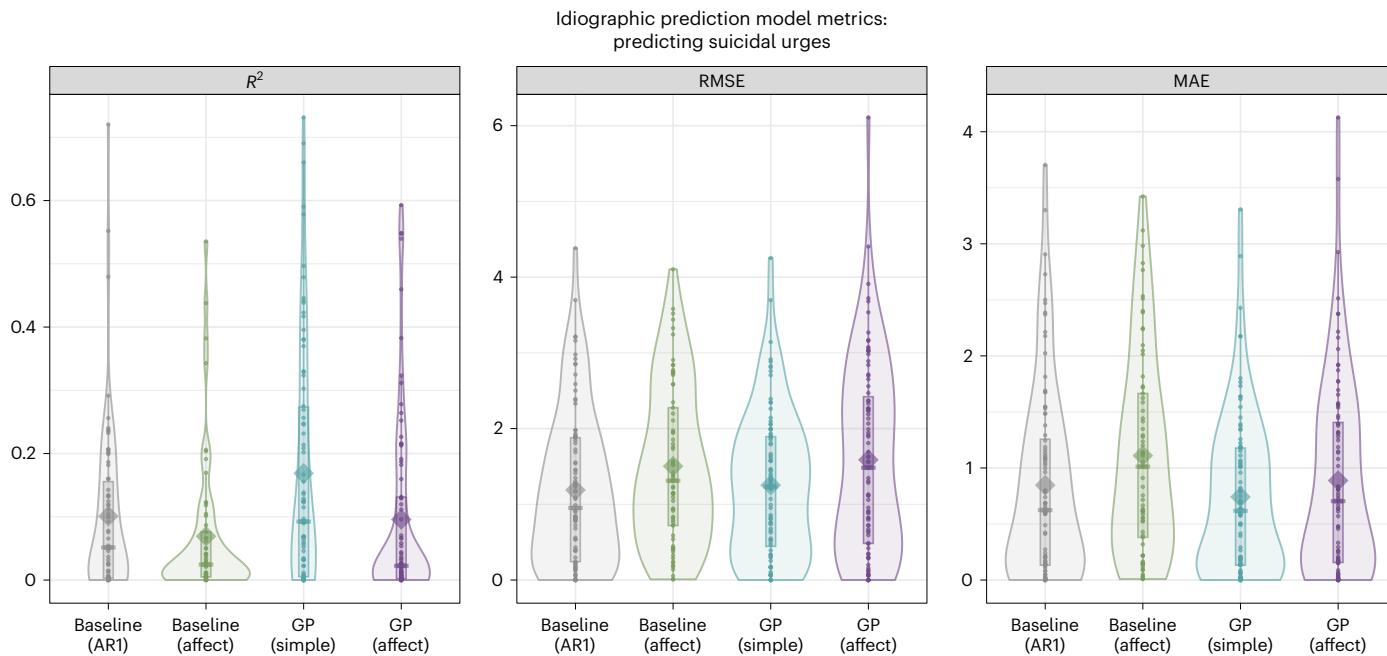


Fig. 1 | Each panel represents descriptive visualizations (violin plots) of the distribution of idiographic model performance predicting suicidal urges. Small dots represent idiographic results for each participant ($N = 88$ total participants). The darker shaded boxes in each violin plot are box plots (where the lower and upper boundaries of the box represent the first and third quartiles, and the whiskers extend to $1.5 \times$ interquartile range, the shaded diamond

marker represents the mean across participants, and the shaded horizontal bar represents the median across participants). Note: for visualization, the y axis for the RMSE plot is truncated, and does not show one higher value ($RMSE = 14.02$); this outlier remains reflected in the summary statistics (mean, median and so on). AR, autoregressive.

average R^2 of 0.06 across participants ($s.d. = 0.11$, range = <0.01–0.54; $n = 10$ participants above medium effect size). Simple GP models (using only the history of suicidal urges) meaningfully improved the prediction overall, with an average R^2 of 0.17 ($s.d. = 0.19$, range = <0.01–0.73, $n = 33$ participants above medium effect size). However, similar to the baseline models, GP models with additional affect features did not improve prediction, with an average R^2 of 0.10 ($s.d. = 0.15$, range = <0.01–0.59, $n = 22$ participants above medium effect size). There were significant differences in R^2 values between the four models ($\chi^2(3) = 52.16$, $P < 0.001$), with the simple GP model outperforming all three others (the baseline autoregressive model ($P < 0.001$), the baseline model with affect ($P < 0.001$) and the GP model with affect ($P < 0.001$)). Other smaller differences also emerged, with the baseline autoregressive model ($P = 0.04$) and the GP model with affect ($P = 0.04$) outperforming the baseline model with affect. Model differences in root-mean-squared error (r.m.s.e.) and mean absolute error (MAE) metrics were less pronounced than those for R^2 (Fig. 1).

For illustrative purposes, idiographic prediction results for one sample patient are shown in Fig. 2 (idiographic predictions for all patients are available at https://github.com/ShirleyBWang/idiographic_prediction ref. 25). As illustrated in Fig. 2 and shown in the idiographic prediction figures in the Supplementary Information, the baseline models tended to provide 'flatter' predictions, with less variability. On the other hand, the GP models tended to provide predictions spanning the entire range of observed data, with more dynamic fluctuations.

Predicting suicidal intent. Results for idiographic models predicting suicidal intent largely mirrored those for models predicting suicidal urges, though with poorer performance overall (all performance metrics shown in Fig. 3). Specifically, the baseline autoregressive models (using only lagged suicidal intent) and the baseline elastic net models (with additional lagged affect features) yielded an average R^2 of 0.08 ($s.d. = 0.10$, range = <0.01–0.37, $n = 11$ participants above medium effect

size) for the simple autoregressive models, and an average R^2 of 0.05 ($s.d. = 0.04$, range = <0.01–0.26, $n = 6$ participants above medium effect size) for the elastic net models. Simple GP models (using history of suicidal intent only) improved prediction beyond both of the baseline models, with an average R^2 of 0.12 ($s.d. = 0.16$, range = <0.01–0.69, $n = 23$ participants above medium effect size). However, GP models including additional lagged affect features provided similar prediction as baseline models, with an average R^2 of 0.08 ($s.d. = 0.06$, range = <0.01–0.38, $n = 15$ participants above medium effect size). There were significant differences in R^2 values between the four models ($\chi^2(3) = 43.07$, $P < 0.001$), with the simple GP model again outperforming all three others (the baseline autoregressive model ($P < 0.001$), the baseline model with affect ($P < 0.001$) and the GP model with affect ($P < 0.001$)). Other model performance metrics (MAE, r.m.s.e.) were quite similar between the baseline and GP models on average (Fig. 3).

