Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Ren B, Balkind EG, Pastro B, Israel ES, Pizzagalli DA, Rahimi-Eichi H, Baker JT, Webb CA (2023). Predicting states of elevated negative affect in adolescents from smartphone sensors: a novel personalized machine learning approach. *Psychological Medicine* **53**, 5146–5154. https://doi.org/10.1017/S0033291722002161

Received: 18 March 2022
Revised: 25 May 2022
Accepted: 23 June 2022
First published online: 27 July 2022

Key words:
Adolescents; ecological momentary assessment; negative affect; prediction; smartphone sensors

Author for correspondence:
Christian A. Webb,
E-mail: cwebb@mclean.harvard.edu

Predicting states of elevated negative affect in adolescents from smartphone sensors: a novel personalized machine learning approach

Boyu Ren[1,2], Emma G. Balkind[1,2], Brianna Pastro[1,2], Elana S. Israel[1,2], Diego A. Pizzagalli[1,2], Habiballah Rahimi-Eichi[1,2], Justin T. Baker[1,2] and Christian A. Webb[1,2]

[1]Department of Psychiatry, Harvard Medical School, Boston, MA, USA and [2]McLean Hospital, Belmont, MA, USA

Abstract

Background. Adolescence is characterized by profound change, including increases in negative emotions. Approximately 84% of American adolescents own a smartphone, which can continuously and unobtrusively track variables potentially predictive of heightened negative emotions (e.g. activity levels, location, pattern of phone usage). The extent to which built-in smartphone sensors can reliably predict states of elevated negative affect in adolescents is an open question.

Methods. Adolescent participants ($n = 22$; ages 13–18) with low to high levels of depressive symptoms were followed for 15 weeks using a combination of ecological momentary assessments (EMAs) and continuously collected passive smartphone sensor data. EMAs probed negative emotional states (i.e. anger, sadness and anxiety) 2–3 times per day every other week throughout the study (total: 1145 EMA measurements). Smartphone accelerometer, location and device state data were collected to derive 14 discrete estimates of behavior, including activity level, percentage of time spent at home, sleep onset and duration, and phone usage.

Results. A personalized ensemble machine learning model derived from smartphone sensor data outperformed other statistical approaches (e.g. linear mixed model) and predicted states of elevated anger and anxiety with acceptable discrimination ability (area under the curve (AUC) = 74% and 71%, respectively), but demonstrated more modest discrimination ability for predicting states of high sadness (AUC = 66%).

Conclusions. To the extent that smartphone data could provide reasonably accurate real-time predictions of states of high negative affect in teens, brief 'just-in-time' interventions could be immediately deployed via smartphone notifications or mental health apps to alleviate these states.

Adolescence is a developmental period characterized by profound change across multiple domains, including emotional experience. Studies reveal that the frequency of negative emotional states increases during adolescence (Bailen, Green, & Thompson, 2019; Frost, Hoyt, Chung, & Adam, 2015; Griffith, Clark, Haraden, Young, & Hankin, 2021; Larson, Moneta, Richards, & Wilson, 2002). Relatedly, studies of personality trait development have also shown that neuroticism (i.e. negative emotionality) increases during adolescence, in particular for girls (Borghuis et al., 2017; Soto, 2016; Soto, John, Gosling, & Potter, 2011). Given longitudinal research indicating that neuroticism is a robust risk factor for the development of depression and anxiety disorders (Vinograd et al., 2020; Zinbarg et al., 2016), increases in this trait during adolescence may help account for the developmental surge in emotional disorders during this same time period (in particular among girls) (Avenevoli, Swendsen, He, Burstein, & Merikangas, 2015). Accordingly, there is an acute need to develop data-driven approaches to reliably predict and ultimately interrupt states of markedly elevated negative emotions as they occur in the daily lives of teens. In addition to the immediate benefits of alleviating acute states of affective distress, reducing the frequency and duration of episodes of high negative affect (HNA) may serve to reduce the risk of future onset of emotional disorders.

Relevant in this context, approximately 84% of American adolescents own a smartphone (Rideout & Robb, 2019), which has the ability to continuously and unobtrusively track variables predictive of shifts in affect, including activity levels, location, phone/screen use and proxies of social interaction, while doing so in real-time in the daily lives of teens. A key benefit of data acquired from these 'passive' built-in smartphone sensors is that no user input is required. When adolescents are immersed in a negative emotional state, they may be less likely to report on their experiences via conventional (and more burdensome) self-report measures.

The extent to which built-in smartphone sensors can reliably predict states of elevated negative affect (e.g. high sadness, anxiety, or anger) in adolescents is an open question. As adolescents transition into a negative emotional state, they may display distinct behavioral and

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

interpersonal changes that can be detected via smartphone sensors, including decreased activity levels (measured via actigraphy), increased time spent at home (GPS and actigraphy), sleep disturbance (phone use patterns and actigraphy) and interpersonal avoidance (decreased initiation and response to calls and texts). Prior research indicates that smartphone sensors can predict states of elevated negative affect, and depressive and anxiety symptoms (e.g. Cao et al., 2020; Jacobson & Chung, 2020; Sequeira et al., 2020; Shah et al., 2021). However, the bulk of this research has been conducted in adults rather than adolescents.

The present study addressed several gaps in the existing literature. First, we recruited a sample of adolescents enriched for heightened experiences of negative emotional states (i.e. adolescents with elevated depressive symptoms). Second, there is likely substantial heterogeneity between adolescents in their behavioral patterns that signal the onset of negative emotional states. For example, teens experiencing heightened anxiety may exhibit decreased *v.* increased activity levels, or increased *v.* decreased interaction with one's smartphone. Thus, the likely idiosyncratic behavioral patterns associated with states of HNA highlight the need for a *person-specific* (idiographic) modeling approach. At the same time, certain predictive relationships may be shared across individuals or subgroups of adolescents. Accordingly, we used a novel statistical approach (Ren, Patil, Dominici, Parmigiani, & Trippa, 2021) which integrates person-specific models (i.e. developing a unique model for each individual) with a data-driven search for the optimal weighting across all subject-specific models (i.e. 'borrowing' information from prediction models for other individuals in the sample). This statistical approach allowed us to estimate the combination of subject-specific models that optimizes predictive performance and to answer the following question: Does a predictive model that is exclusively trained on one's own data outperform a weighted model which incorporates information from other individuals' models? Third, recent advances in machine learning (ML) provide statistical approaches well-suited to model complex smartphone-acquired passive data and their association with negative affect. For example, several decision-tree-based algorithms (e.g. random forest) and multiple kernel learning algorithms (e.g. support vector machines) can (1) handle large numbers of predictors (whether continuous or categorical), (2) have robust performance in the presence of multicollinearity, (3) handle non-linear associations and complex interactions and, critically, can (4) be tuned to prevent overfitting (i.e. tend to generalize well to new samples; Boehmke & Greenwell, 2019; Squeglia et al., 2016). In the present study, an ML ensembling method was implemented, which assigns weights to a set of preselected ML approaches to develop a consolidated predictive algorithm which optimizes cross-validated predictive performance. A key benefit of this approach is that it can achieve improved prediction performance by combining a range of ML approaches which rely on very different algorithms to generate predictions (e.g. variants of conventional regression such as elastic net *v.* tree-ensemble approaches such as a random forest). Given the substantial differences across algorithms (e.g. in the selection and weighting of variables, in testing nonlinear relations and higher-order interactions), it is likely that they would capture different aspects of the relationship between the outcome of interest and predictors. Thus, a combination of these diverse models tends to be more flexible than any of the individual models and has the potential to attain better prediction performance. Here, we tested whether the consolidated approach weighting these different algorithms outperforms each of the individual ML approaches, as well as conventional statistical approaches (i.e. linear mixed model). We hypothesized that an ensemble ML approach would outperform common modeling approaches in predicting states of elevated negative affect. In addition, we expected that the former ensemble ML approach (i.e. searching for the optimal combination of subject-specific models) would outperform a fully idiographic ML model (i.e. person-specific models that do not borrow any information from other individuals' models). Finally, we expected our final model would predict states of elevated negative emotions with acceptable accuracy (area under the curve (AUC) ⩾ 0.70).

Method

Participants

Adolescents ages 13–18 were recruited from the greater Boston area. Given that the sample in this study was derived from a larger ongoing trial focused on adolescents with anhedonia, participants in this study had high (Anhedonic, AH; $n = 13$) or low (Typically Developing, TD; $n = 9$) levels of anhedonia. The following exclusion criteria were applicable to both groups: history of head trauma with loss of consciousness > 2 min, history of seizure disorder, serious or unstable medical illness, or current use of cocaine, stimulant, or dopaminergic drugs. The AH group were required to have elevated anhedonia according to the Snaith Hamilton Pleasure Scale (SHAPS) (total score ⩾ 3, as assessed by the original binary [0–1] scoring (Snaith et al., 1995)) and the Schedule for Affective Disorders and Schizophrenia for School-Age Children Present and Lifetime Version (K-SADS-PL) (Kaufman et al., 2013) clinical interview (anhedonia item score > 1). TD participants were required to have a SHAPS score = 0 (i.e. no anhedonia). See online Supplement for additional information on exclusion criteria.

Procedure

All procedures were approved by the Mass General Brigham (MGB) IRB. All participating parents provided written informed consent and all children provided assent. Participants completed an initial assessment session with self-report measures, including assessments of anhedonia (SHAPS) (Snaith et al., 1995) and depressive symptom severity (Center for Epidemiologic Studies Depression Scale; CES-D) (Radloff, 1977). The K-SADS-PL for the DSM-5 was subsequently administered (see online Supplement for additional details). Following the initial assessment, participants completed an average of 15 weeks (mean = 106.2 days; S.D. = 16.1) of ecological momentary assessment (EMA) surveys. Surveys were triggered 2–3 times a day via the Metricwire smartphone application. Throughout this study period, passive data were collected from participants' smartphones using the Beiwe application (see *Passive Smartphone Measures* below) (Rahimi-Eichi et al., 2021; Torous, Kiang, Lorme, & Onnela, 2016).

Measures

Ecological Momentary Assessment measures

EMA surveys assessing current affect were triggered on participant's smartphones 2–3 times per day, every other week, resulting in 1145 measurements (mean of 52 EMA surveys per participant). Similar to previous EMA studies in adolescents (Forbes et al.,

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

2009, 2012; Webb et al., 2021), items were derived from the Positive and Negative Affect Schedule for Children (Laurent et al., 1999) and surveys were triggered Thursday through Monday to sample affect on weekdays and weekends (see online Supplement for additional details on the EMA protocol). At each EMA survey, participants were asked to rate how much they were experiencing a given emotion immediately prior to receiving the EMA survey notification. Items were rated on a 5-point Likert scale from 1 (Very slightly/not at all) to 5 (Extremely). The present study was focused on the prediction of negative affect items: 'Sad', 'Nervous', and 'Angry'.

Passive smartphone measures

The following fourteen variables were included as predictors of emotional states: (1) Phone Accelerometer Score: Hourly activity score of the phone accelerometer data, calculated from the standard deviation of the triaxial accelerometer data, root mean squared and scored as 1 when less than 70% (percentile), 2 when between 70% and 90%, and 3 when greater than 90% (Rahimi-Eichi et al., 2021). The scores are recorded for every minute and then averaged to calculate the hourly values; (2) Phone Use (MinsHr): The number of minutes the phone is used (unlocked) during each hour; (3) Phone Use (MinsDay): Number of minutes that the phone has been used (unlocked) during the day; (4) Phone Use (Hr): Number of hours during the day that the phone has been used (unlocked) at least for a minute during that hour; (5) Sleep Onset Time: Beginning of a sleep episode (hh:mm) (see online Supplement for details on the calculation of sleep episodes); (6) Wake Time: End of sleep episode (hh:mm); (7) Sleep Duration: Difference between Sleep Onset Time and Wake Time per day (min); (8) Daily Percent Home: Percentage of time the participant spends at home which could be the Home (the most visited POI of the study) if visited on that day or otherwise, the most visited POI of the day (therefore, when the individual is traveling for several days the home location is adjusted). POIs: Maximum 50 points of interest that were detected by spatial clustering of the temporally filtered GPS coordinates visited by the participant during the entire study; (9) Distance from Home: The farthest distance from Home visited by the participant during the day (km); (10) Daily Mobility Area: The radius of a circle that encompasses all visited locations during the day (km); (11) Places Visited Daily: Number of POIs other than Home that the participant has visited during the day; (12) Places Visited Hourly: The number of places that the participant has visited during each hour; (13) GPS Available: Number of hours the GPS signal is available (not missing) in 24 h; (14) Time of Day [Morning (5am-11:59am), Afternoon (noon-5:59pm) or Evening/Night 6pm-4:59am].