For illustrative purposes, idiographic prediction results for one sample patient are shown in Fig. 4 (figures displaying idiographic predictions for all patients are available at https://github.com/ShirleyBWang/idiographic_prediction ref. 25). Similar to suicidal urge models, baseline models predicting suicidal intent tended to yield flatter predictions (with a narrower range of predicted values compared to observed values), whereas the GP models produced predictions with more dynamic fluctuations, which better reflected (but sometimes exceeded) the full range of observed EMA data.

Model performance and suicidal thinking dynamics. Correlations between EMA features and model performance metrics are shown in Supplementary Figs. 1–8. Across all baseline and GP models, for both suicidal urges and intent, higher R^2 values were consistently associated with greater marginal entropy, marginal mean and with lower percent zero responses. Across all GP models (with and without affect), for both suicidal urges and intent, higher R^2 values were also associated with higher percent big jumps, percent high EMA responses, and ratio of percent high responses to marginal entropy, and with lower percent

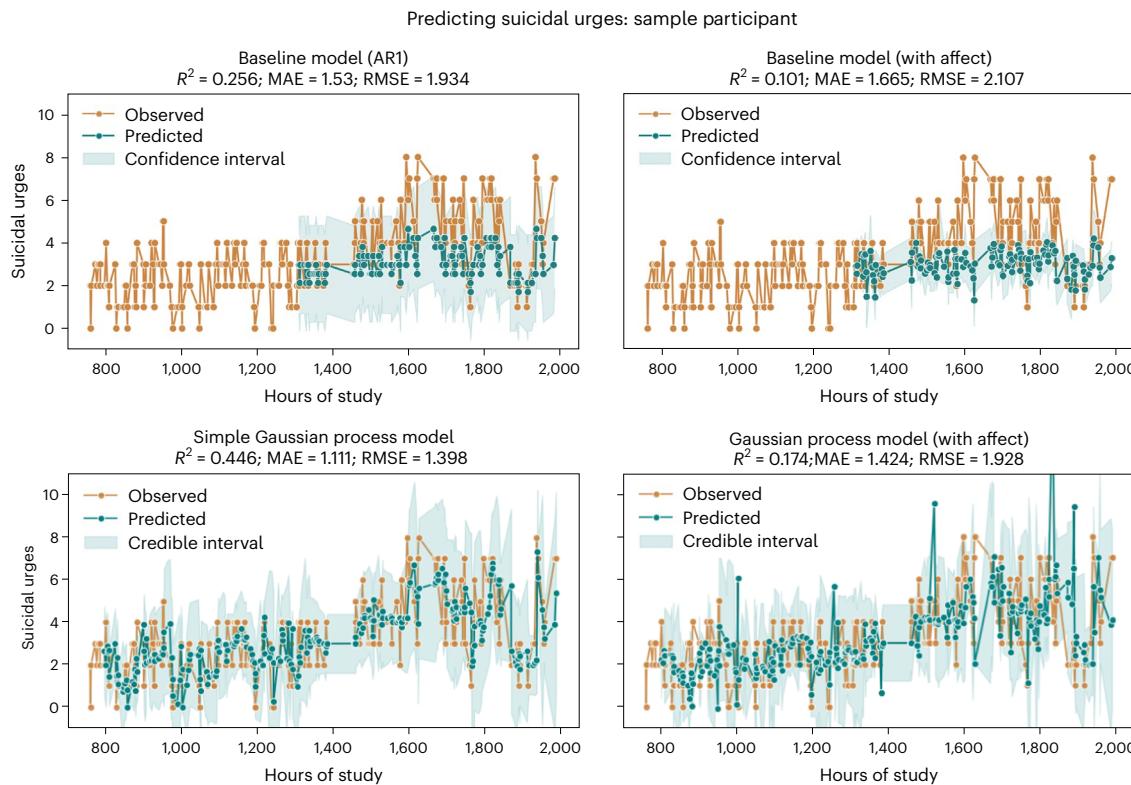


Fig. 2 | Illustrative example of idiographic prediction of suicidal urges for one participant. Illustrative example, modeled using (1) baseline (single train/test split) simple autoregressive models, (2) baseline elastic net models with

additional lagged affect features, (3) simple GP models (using iterative rolling-forward prediction) with history of suicidal urges only and (4) GP models with additional lagged affect features.

small jumps. Interestingly, there was a similar pattern of correlations between EMA features and MAE and RMSE metrics (that is, higher values of MAE and r.m.s.e. correlated with greater marginal entropy, marginal mean and lower percent zero responses).

Discussion

In this Article we built personalized prediction models to forecast the severity of future suicidal urges and intent for individual participants. There were three key findings. First, on average, simple GP models provided better idiographic prediction of suicidal thoughts than the baseline autoregressive and elastic net models. Specifically, compared to baseline models, GP models tended to provide predictions with more dynamic fluctuations, and that better reflected the full range of observed data for suicidal thoughts. Second, the addition of affective and cognitive-affective EMA items did not improve prediction for either the baseline or GP models. Third, idiographic models tended to provide more accurate predictions for patients whose suicidal thoughts were characterized by higher spread (that is, a more equal distribution of responses across all Likert-scale responses) and higher average severity, and poorer predictions for patients with highly zero-inflated data. Each of these findings merits further discussion.

On average across participants, simple GP models using an iterative rolling-forward prediction method provided better prediction than baseline autoregressive and elastic net models (using a single train/test split, as is common in existing idiographic prediction methods in psychopathology research^{19,20}). At the individual level, we also observed the largest number of participants with an above medium effect size for prediction of both suicidal urges and intent (that is, $R^2 > 0.13$; Figs. 1 and 3)²⁴ with simple GP models. This suggests that the dynamics of participants' suicidal thinking may be changing over time, such that a model trained on the first half of a participant's data (for example,

weeks 1–6) may not generalize to the second half of their data (for example, weeks 7–12), and would not respond dynamically to any changes in suicidal thinking that occur beyond the baseline period, substantially limiting the clinical utility of such models. Rather, iteratively training and updating models (as more data are collected) may better capture and predict real-time suicidal thoughts that evolve over time. GP models are especially well-suited for this problem of modeling complex functions and updating over time^{26–28}, even with limited initial training data (for example, $N = 10$ EMA observations), as Bayesian methods are well-equipped to work with small sample sizes.

Our findings are encouraging, because iteratively developed models may also be especially useful in clinical and community settings. For instance, models that can more quickly start making predictions (for example, after several days, rather than several weeks) could provide important data to inform clinical care almost immediately (for example, by alerting clinicians to potentially high-risk time periods). Of note, daily symptom monitoring is already a component of many existing interventions (for example, self-monitoring records in cognitive-behavioral therapy, diary cards in dialectical behavioral therapy), and recent efforts to integrate digital clinics into routine healthcare settings offer exciting opportunities to collect more granular time-series data in a sustainable and scalable manner^{29,30}. As the nature of individuals' suicidal thinking is likely to change over the course of treatment, models that continuously update and learn symptom dynamics as new data are observed may provide more accurate predictions over time.

The benefit of using iterative rolling-forward GP models for idiographic prediction of suicidal thoughts (beyond our baseline autoregressive and elastic net models) was readily observable when visualizing the model predictions and observed EMA data (Figs. 2 and 4 present illustrative examples; all figures are available at https://github.com/ShirleyBWang/idiographic_prediction.git)²⁵. Specifically,

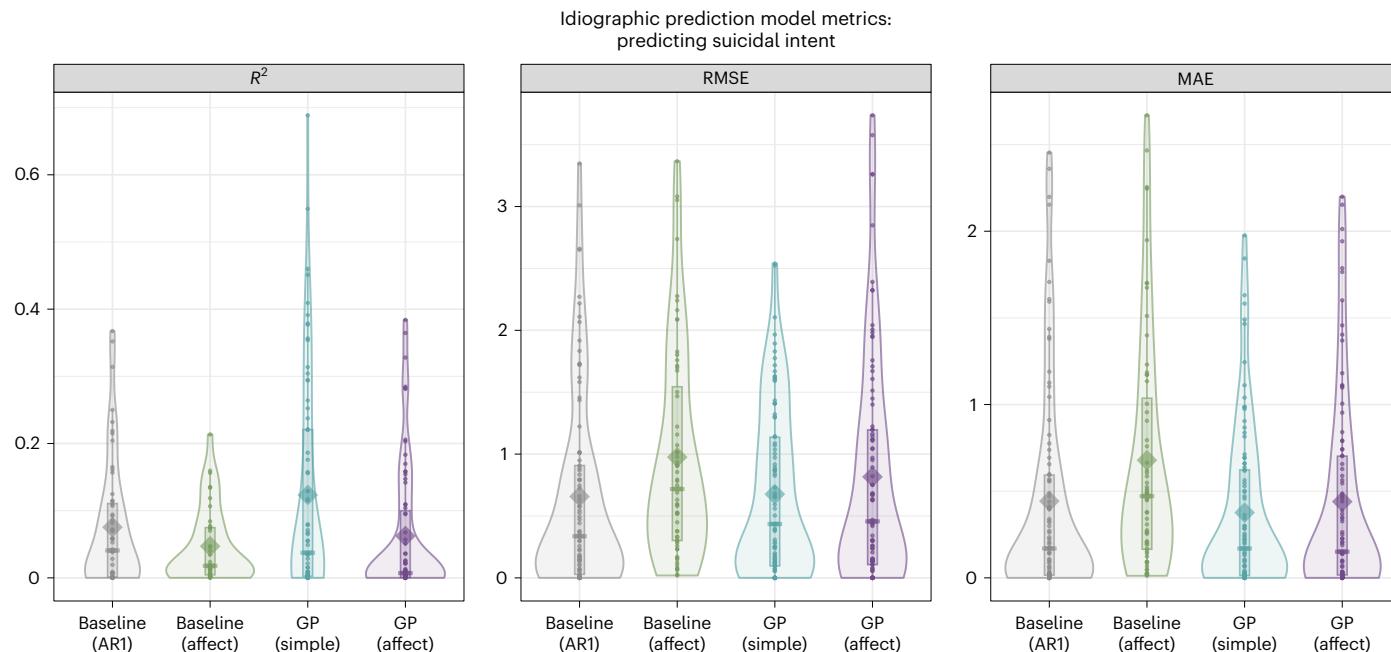


Fig. 3 | Descriptive visualizations (violin plots) of the distribution of idiographic model performance predicting suicidal intent. Small dots represent idiographic results for each participant ($N = 89$ total participants). The darker shaded boxes in each violin plot are box plots (where the lower and upper boundaries of the box represent the first and third quartiles, and the whiskers

extend to $1.5 \times$ interquartile range, the shaded diamond marker represents the mean across participants, and the shaded horizontal bar represents the median across participants). Note: for visualization, the y axis for the RMSE plot is truncated, and does not show one higher value (RMSE = 33.85); this outlier remains reflected in the summary statistics (mean, median and so on).

GP models tended to provide predictions with more dynamic fluctuations, reflecting the full range of EMA data, whereas baseline models tended to yield more stable predictions with less variability. This may also explain our observed difference in quantitative performance metrics, with differences emerging in R^2 —but not MAE or r.m.s.e.—between the GP and baseline models. Considering these metrics in concert with data visualizations, it appears that GP models provided predictions that fluctuated in tandem with the observed data (reflected in a higher squared correlation coefficient between the observed and predicted data). On the other hand, the baseline models provided predictions that were largely concentrated around the mean of the observed data, yielding relatively low (that is, good) MAE and r.m.s.e. metrics, despite not fully capturing the true dynamic fluctuations in suicidal thinking.