Data analytic plan

Missing data imputation

We imputed missing observations of predictors using multiple imputation by chained equations (MICE package in R) (van Buuren & Groothuis-Oudshoorn, 2011) on the merged dataset across subjects. We excluded outcome variables (i.e. anger, sadness, and nervousness) from the imputations to avoid overfitting of the final prediction models.

Definition of high negative affect (HNA) states

The goal of this study was to predict when adolescents were in states of relatively high sadness, anxiety, or nervousness. For simplicity, we henceforth refer to these states as HNA states. We define HNA states based on deviations from the *person-specific* average emotion score for a given adolescent. More specifically, for each emotion, if the observed score of a participant at time *t* exceeds their overall average by at least 1/2 point, we define this as an HNA state of that emotion (see Table 1 and Results). Note that there is a tradeoff between the cutoff value and the proportion of HNA states. High cutoff values lead to more confident identification of HNA states but as a result, the proportion of HNA states might be too low to train any generalizable classification algorithm. On the other hand, low cutoff values provide us with more HNA states for predictive modeling, but a number of these identified states might be questionable (i.e. too low to be considered states of 'high' negative affect). A sensitivity analysis with an alternative definition of HNA based on subject-specific quantiles yielded similar results (see online Supplement).

Table 1. Summary statistics of high negative affect (HNA) states for each negative emotion

	HNA proportion (%)	Average elevation above subject-specific mean
Anger	27.56	1.12
Sadness	32.18	1.06
Nervousness	32.28	1.25

Personalized predictions of HNA states

HNA states were predicted using passive smartphone data (see passive measures above) from the same day. We constructed the subject-specific prediction models using two approaches. The first was a generalized linear mixed-effects regression (GLMER), which captured the heterogeneity across subjects with random effects and predicted the person-specific outcomes (HNA states) by combining the estimated fixed effects with the random effects. The second model used a recently developed ensemble learning approach (Ren et al., 2021) that prioritizes predictive performance at the individual level while borrowing information across individuals to supplement idiosyncratic prediction models. Specifically, this approach utilizes an ensemble of idiosyncratic prediction models $f_i^l(x),\ \ i = 1,\ \ldots, K,\ \ l = 1,\ \ldots, L$, where K is the number of subjects and L is the number of different learning algorithms (e.g. logistic regression). $f_i^l(x)$ is trained on data from subject i with algorithm l. The personalized ensemble model (PEM) $f_i(x;w^i)$ for subject i is a linear combination of all idiosyncratic models (IM):

$$f_i(x;\, w^i) = \sum_{i'=1}^{K} \sum_{l=1}^{L} w_{i',l}^i f_{i'}^l(x)$$

and the combination weights $w^i = (w_{i',l}^i;\ i' = 1,\ \ldots K, l = 1,\ \ldots, L)$ with the constraints that $\sum_{i',l} w_{i',l}^i = 1$ and $w_{i',l}^i \geq 0$ for all $i,\ i' \in \{1, \ldots, K\}$ and $l \in \{1, \ldots, L\}$ are selected to minimize a cross-validated loss function:

$$\hat{\mathrm{w}}^i = argmin_w \sum_{j=1}^{n_i} \mathcal{L}\left(y_{i,j};\ \sum_{i' \neq i,l} w_{i',l}^i f_{i'}^l(x_{i,j}) + w_{i,l}^i f_{i,-j}^l(x_{i,j}) \right),$$

where n_i is the number of observations for subject i and $f_{i,-j}^l(x)$ is the IM trained on all data from subject i except for the j-th observation

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

(or a fold containing the j-th observation) with algorithm l. $\mathcal{L}$ is a loss function and here we use log-loss for binary outcomes. In summary, this ensemble approach develops a unique model for *each* individual through a data-driven search for the optimal weighting of IMs (i.e. 'borrowing' information from the prediction models for other individuals in the sample in an effort to improve predictive performance). A summary of the PEM workflow is illustrated in online Supplementary Fig. S1.

For the GLMER approach, we used a subject-specific random intercept and for the PEM approach, we conducted 10-fold cross-validation to estimate the combination weights $\hat{w}^i$ and considered three different learning algorithms: support vector machine (SVM), GLM with elastic net penalty (ENet) and random forest (RF). We used them individually ($L = 1$) in three separate ensemble (PEM) models and in combination ($L = 3$) (i.e. a total of 4 separate ensemble models were tested). We refer to the PEMs with $L = 3$ as personalized double ensemble models (PDEM). See online supplement for additional details. R code for all analyses is available online (https://github.com/csfm269/PEM-emotion).

Evaluating prediction performance

We used 10-fold cross-validation (CV) at the individual level to evaluate the subject-specific performance of the PEMs and report the average prediction accuracy derived from the 10-fold CV across all participants. We used Area Under the Receiver Operating Characteristic (AUROC) curve as the metric of accuracy and visualize the average ROCs across subjects for different models.

Decision curve analysis

Decision curves (Perry et al., 2021; Vickers, van Calster, & Steyerberg, 2019) are plotted to show how the prediction models could benefit adolescents if used to inform the timing of smartphone-delivered interventions for elevated negative emotions. Ultimately, the goal of developing prediction models for HNA states is to provide just-in-time (JIT) intervention for adolescents. Specifically, if the predicted probability of a HNA state for an individual is larger than a pre-defined threshold p^*, a JIT smartphone-based intervention could be automatically triggered (e.g. via a smartphone notification or suggested exercise via a mental health app). However, since such predictions cannot be perfect, there will be false positives (FP; i.e. the model falsely predicts that an individual is currently in a HNA state, which triggers an unneeded smartphone-based intervention) and false negatives (FN; i.e. the model falsely predicts that an individual is *not* in a HNA state, and thus no intervention is triggered when one could be helpful) associated with a given prediction model. Individuals likely differ substantially in how much weight they personally place on avoiding FP *v*. FN: some put more weight on avoiding FN over FP, which suggests the individual is relatively more tolerant of receiving smartphone-triggered interventions, even if some of the interventions are unnecessary; whereas other individuals may place more weight on avoiding FP and thus interventions should only be delivered when models predict HNA states with high confidence. These differences in preference can be quantified by the threshold probability p^*: $p^* < 0.5$ corresponds to individuals who place more weight on avoiding FN over FP, while $p^* > 0.5$ corresponds to individuals who place relatively more weight on avoiding FP. We use a decision curve to capture the utility of a prediction model when the relative impact of FN *v*. FP varies, as captured by p^*. Consistent with prior studies (Vickers, Calster, & Steyerberg, 2016), we define the benefit of a prediction model at a specific p^* as

$$Benefit = \frac{\#TP}{n} - \frac{\#FP}{n} \times \frac{p^*}{1 - p^*},$$

where #TP is the number of true positives, #FP is the number of false positives and n is the total number of events to be considered. The prediction model with the highest benefit at a particular threshold probability p^* has the highest clinical value.

Results

Depressive symptom scores ranged from none to severe (CES-D range = 1–51) with a mean of 6.3 (S.D. = 3.5) for the TD group and 40.7 (S.D. = 7.6) for the AH group. Fifty-nine percent of the sample had clinically significant levels of depressive symptoms (CES-D ⩾ 16) (Chwastiak et al., 2002). Anhedonia (SHAPS) scores ranged from 14 to 48 with a mean of 17.3 (S.D. = 3.5) for the TD group and 37.0 (S.D. = 6.5) for the AH group. There are no established norms for severity ranges for the SHAPS. However, a recent meta-analysis of 168 studies ($n > 16\,000$) revealed that a score of 25 represented the 99th percentile for healthy controls (Trøstheim et al., 2020). Fifty-five percent of our sample had a SHAPS score > 25 (see Table 2 for demographic and clinical characteristics of the sample).

Personalized predictions of HNA states

The mean proportion of HNA states across participants, as well as the mean difference between the observed emotion scores of HNA states and the person-specific mean emotion score are listed in Table 1. We can see that the HNA states were present in 28–32% of all EMA survey timepoints. On average, HNA states were approximately 1.06 to 1.25 points above their person-specific mean emotion score. Given that the mean within-person standard deviation (S.D.) of emotions ranged from 0.79 (anger) to 1.00 (nervousness), HNA states were, on average, approximately 1.2 to 1.4 S.D. above an individual's mean negative emotion score.

As described above, we considered four PEMs (ENet, SVM, RF and PDEM) as well as GLMER to predict HNA states for each emotion separately. The ROCs for all models and emotions are shown in Fig. 1 and the corresponding AUCs are listed in Table 3. The table also lists the accuracy of each model at their optimal cut-off values, defined as the values that minimize the Euclidean distance between (specificity, sensitivity) and (1, 1), i.e. the point on the ROC that is closest to the top left corner (Unal, 2017). PDEM and PEM-RF have superior predictive performance in comparison to the other three models. The linear mixed model (GLMER) had the poorest performance across all models.

Idiosyncratic (subject-specific) models (IMs) v. personalized ensemble models (PEM)

In Fig. 2, we plot the difference in Brier scores and AUCs between PEMs and IMs when applied to predict outcomes for each subject. We prioritize Brier scores since they capture both the calibration and discrimination aspects of a prediction model (Steyerberg et al., 2010). In addition, the Brier score is more reliable than AUC when the number of observations is relatively small (Steyerberg et al., 2010). The boxplot displays the distribution

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

Table 2. Demographics and clinical characteristics of the sample

Sample characteristics					
Anhedonic	*N*	%	Typically developing	*N*	%
Biological Sex			Biological Sex		
Female	12	92.3	Female	6	66.7
Male	1	7.6	Male	3	33.4
Race			Race		
American Indian or Alaska Native	0	0.0	American Indian or Alaska Native	0	0.0
Asian	1	7.6	Asian	0	0.0
Black or African American	1	7.6	Black or African American	0	0.0
Native Hawaiian or Other	0	0.0	Native Hawaiian or Other	1	11.1
Pacific Islander White	11	84.6	Pacific Islander White	8	88.9
Ethnicity			Ethnicity		
Hispanic or Latino.	1	7.6	Hispanic or Latino	1	11.1
Not Hispanic or Latino.	12	92.3	Not Hispanic or Latino.	8	88.9
Current Diagnoses (DSM-V)			Current Diagnoses (DSM-V)		
Major Depressive Episode	10	76.9	Major Depressive Episode	0	0.0
Generalized Anxiety Disorder	4	30.8	Generalized Anxiety Disorder	0	0.0
Social Anxiety Disorder	2	15.4	Social Anxiety Disorder	0	0.0
Panic Disorder	0	3.5	Panic Disorder	0	0.0
Specific Phobia	0	3.5	Specific Phobia	0	0.0
Attention-Deficit/Hyperactivity Disorder	1	7.6	Attention-Deficit/Hyperactivity Disorder	0	0.0
Oppositional Defiant Disorder	0	3.5	Oppositional Defiant Disorder	0	0.0
Medication			Medication		
SSRI	4	30.8	SSRI	0	0.0
	M	S.D.		*M*	S.D.
Age (in years)	15.9	1.8	Age (in years)	17.0	1.0
Family Income (dollars)	122 727	85 801	Family Income (dollars)	154 777	77 748
SHAPS Score	37.0	6.5	SHAPS Score	17.3	3.5
CES-D Score	40.7	7.6	CES-D Score	6.3	3.5