Interestingly, adding additional affective and cognitive-affective EMA did not improve idiographic prediction for either the baseline or GP models. Rather, models including affect seemed to produce equal or poorer prediction than models including suicidal thoughts only. One potential explanation for this finding involves the time-scale of the expected causal effects of affective states on suicidal thoughts. Theoretically, we might expect affect to predict momentary suicidal thoughts over a short timeframe (for example, increased feelings of agitation and hopelessness predicting increased suicidal thoughts over the next hour). Indeed, emerging evidence suggests that suicidal thoughts are transient and change rapidly, with elevated suicidal thoughts lasting only 1–3 hours on average³¹. However, in the present study, the average time between EMA responses was 5.33 hours. Over this longer timeframe, the added predictive power of lagged affect (beyond the predictive power of lagged suicidal thoughts alone) may be relatively weak or nonexistent, and elastic net may not have provided sufficient regularization to prevent overfitting in these models. Future research with higher-density sampling methods may provide increased granularity for affective features (and other relevant variables, such as stress level/context) to improve the prediction of short-term future suicidal thoughts. For instance, novel micro-EMA (μ EMA) methods

involve single-question prompts delivered on smartwatches (rather than smartphones), and enable much higher-density sampling—up to six prompts per hour, rather than six prompts per day—with higher compliance and lower reported participant burden^{32,33}. This offers an exciting new possibility for gathering high-resolution data on suicidal thoughts and their putative risk factors, on timescales that may better align with our theoretical models and yield more accurate prediction.

As a secondary aim to examine differences in model performance across participants, we examined the relationship between the model metrics (R^2 , MAE, RMSE) and time-series features of each patient's suicidal urges and intent. Notably, all models performed more poorly for individuals with greater zero-inflation in suicidal thoughts. This highlights the inherent difficulty of idiographic prediction for suicide risk, given that the dynamics of suicidal thinking for most individuals (even at high levels of clinical severity) is characterized by high zero-inflation. However, these analyses also found that models tended to provide better prediction for individuals whose EMA data were characterized by greater marginal entropy and mean (that is, individuals with more equal distribution of responses across the Likert scale, and higher average severity). These results are promising, as predicting suicidal thoughts is probably most important for individuals with higher severity and fluctuations in suicidal thinking (who may be at greatest risk for short-term suicidal behavior).

Findings from this study should be considered alongside several limitations. First, this sample was composed primarily of white cisgender individuals, and inclusion criteria included English fluency. Future research focused on predicting suicidal thoughts for individuals from marginalized backgrounds (for example, individuals from racial, ethnic and linguistic minority groups) is crucial, and can also shed light on additional features (for example, daily life experiences with racism and discrimination) that may be salient predictors of suicide risk in diverse populations^{34,35}. Second, despite observed differences in model performance between the baseline and GP models, predictive accuracy was relatively low overall, with all average R^2 values being <0.20 (compared

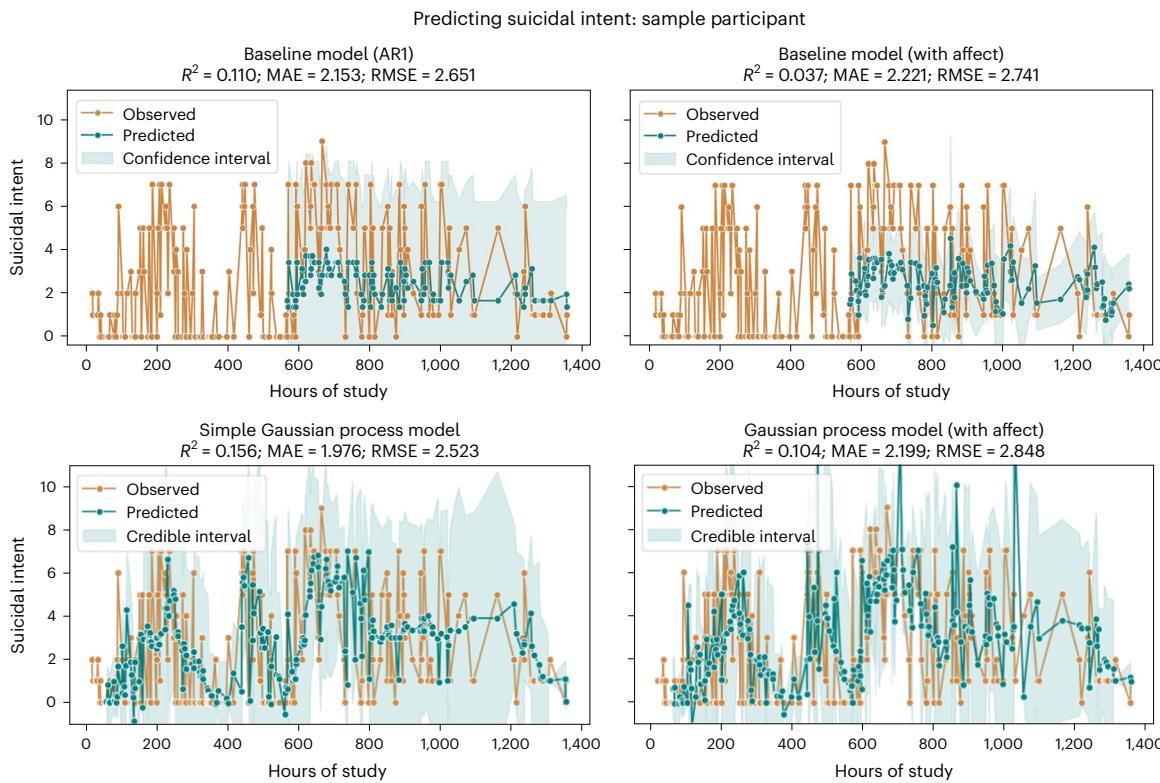


Fig. 4 | Illustrative example of idiographic prediction of suicidal intent for one participant. Illustrative example, modeled using (1) baseline (single train/test split) simple autoregressive models, (2) baseline elastic net models with

additional lagged affect features, (3) simple GP models (using iterative rolling-forward prediction) with history of suicidal urges only and (4) GP models with additional lagged affect features.

to, for example, average R^2 values of 0.27 for idiographic prediction of alcohol craving¹⁹). This was not surprising, as suicidal thoughts are extremely heterogeneous and transient phenomena and characterized by dynamic fluctuations, which makes accurate prediction quite challenging^{5,6,31}. Indeed, in many cases, our models were not able to anticipate extreme changes in suicidal thoughts, but only captured these changes after they had occurred. Future work incorporating additional data streams (for example, from wearable biosensors, passive smartphone monitoring) and self-reported EMA data (including with higher-density sampling methods) may assist in producing stronger models³⁶. Third, most participants (81%) who enrolled in the larger intensive longitudinal study did not meet inclusion criteria for our present analyses due to high amounts of missing data and low compliance across the full sample. While concerns about generalizability are tempered by the idiographic nature of our Article, and there were minimal differences in clinical and demographic characteristics between included versus excluded participants, it is still possible that these participants differed meaningfully in other dimensions that we did not assess in the present study (for example, conscientiousness). Finally, although compliance was relatively high for all participants in the present sample, it is possible that the true underlying dynamics of suicidal thinking were not fully captured in our models due to missing data. Future studies would benefit from methods to reduce participant burden and facilitate higher compliance rates overall.