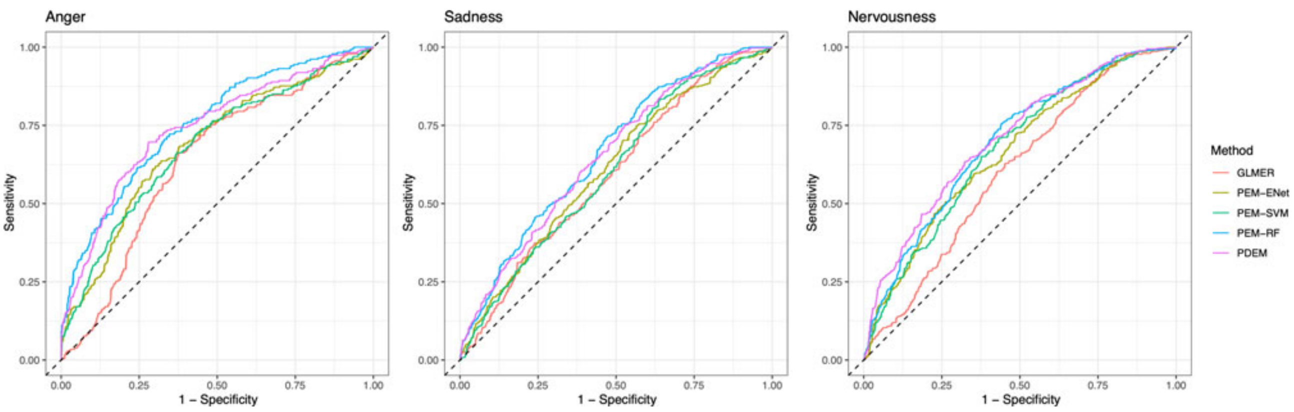

Fig. 1. Receiver operating characteristic (ROC) curve of different models for the prediction of high negative affect (HNA) states for each emotion.

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

Table 3. Prediction performance of different models when predicting HNA states for each of the three emotions

	AUC			Accuracy		
Method	Anger	Sadness	Nervousness	Anger	Sadness	Nervousness
GLMER	0.64	0.59	0.60	0.63	0.56	0.58
PEM-ENet	0.69	0.61	0.67	0.67	0.59	0.63
PEM-SVM	0.70	0.60	0.67	0.64	0.57	0.63
PEM-RF	0.73	0.66	0.70	0.69	0.63	0.66
PDEM	0.74	0.66	0.71	0.71	0.65	0.67

Note: GLMER, Generalized linear mixed-effects regression; PEM-ENet, Personalized ensemble model with Elastic net; PEM-SVM, PEM with support vector machine; PEM-RF, PEM with random forest; PDEM, Personalized double ensemble model. AUC, area under the curve. Accuracies (right three columns) are based on optimal cut-off values (0.3–0.4 across models).

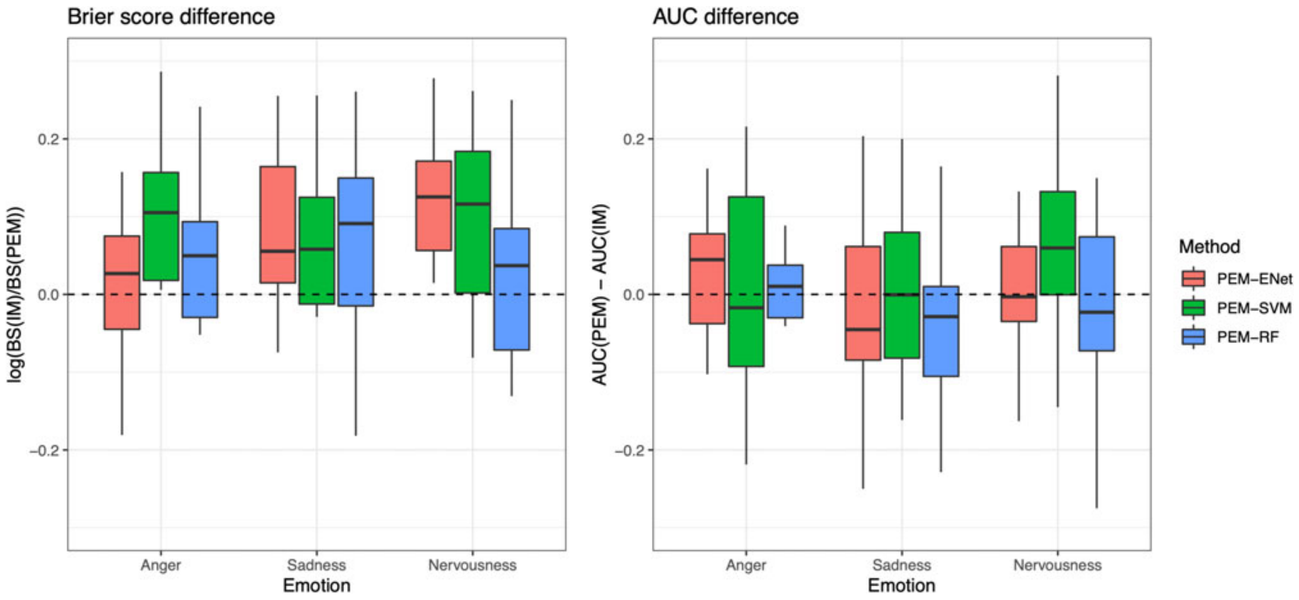

Fig. 2. Differences in Brier score (BS) and area under the curve (AUC) of idiosyncratic models (IM) *v.* personalized ensemble models (PEM). The boxplot displays the distribution of the difference across all subjects.

of the difference across all subjects. From the plots we see that, for each emotion, PEM was associated with superior Brier scores over IM for most subjects (i.e. scores above the horizontal dotted line). The most consistent improvement is observed for PEMs combining IMs trained with SVM (green boxplots). The improvement in subject-specific AUCs however is not as pronounced as the Brier scores, which is due to bias and instability of AUC estimates derived from cross-validation in small sample scenarios (Airola, Pahikkala, Waegeman, De Baets, & Salakoski, 2009).

To see more explicitly the difference between a PEM and the corresponding IMs, we plot the combination weights for every PEM-RF in online Supplementary Fig. S5. Each row in the plot displays (with red shading) the weights assigned to each IM. Notice that if IMs yielded the best predictive performance, all weights (red) would be assigned to the squares along the diagonal dotted line (i.e. all weights assigned to the target individual's IM). However, we find that subject *k*'s PEM-RF typically assigns non-trivial weights to *other* subjects' IMs. In summary, these findings combined with the results summarized in Fig. 2 suggest that personalized predictions for a given participant benefit from borrowing information from other individuals.

Decision curve analysis

In Fig. 3, we illustrate the decision curves of different prediction models, as the threshold p^* varies. We can see that for anger, all PEMs will benefit individuals when their concern over FP does not outweighs FN too much ($p^* < 0.9$). PEM-RF and PDEM have the largest benefit overall and PDEM outperform PEM-RF when $p^* < 0.4$. For sadness, none of the prediction models will benefit individuals if their concern over FP outweighs FN ($p^* > 0.5$). For nervousness, PEMs have positive benefit when $p^* <$ 0.75 and PEM-RF and PDEM are the best performing models with indistinguishable decision curves. The advantages of the best performing PEMs over other models are maximized when $p^* \approx 0.3$ (i.e. a predicted probability of 0.3 or greater is used to define the presence of a HNA state) for all three emotions.

Discussion

In the present study, passive smartphone sensor-derived variables were submitted to a novel personalized ML approach to predict states of heightened negative affect in adolescents. Models

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

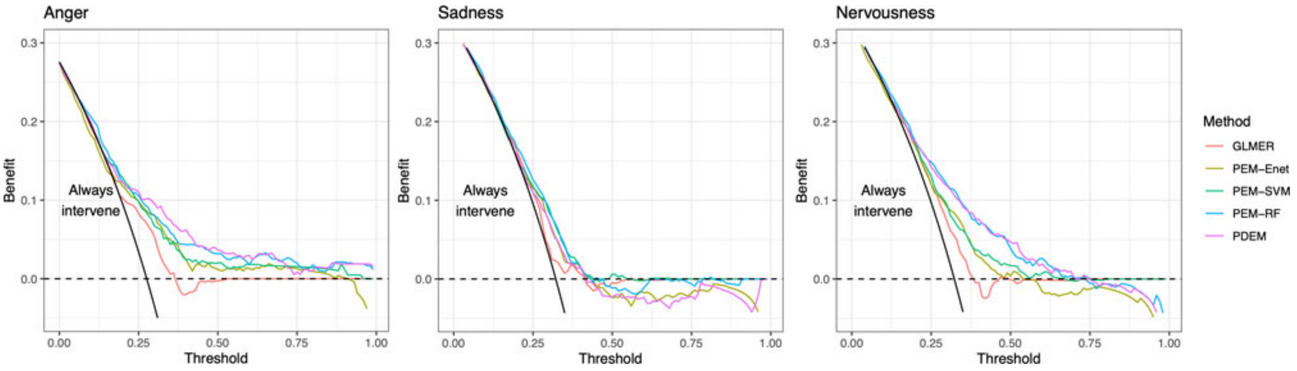

Fig. 3. Decision curve analysis: Benefit of decision rules based on different prediction models for all three emotions. The solid black lines reflect the decision rule that always predicts HNA states at every timepoint and the dashed black lines corresponds to the decision rule that predicts none of the time points to be HNA.

predicted states of elevated anger with the highest accuracy (71%; AUC = 0.74), with relatively more modest accuracy for predicting states of high anxiety (67%; AUC = 0.71) or sadness (65%; AUC = 0.66). To the extent that these findings are replicated – and importantly, predictive performance further improved by incorporating additional smartphone-derived features potentially predictive of heightened negative affect (e.g. call/text metadata, patterns of social media usage, or peripheral measures of physiological arousal from wearables) – these predictive models could ultimately be used to trigger brief 'just-in-time' (JIT) interventions via smartphone notifications or mental health apps to alleviate heightened negative emotions in real-time in the daily lives of teens.

Several notable findings emerged. First, our ML model outperformed a conventional statistical approach (GLMER) in predicting each negative emotion. In contrast, there are other examples in which ML does not improve predictive performance over conventional statistical approaches, such as ordinary least squares (OLS) regression and GLMER (Christodoulou et al., 2019; Webb & Cohen, 2021). ML approaches, such as random forest, are more likely to outperform conventional statistical approaches when a relatively large number of predictor variables are included and complex predictor-outcome associations are present in the data, including non-linear relations and higher-order interactions (Kessler & Luedtke, 2021), which may be the case with respect to the association between affective states and smartphone-derived variables such as activity level (accelerometer) and locations visited (GPS). Second, our ML approach (i.e. which involved a data-driven search of the optimal combination of subject-specific models) outperformed a fully idiographic model (i.e. person-specific models that do not borrow any information from other individuals' models) (Fig. 2). Idiographic models are becoming increasingly popular in clinical psychological science (Fisher, Medaglia, & Jeronimus, 2018; Wright & Woods, 2020), in line with the intuitive notion that every individual is different and thus requires a personalized modeling approach. Conventional (e.g. regression models) statistical approaches ignore this fact by aggregating across individuals (Fisher et al., 2018). The ensemble ML approach in this study used a data-driven search to identify the combination of subject-specific models that yielded the best predictive performance. If it were the case that the models that yielded the best predictive performance were the fully idiographic models, then we would expect the ensemble approach to assign all model weights to the individual's own model (i.e. in online Supplementary Fig. S5, all weights (red shading) would be assigned to the squares along the diagonal dotted line). Instead, as seen in this figure, the data-driven search applied non-zero weights to the subject-specific models of other individuals, indicating that predictive performance was improved by 'borrowing' information from other individual's models. In other words, these findings suggest that there is meaningful shared signal between individuals in the predictive relationships between smartphone sensor data and negative affective states.

As seen in online Supplementary Fig. S4, averaging across all participants, activity level ('accelerometer score') was the top predictor for all three emotions. Given the person-specific modeling approach, individuals could have very different rankings of which variables contributed most strongly to the predictions of negative emotional states. Activity level ranked among the top three strongest predictors in 64% (outcome: anger), 79% (outcome: sadness) and 72% (outcome: nervousness) of participants. However, and in support of personalized modeling approaches, there were substantial differences across individuals in the directionality of the association between predictors and outcome.

The fact that predictive performance differed meaningfully across affect variables (highest for anger and lowest for sadness) suggests that some negative emotions may be relatively more strongly linked to passive smartphone features such as activity level, location and phone use patterns. In addition, these findings suggest that it may be important for future studies to examine sadness, nervousness and anger separately rather than averaging them into a single negative affect variable, as is commonly done in studies. No clear pattern emerged in terms of which variables were most predictive of each negative emotion. Thus, it is unclear why differences in predictive performance emerged between emotions. Additional research is needed to test whether our pattern of findings is replicated.