Despite these limitations, this study provides encouraging results about the potential of idiographic prediction of suicidal thoughts, and extends recent work building personalized explanatory models of suicide¹⁸ to suicide prediction. Our findings suggest that iteratively training and updating models as data are collected may potentially be a promising method for optimizing predictive accuracy, although idiographic prediction of suicidal thoughts remains quite challenging, and our models had only modest performance overall. Future work combining idiographic

and nomothetic approaches (for example, stratifying patients based on between-person risk factors and building prediction models within each group) may further improve predictive accuracy. In addition, continued research incorporating higher-density sampling of suicidal thoughts and putative risk factors—alongside integration of actively and passively collected data streams—can provide critical information to guide the development of just-in-time adaptive interventions²², to detect risk and intervene when support may be most needed.

Methods

Participants and procedure

Participants were drawn from an ongoing intensive longitudinal study of suicidal thoughts and behaviors among individuals presenting for psychiatric hospital treatment. All procedures were approved by the Harvard University Institutional Review Board (IRB) and governing hospital IRBs. Adult participants were recruited from the psychiatric emergency department at Massachusetts General Hospital and provided written informed consent. Adolescent participants were recruited from the inpatient psychiatric unit at Franciscan Children's and provided written assent, and their parents/guardians provided written informed consent. All participants completed a battery of self-report questionnaires upon enrollment into the study, followed by a three-month EMA protocol including six prompts per day (including morning and evening prompts at scheduled times, and four semi-random daytime prompts during predefined intervals). Participant compensation included US\$10 for the initial survey and US\$1 per EMA survey. Inclusion criteria in the larger intensive longitudinal study included English fluency, presentation to the hospital with suicidal thoughts, not being acutely agitated or exhibiting violent behavior, cognitive capacity to provide informed consent, and possession of a smartphone. Participants in this study were recruited between May 2019 and October 2022.

Our goal in this specific investigation was to use participants' history of suicidal thoughts and affect to predict their future suicidal thinking, such that the relevant sample size became the number of EMA responses provided by each participant. Therefore, our inclusion criteria for the current investigation also included (1) providing at least 100 EMA responses and (2) completing at least one EMA survey every 72 h, given our goal of idiographic short-term prediction. Of 477 participants in the larger study, a total of 89 participants (45 adults, 44 adolescents) met these inclusion criteria and were included in our current analyses for suicidal intent, and a total of 88 participants (44 adults, 44 adolescents) met these inclusion criteria and were included in our current analyses for suicidal urges.

There were minimal differences in baseline characteristics between included versus excluded participants. A series of *t*-tests indicated a small difference in the number of past-week suicide plans ($P = 0.04$), with slightly fewer past-week plans among included ($M = 2.04$) compared to excluded ($M = 2.74$) participants, but no differences in past-year or past-month suicide plans, or in past-year, past-month or past-week suicide thoughts or attempts (all $P > 0.05$). Similarly, a series of χ^2 tests found no significant differences in demographic characteristics (including race, ethnicity and gender) between included and excluded participants in this study (all $P > 0.05$).

Measures

Demographics. Participants self-reported demographic data, including age, race, ethnicity and gender at study enrollment.

Baseline assessment of self-injurious thoughts and behaviors.

Participants completed the self-report version of the SITBI³⁷ at the time of study enrollment, assessing their history (presence and frequency) of suicidal thoughts, plans and attempts.

Ecological momentary assessment of suicidal thoughts. Following previous work from our laboratory^{6,38,39}, we used two items to assess suicidal thinking. We first assessed suicidal urges with a single item ('Right now, how strong is your urge to kill yourself?') rated on a scale from 0 (not at all) to 10 (very strong). We assessed suicide intent with a second single item ('Right now, how strong is your intention to kill yourself today?') rated on a scale from 0 (I am definitely not going to kill myself today) to 10 (I am definitely going to kill myself today). Given the different question phrasing and anchoring between these two items, we modeled suicidal urges and intent separately in all analyses, building separate idiographic models to predict suicidal urges versus suicidal intent (rather than collapsing the two items into a sum score of 'suicidal thought intensity').

Ecological momentary assessment of affect. We used 15 items to measure momentary current affect (or cognitive-affective) states, similar to previous work from our laboratory^{39,40}. In the present study we measured the following affective/cognitive-affective states with single-item questions, each rated on a scale from 0 (not at all) to 10 (very much): agitated, angry, burdensome, desire to escape, energetic, fatigued, hopeless, humiliated, isolated, negative, numb, positive, self-hate, trapped and worried.

Data analyses

All data preparation and analyses were performed in R (version 4.2.0) and Python (version 3.10.6), with the tidyverse and tidymodels meta-packages in R, and the numpy, pandas, scikit-learn and matplotlib packages in Python^{41–46}. All code is publicly available at https://github.com/ShirleyBWang/idiographic_prediction (ref. 25).

Data preparation. We prepared each participant's EMA data for analysis separately, as the idiographic approach treats each person's time-series data as its own dataset for modeling. First, we created a

lagged dataset for each participant, with columns representing suicidal thoughts and affect measured at time t , and suicidal thoughts measured at time $t + 1$. For the baseline (single train/test split) models, we then split each participant's data into a training dataset (first 50% of data) and testing dataset (last 50% of data).

For the iterative rolling-forward models, we created an initial training dataset of the first ten rows for each participant, and held out the next (11th) row as the test observation. After making a prediction for the held-out observation, we created another training dataset of the subsequent ten rows (that is, rows 1–11), and held out the next (12th) row as the test observation (so that each prediction of next time-point suicidal thought severity was made using data from the prior ten observations). We iterated through each participant's dataset in this manner, such that each participant had a total of $N - 9$ training datasets and $N - 10$ held-out test observations, where N represents the total number of rows (that is, EMA observations) for each participant.