Ultimately, prediction models could be used to trigger interventions directly to an adolescent's smartphone during states of HNA. Research is needed to test which brief smartphone-delivered interventions (e.g. app-based mindfulness exercise or behavioral activation) would be beneficial for an adolescent experiencing a particular negative emotional state (high anger *v.* sadness *v.* nervousness). A decision curve analysis was conducted as an initial step to estimate under what conditions (i.e. rate of false positive (FN) *v.* false negatives (FN), and their relative weighting by an individual) such a predictive model could be beneficial. Decision curves revealed that the prediction models developed in this study would be most beneficial for adolescents with a relatively high tolerance for FP (i.e. the model predicts that an adolescent is in a state of HNA when they are not) relative

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press

to FN (i.e. falsely predicting that an adolescent is *not* in a state of HNA). In other words, our prediction models would not be beneficial for those adolescents who are relatively sensitive to FP (i.e. receiving a smartphone-delivered intervention when they are not in a state of HNA). As noted above, we did not assess participants' individual tolerance for FP relative to FN. A future study could assess these personal preferences and could individually titrate the threshold (i.e. predicted probability of being in a HNA state) which triggers a smartphone intervention. Namely, users could be given the option to adjust the threshold over time based on their personal preferences (e.g. via a dial on an app which controls the user-specific threshold needed to be reached to trigger an intervention).

There were several limitations to this study. First, predictions of affect were based on a limited set of smartphone-derived variables. Other relevant features extracted from smartphones may improve predictive performance (e.g. meta-data on calls/texts, social media use, recorded vocal tone/characteristics). Future studies could consider how these variables and others (e.g. peripheral measures of physiological arousal from wearables) could improve predictions of heightened negative affect, in particular for sadness which was associated with the poorest predictive performance. Furthermore, although this study focused on *within*-person predictions, the sample size was small. Finally, a denser EMA sampling strategy with participants completing more affect surveys per day would provide more granularity in assessments of affect fluctuations and a larger within-person dataset to use for modeling. These limitations notwithstanding, the present findings provide preliminary support for the use of passively collected smartphone data to predict states of affective distress in adolescents, which could ultimately be translated into timely interventions to alleviate these states and, perhaps, reduce future risk of affective disorder onset.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291722002161.

Financial support. This research was supported by NIMH K23MH108752 (C. A. W.). C. A. W. was partially supported by NIMH R01MH116969, NCCIH R01AT011002, the Tommy Fuss Fund and a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation.

Conflict of interest. Over the past 3 years, Dr Pizzagalli has received consulting fees from Albright Stonebridge Group, BlackThorn Therapeutics, Boehringer Ingelheim, Compass Pathway, Concert Pharmaceuticals, Engrail Therapeutics, Neuroscience Software, Neurocrine Biosciences, Otsuka Pharmaceuticals, and Takeda Pharmaceuticals, as well as one honorarium from the Psychonomic Society (for editorial work) and Alkermes. In addition, he has received stock options from BlackThorn Therapeutics and Compass Pathways, as well as research support from National Institute of Mental Health, Dana Foundation, Brain and Behavior Research Foundation, and Millennium Pharmaceuticals. Dr Baker has received consulting fees from Verily Life Sciences, Inc., as well as consulting fees and equity from Mindstrong Health, Inc., for unrelated work.

References

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., & Salakoski, T. (2009). A comparison of AUC estimators in small-sample studies. *Machine Learning in Systems Biology*, *1*, 3–13.

Avenevoli, S., Swendsen, J., He, J.-P., Burstein, M., & Merikangas, K. R. (2015). Major depression in the national comorbidity survey–adolescent supplement: Prevalence, correlates, and treatment. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*(1), 37–44.e2. https://doi.org/10.1016/j.jaac.2014.10.010.

Bailen, N. H., Green, L. M., & Thompson, R. J. (2019). Understanding emotion in adolescents: A review of emotional frequency, intensity, instability, and clarity. *Emotion Review*, *11*(1), 63–73.

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R* (1st ed.). Boca Raton: Chapman and Hall/CRC.

Borghuis, J., Denissen, J. J., Oberski, D., Sijtsma, K., Meeus, W. H., Branje, S., … Bleidorn, W. (2017). Big five personality stability, change, and codevelopment across adolescence and early adulthood. *Journal of Personality and Social Psychology*, *113*(4), 641.

Cao, J., Truong, A. L., Banu, S., Shah, A. A., Sabharwal, A., & Moukaddam, N. (2020). Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: Development and usability study. *JMIR Mental Health*, *7*(1), e14045.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22.

Chwastiak, L., Ehde, D. M., Gibbons, L. E., Sullivan, M., Bowen, J. D., & Kraft, G. H. (2002). Depressive symptoms and severity of illness in multiple sclerosis: Epidemiologic study of a large community sample. *American Journal of Psychiatry*, *159*(11), 1862–1868. https://doi.org/10.1176/appi.ajp.159.11.1862.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115.

Forbes, E. E., Hariri, A. R., Martin, S. L., Silk, J. S., Moyles, D. L., Fisher, P. M., … Dahl, R. E. (2009). Altered striatal activation predicting real-world positive affect in adolescent major depressive disorder. *American Journal of Psychiatry*, *166*(1), 64–73. https://doi.org/10.1176/appi.ajp.2008.07081336.

Forbes, E. E., Stepp, S. D., Dahl, R. E., Ryan, N. D., Whalen, D., Axelson, D. A., … Silk, J. S. (2012). Real-world affect and social context as predictors of treatment response in child and adolescent depression and anxiety: An ecological momentary assessment study. *Journal of Child and Adolescent Psychopharmacology*, *22*(1), 37–47. https://doi.org/10.1089/cap.2011.0085.

Frost, A., Hoyt, L. T., Chung, A. L., & Adam, E. K. (2015). Daily life with depressive symptoms: Gender differences in adolescents' everyday emotional experiences. *Journal of Adolescence*, *43*, 132–141. https://doi.org/10.1016/j.adolescence.2015.06.001.