Modeling approach. Our baseline models followed existing approaches in single-subject machine learning with EMA data, using a single train-test split of each participant's data^{19,20}. We first fit a simple autoregressive (linear regression) model, using participants' suicidal thinking at time t to predict their suicidal thoughts at the subsequent time point, time $t + 1$. We next fit an elastic net model with suicidal thoughts and additional affect/cognitive-affective features (all 15, as described above) at time t to predict suicidal thoughts at time $t + 1$. The elastic net algorithm is a relatively simple machine learning method that combines the λ_1 (lasso) and λ_2 (ridge) penalties, to shrink parameter estimates towards zero:

$$\text{SSE}_{\text{EN}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^P |\beta_j| + \lambda_2 \sum_{j=1}^P \beta_j^2 \quad (1)$$

where SSE is the sum of squared errors, y_i is the value of the outcome variable y for the i th observation and \hat{y}_i is the predicted value of the outcome variable y for the i th observation. P is the number of predictors (features) and β is the regression parameter weight.

Our preprocessing steps included k -nearest neighbors imputation for missing features, removing features with near-zero variance, removing redundant features (highly correlated at $r \geq 0.9$ or linear combinations of other features), and normalizing all features. Following recommendations by Kuhn and Johnson⁴⁷, we used fivefold cross-validation with three repetitions within each training dataset (that is, the first 50% of each participant's EMA data) to select the optimal lambda (shrinkage) and alpha (mixing) parameters. We then used the best cross-validated model to predict suicidal thoughts in each participant's testing data (that is, the last 50% of each participant's EMA data). Because we build idiographic models predicting suicidal urges and suicidal intent separately, this resulted in a total of four baseline models being fit for each participant (two simple autoregressive models and two elastic net models). However, we note that elastic net models could not be fit for a subset of participants ($N = 19$ for suicidal urges and $N = 29$ for suicidal intent) due to constant values of the outcome variable and/or preprocessing steps removing all but one feature. All baseline models were fit and evaluated using the tidymodels meta-package in R.

For the iterative rolling-forward models, we fit GP regression models using participants' history of suicidal thoughts and affect to predict their suicidal thoughts at the next time point. A complete introduction to GP models is beyond the scope of this Article, and we refer interested readers to excellent textbooks and tutorial papers (including their application to time-series data)^{26–28}. However, in brief, GP models are powerful Bayesian machine learning models that are able to infer a distribution over functions $f(x)$ that map inputs x to outputs y . In contrast to parametric Bayesian models, which place a prior over the coefficients/weights of the predictor $f(x)$ (for example, a prior over

the coefficients of a linear model), GPs are non-parametric, and assume a prior distribution directly over the functions themselves:

$$f(x) \sim \text{GP}(m(x), k(x, x')) \quad (2)$$

The above equation denotes a GP prior with mean $m(x)$ and a covariance matrix induced by a kernel function $k(x, x')$, which implies the distribution over functions. For the current models, as is typically done, we assumed a constant mean prior (specified as the mean of the training data). We specified a kernel composed as the sum of three commonly used kernels, including (1) a radial basis function kernel (the most commonly used kernel, representing the intuition that observations close in time are more correlated than observations far away in time), (2) a periodic kernel (representing periodic trends with respect to time, for example, suicidal thinking that may fluctuate higher in the evenings and lower in the mornings) and (3) a white-noise kernel (to model any other stochastic fluctuations in suicidal thinking that are independent of time). The hyperparameters for each kernel were optimized for each participant, within each training dataset, by maximizing the log marginal likelihood, using the ‘L-BFGS-B’ optimizer with five restarts, using the scikit-learn package in Python.

Given this prior over functions, we then assume that our outputs relate to our inputs under the following likelihood:

$$y = f(x) + \epsilon \quad (3)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ represents observation noise. Using this prior and likelihood, we can then implement Bayes’ rule to infer a posterior distribution over functions given the observed data (that is, inputs and outputs), $P(f|data)$. This distribution represents functions $f(x)$ that both capture the data well (that is, make accurate predictions) and are likely under the prior. Using this posterior, we can compute the ultimate object of interest—the posterior predictive distribution—which is used to make predictions for new data points. For brevity, we have omitted the analytic forms of the posterior and posterior predictive distributions, though they can be found in standard machine learning textbooks.

Similar to the baseline models, we fit two sets of GP models for each participant: (1) only using participants’ history of suicidal thoughts and (2) with additional lagged affect features. Because we aimed to predict suicidal urges and suicidal intent separately, this also resulted in a total of four GP models being fit for each participant.

Model performance. Our primary model performance metric was R^2 (calculated as the square of the correlation coefficient) between the observed versus predicted values of suicidal thoughts in each participant’s testing dataset. We statistically compared R^2 values with multilevel regression models (adjusted with Tukey’s honestly significant difference (HSD) to evaluate pairwise significant differences). We also used r.m.s.e. and MAE as secondary model performance metrics. Higher scores for R^2 indicate better model performance, and lower scores for r.m.s.e. and MAE indicate better performance. Of note, we chose R^2 as our primary metric to evaluate our models more rigorously. As shown in the results in the next section, relying on r.m.s.e. and MAE for model evaluation would paint an overly optimistic picture (for example, that our models could predict next time-point suicidal thought severity with an average error of <1 point on our 10-point Likert scales). We believed that R^2 would present a realistic view of our models’ abilities to predict the dynamic variability in suicidal thinking, without being overly optimistic.

Secondary analysis. As a secondary analysis, to explore differences in model accuracy across participants, we constructed features of each participant’s EMA data and correlated these features with our evaluation metrics (R^2 , MAE, r.m.s.e.). We constructed ten features for each

participant’s suicidal urges and suicidal intent (that is, 20 features total per participant), including average time between EMA observations, variance of time between EMA observations, number of EMA observations, percent big jumps (in EMA responses from one time point to the next), percent small jumps, percent zero, percent high, marginal mean, marginal entropy and the ratio of percent high to marginal entropy (that is, with higher values representing patients with more chronic suicidal thoughts). More details on feature construction are provided in the Supplementary Methods.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

This study includes data from a larger existing project. Access to anonymized data for the larger project (of which this study is a component) will be available through the National Institute of Mental Health Data Archive upon its completion.

Code availability

All code is publicly available at github.com/ShirleyBWang/idiographic_prediction (ref. 25).