Griffith, J. M., Clark, H. M., Haraden, D. A., Young, J. F., & Hankin, B. L. (2021). Affective development from middle childhood to late adolescence: Trajectories of mean-level change in negative and positive affect. *Journal of Youth and Adolescence*, *50*(8), 1550–1563.

Jacobson, N. C., & Chung, Y. J. (2020). Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors*, *20*(12), 3572.

Kaufman, J., Birmaher, B., Axelson, D., Perepletchikova, F., Brent, D., & Ryan, N. (2013). *Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version* (*K-SADS-PL 2013, DSM-5*). Pittsburgh, PA: Western Psychiatric Institute and Yale University.

Kessler, R. C., & Luedtke, A. (2021). Pragmatic precision psychiatry-a new direction for optimizing treatment selection. *JAMA Psychiatry*, *78*(12), 1384–1390. https://doi.org/10.1001/jamapsychiatry.2021.2500.

Larson, R. W., Moneta, G., Richards, M. H., & Wilson, S. (2002). Continuity, stability, and change in daily emotional experience across adolescence. *Child Development*, *73*(4), 1151–1165.

Laurent, J., Catanzaro, S. J., Joiner Jr., T. E., Rudolph, K. D., Potter, K. I., Lambert, S., … Gathright, T. (1999). A measure of positive and negative affect for children: Scale development and preliminary validation. *Psychological Assessment*, *11*(3), 326–338. https://doi.org/10.1037/1040-3590.11.3.326.

Perry, B. I., Osimo, E. F., Upthegrove, R., Mallikarjun, P. K., Yorke, J., Stochl, J., … Khandaker, G. M. (2021). Development and external validation of the psychosis metabolic risk calculator (PsyMetRiC): A cardiometabolic risk prediction algorithm for young people with psychosis. *The Lancet Psychiatry*, *8*(7), 589–598. https://doi.org/10.1016/S2215-0366(21)00114-0.

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1* (3), 385–401. https://doi.org/10.1177/014662167700100306.

Rahimi-Eichi, H., Iii, G. C., Bustamante, C. M. V., Onnela, J.-P., Baker, J. T., & Buckner, R. L. (2021). Open-source longitudinal sleep analysis from accelerometer data (DPSleep): Algorithm development and validation. *JMIR MHealth and UHealth*, *9*(10), e29849. https://doi.org/10.2196/29849.

Ren, B., Patil, P., Dominici, F., Parmigiani, G., & Trippa, L. (2021). Cross-study learning for generalist and specialist predictions. *ArXiv:2007.12807 [Math, Stat]*. Retrieved from http://arxiv.org/abs/2007.12807.

Rideout, V., & Robb, M. B. (2019). The common sense census: Media use by tweens and teens, 2019 | Common Sense Media. Retrieved August 19, 2020, from https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2019.

Sequeira, L., Perrotta, S., LaGrassa, J., Merikangas, K., Kreindler, D., Kundur, D., … Strauss, J. (2020). Mobile and wearable technology for monitoring depressive symptoms in children and adolescents: A scoping review. *Journal of Affective Disorders*, *265*, 314–324.

Shah, R. V., Grennan, G., Zafar-Khan, M., Alim, F., Dey, S., Ramanathan, D., & Mishra, J. (2021). Personalized machine learning of depressed mood using wearables. *Translational Psychiatry*, *11*(1), 1–18. https://doi.org/10.1038/s41398-021-01445-0.

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. *The British Journal of Psychiatry*, *167*(1), 99–103. https://doi.org/10.1192/bjp.167.1.99.

Soto, C. J. (2016). The little six personality dimensions from early childhood to early adulthood: Mean-level age and gender differences in parents' reports. *Journal of Personality*, *84*(4), 409–422.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*(2), 330.

Squeglia, L. M., Ball, T. M., Jacobus, J., Brumback, T., McKenna, B. S., Nguyen-Louie, T. T., … Tapert, S. F. (2016). Neural predictors of initiating alcohol use during adolescence. *American Journal of Psychiatry*, *174*(2), 172–185. https://doi.org/10.1176/appi.ajp.2016.15121587.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data-driven smartphone research. *JMIR Mental Health*, *3*(2), e5165. https://doi.org/10.2196/mental.5165.

Trøstheim, M., Eikemo, M., Meir, R., Hansen, I., Paul, E., Kroll, S. L., … Leknes, S. (2020). Assessment of anhedonia in adults with and without mental illness: A systematic review and meta-analysis. *JAMA Network Open*, *3*(8), e2013233. https://doi.org/10.1001/jamanetworkopen.2020.13233.

Unal, I. (2017). Defining an optimal cut-point value in ROC analysis: An alternative approach. *Computational and Mathematical Methods in Medicine*, *6*, 1–14. https://doi.org/10.1155/2017/3762651.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03.

Vickers, A. J., Calster, B. V., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, *352*, i6. https://doi.org/10.1136/bmj.i6.

Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research*, *3*(1), 18. https://doi.org/10.1186/s41512-019-0064-7.

Vinograd, M., Williams, A., Sun, M., Bobova, L., Wolitzky-Taylor, K. B., Vrshek-Schallhorn, S., … Craske, M. G. (2020). Neuroticism and interpretive bias as risk factors for anxiety and depression. *Clinical Psychological Science*, *8*(4), 641–656.

Webb, C. A., & Cohen, Z. D. (2021). Progress towards clinically informative data-driven decision support tools in psychotherapy. *The Lancet Digital Health*, *3*(4), e207–e208.

Webb, C. A., Israel, E. S., Belleau, E., Appleman, L., Forbes, E. E., & Pizzagalli, D. A. (2021). Mind-wandering in adolescents predicts worse affect and is linked to aberrant default mode network–salience network connectivity. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(3), 377–387. https://doi.org/10.1016/j.jaac.2020.03.010.

Wright, A. G., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology*, *16*, 49–74.

Zinbarg, R. E., Mineka, S., Bobova, L., Craske, M. G., Vrshek-Schallhorn, S., Griffith, J. W., … Anand, D. (2016). Testing a hierarchical model of neuroticism and its cognitive facets: Latent structure and prospective prediction of first onsets of anxiety and unipolar mood disorders during 3 years in late adolescence. *Clinical Psychological Science*, *4*(5), 805–824.

https://doi.org/10.1017/S0033291722002161 Published online by Cambridge University Press