References

1. CDC. Facts About Suicide <https://www.cdc.gov/suicide/facts/index.html> (CDC, 2021).
2. Nock, M. K. et al. Suicide and suicidal behavior. *Epidemiol. Rev.* **30**, 133–154 (2008).
3. WHO. *National Suicide Prevention Strategies: Progress, Examples and Indicators* (World Health Organization, 2018).
4. Franklin, J. C. et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol. Bull.* **143**, 187–232 (2017).
5. Kleiman, E. M. et al. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. *J. Abnorm. Psychol.* **126**, 726–738 (2017).
6. Wang, S. B. et al. A pilot study using frequent inpatient assessments of suicidal thinking to predict short-term postdischarge suicidal behavior. *JAMA Netw. Open* **4**, e210591 (2021).
7. Bryan, C. J., Rozek, D. C., Butner, J. & Rudd, M. D. Patterns of change in suicide ideation signal the recurrence of suicide attempts among high-risk psychiatric outpatients. *Behav. Res. Ther.* **120**, 103392 (2019).
8. Wright, A. G. C. & Woods, W. C. Personalized models of psychopathology. *Annu. Rev. Clin. Psychol.* **16**, 49–74 (2020).
9. Piccirillo, M. L. & Rodebaugh, T. L. Foundations of idiographic methods in psychology and applications for psychotherapy. *Clin. Psychol. Rev.* **71**, 90–100 (2019).
10. Molenaar, P. C. A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Measurement* **2**, 201–218 (2004).
11. Molenaar, P. C. M. & Campbell, C. G. The new person-specific paradigm in psychology. *Curr. Dir. Psychol. Sci.* **18**, 112–117 (2009).
12. Fisher, A. J., Medaglia, J. D. & Jeronimus, B. F. Lack of group-to-individual generalizability is a threat to human subjects research. *Proc. Natl Acad. Sci. USA* **115**, E6106–E6115 (2018).
13. Berman, A. L., King, R. A. & Aptek, A. in *Suicide in Children and Adolescents* (eds King, R. A. & Aptek, A.) 198–210 (Cambridge Univ. Press, 2003).
14. Leenars, A. A. In defense of the idiographic approach: studies of suicide notes and personal documents. *Arch. Suicide Res.* **6**, 19–30 (2002).

15. Barlow, D. H. & Nock, M. K. Why can't we be more idiographic in our research? *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **4**, 19–21 (2009).
16. Ozomaro, U., Wahlestedt, C. & Nemeroff, C. B. Personalized medicine in psychiatry: problems and promises. *BMC Med.* **11**, 132 (2013).
17. Simon, G. E. & Perlis, R. H. Personalized medicine for depression: can we match patients with treatments? *Am. J. Psychiatry* **167**, 1445–1455 (2010).
18. Kaurin, A., Dombrovski, A. Y., Hallquist, M. N. & Wright, A. G. Integrating a functional view on suicide risk into idiographic statistical models. *Behav. Res. Ther.* **150**, 104012 (2022).
19. Soyster, P. D., Ashlock, L. & Fisher, A. J. Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: a demonstration of parallel utility. *Psychol. Addict. Behav.* **36**, 296–306 (2021).
20. Fisher, A. J. & Soyster, P. D. Generating accurate personalized predictions of future behavior: a smoking exemplar. Preprint at <https://doi.org/10.31234/osf.io/e24v6> (2019).
21. Beck, E. D. & Jackson, J. J. Personalized prediction of behaviors and experiences: an idiographic person-situation test. *Psychol. Sci.* **33**, 1767–1782 (2022).
22. Nahum-Shani, I. et al. Just-in-Time Adaptive Interventions (JITAl)s in mobile health: key components and design principles for ongoing health behavior support. *Ann. Behav. Med.* **52**, 446–462 (2018).
23. Howe, E. S. & Fisher, A. J. Identifying and predicting posttraumatic stress symptom states in adults with posttraumatic stress disorder. *J. Trauma. Stress* **35**, 1508–1520 (2022).
24. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, 2013).
25. Wang, S. B. et al. Idiographic prediction of suicidal thoughts (GitHub); https://github.com/ShirleyBWang/idiographic_prediction
26. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
27. Roberts, S. et al. Gaussian processes for time-series modelling. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **371**, 20110550 (2013).
28. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on Gaussian process regression: modelling, exploring and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
29. Torous, J. & Hsin, H. Empowering the digital therapeutic relationship: virtual clinics for digital health interventions. *Npj Digit. Med.* **1**, 16 (2018).
30. Rodriguez-Villa, E. et al. The digital clinic: implementing technology and augmenting care for mental health. *Gen. Hosp. Psychiatry* **66**, 59–66 (2020).
31. Coppersmith, D. D. L. et al. Mapping the timescale of suicidal thinking. *Proc. Natl Acad. Sci. USA* **120**, e2215434120 (2023).
32. Intille, S., Haynes, C., Maniar, D., Ponnada, A. & Manjourides, J. μEMA: microinteraction-based Ecological Momentary Assessment (EMA) using a smartwatch. *Proc. ACM Int. Conf. Ubiquitous Comput.* **2016**, 1124–1128 (2016).
33. Ponnada, A., Haynes, C., Maniar, D., Manjourides, J. & Intille, S. Microinteraction ecological momentary assessment response rates: effect of microinteractions or the smartwatch? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **1**, 92 (2017).
34. Adams, L. et al. Assessing the real-time influence of racism-related stress and suicidality among black men: protocol for an ecological momentary assessment study. *JMIR Res. Protoc.* **10**, e31241 (2021).
35. Alvarez, K., Polanco-Roman, L., Samuel Breslow, A. & Molock, S. Structural racism and suicide prevention for ethnoracially minoritized youth: a conceptual framework and illustration across systems. *Am. J. Psychiatry* **179**, 422–433 (2022).
36. Kleiman, E. M. et al. Can passive measurement of physiological distress help better predict suicidal thinking?. *Transl. Psychiatry* **11**, 611 (2021).
37. Nock, M. K., Holmberg, E. B., Photos, V. I. & Michel, B. D. Self-injurious thoughts and behaviors interview: development, reliability and validity in an adolescent sample. *Psychol. Assess.* **19**, 309–317 (2007).
38. Fortgang, R. G. et al. Increase in suicidal thinking during COVID-19. *Clin. Psychol. Sci.* **9**, 482–488 (2021).
39. Bentley, K. H. et al. Do patterns and types of negative affect during hospitalization predict short-term post-discharge suicidal thoughts and behaviors? *Affect. Sci.* **2**, 484–494 (2021).
40. Mou, D. et al. Negative affect is more strongly associated with suicidal thinking among suicidal patients with borderline personality disorder than those without. *J. Psychiatr. Res.* **104**, 198–201 (2018).
41. R Core Team. R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2019); <https://www.R-project.org/>
42. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
43. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
44. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
46. Kuhn, M. & Wickham, H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles (Tidymodels, 2020); <https://www.tidymodels.org>
47. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer, 2013).

Acknowledgements

This research was supported by funding from the National Institute of Mental Health (F31MH125495 to S.B.W., U01MH116928 to M.K.N., K23MH120436 to K.H.B., K23MH132766 to R.G.F., K23MH120439 to K.L.Z. and R01MH117599 to J.W.S.). This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

All of the authors made a substantial contribution to this study. S.B.W. was responsible for data analysis and writing the paper. M.K.N. was responsible for study conception, design, funding acquisition and supervision of all activities. R.D.I.V.G., YY. and W.P. contributed to data analysis. A.H. contributed to study design. K.H.B., S.A.B., R.J.B., R.G.F., E.M.K., A.J.M., J.W.S. and K.L.Z. contributed to study design and supervision. D.D. and J.P.O. contributed to software and data management. A.C., M.D., L.F., F.K.-B., O.O.-O., N.R., J.R.R. and T.T. contributed to data collection. All authors contributed to reviewing and revising the manuscript, and all authors approved the final version of the paper for submission.

Competing interests

M.K.N. receives publication royalties from Macmillan, Pearson and UpToDate. He has been a paid consultant in the past three years for Microsoft Corporation, the Veterans Health Administration and COMPASS Pathways, and for legal cases regarding a death by suicide. He has stock options in Cerebral Inc. He is an unpaid scientific advisor for Empatica, Koko and TalkLife. E.M.K. has been a paid consultant in the past three years for Boehringer Ingelheim Pharmaceuticals. J.W.S. is a member of the Scientific Advisory Board of Sensorium Therapeutics (with equity), and has received grant support from Biogen, Inc. He is PI of a collaborative study of the genetics of depression and bipolar disorder sponsored by 23andMe for which

23andMe provides analysis time as in-kind support but no payments. J.P.O. has been a paid consultant in the past three years for Boehringer Ingelheim and has received research funding from them. D.D. is the founder and CEO of Apoth.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44220-024-00335-w>.

Correspondence and requests for materials should be addressed to Shirley B. Wang.

Peer review information *Nature Mental Health* thanks Zhongzhi Xu and the other, anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2024

¹Department of Psychology, Yale University, New Haven, CT, USA. ²Institute for Experiential Artificial Intelligence, Northeastern University, Boston, MA, USA. ³Department of Computer Science, Wellesley College, Wellesley, MA, USA. ⁴John A. Paulson School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA. ⁵Department of Psychology, Harvard University, Cambridge, MA, USA. ⁶Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA. ⁷Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA. ⁸Mental Health Research, Franciscan Children's, Boston, MA, USA. ⁹National Institute of Mental Health, Rockville, MD, USA. ¹⁰Department of Psychology, University of Georgia, Athens, GA, USA. ¹¹Department of Psychology, University of Tennessee, Knoxville, TN, USA. ¹²Department of Psychology, Rutgers University, New Brunswick, NJ, USA. ¹³Department of Psychology, The Pennsylvania State University, University Park, PA, USA. ¹⁴Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.  e-mail: shirley.wang@yale.edu

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Software used for data collection includes LifeData and Qualtrics.

Data analysis All data preparation and analyses were performed in R (version 4.2.0) and Python (version 3.10.6), with tidyverse and tidymodels meta-packages in R, and numpy, pandas, scikit-learn, and matplotlib packages in Python.^{28–33} All code is publicly available at github.com/ShirleyBWang/idiographic_prediction.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

This study includes data from a larger existing project. Access to anonymized data for the larger project (of which this study is a component) will be available through the National Institute of Mental Health Data Archive upon its completion.

Human research participants

Policy information about [studies involving human research participants](#) and [Sex and Gender in Research](#).

Reporting on sex and gender

[Reported in Table 1](#)

Population characteristics

[Reported in Table 1](#)

Recruitment

Adult participants were recruited from the psychiatry emergency department at Massachusetts General Hospital and provided written informed consent. Adolescent participants were recruited from the inpatient psychiatric unit at Franciscan Children's and provided written assent, and their parents/guardians provided written informed consent. Participant compensation included \$10 for the initial survey and \$1 per EMA survey. Recruitment procedures, inclusion, and exclusion criteria are presented in the Methods, with potential limitations discussed in the Discussion.

Ethics oversight

All procedures were approved by the Harvard University Institutional Review Board (IRB) and governing hospital IRBs.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Intensive longitudinal study using ecological momentary assessment (EMA). Data are quantitative. More details provided in Methods.

Research sample

The sample included individuals presenting for psychiatric hospital treatment. Average age was 22.10 (SD = 10.02) and the sample was predominantly White (80%) and cisgender female (59%). More details are provided in the Methods and Table 1, with limitations discussed in the Discussion.

Sampling strategy

Individuals who presented to psychiatric hospital settings with suicidal thoughts and/or recent suicidal behavior.

Data collection

Ecological momentary assessment (EMA) with LifeData.

Timing

May 2019 - October 2022

Data exclusions

In the larger longitudinal study: not being acutely agitated or exhibiting violent behavior, not having cognitive capacity to provide informed consent, not owning a smartphone. For the present analyses: not providing at least 100 EMA responses with at least one response every 72 hours (given our goal of short-term idiographic prediction). More details in the Methods.

Non-participation

No individuals in the current study were non-participating

Randomization

No randomization.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	Antibodies
<input checked="" type="checkbox"/>	Eukaryotic cell lines
<input checked="" type="checkbox"/>	Palaeontology and archaeology
<input checked="" type="checkbox"/>	Animals and other organisms
<input checked="" type="checkbox"/>	Clinical data
<input checked="" type="checkbox"/>	Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	ChIP-seq
<input checked="" type="checkbox"/>	Flow cytometry
<input checked="" type="checkbox"/>	MRI-based neuroimaging