Journal of Affective Disorders 320 (2023) 201–210

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research paper

Applying ensemble machine learning models to predict individual response to a digitally delivered worry postponement intervention

Joseph A. Gyorda [a,b,*], Matthew D. Nemesure [a,c], George Price [a,c], Nicholas C. Jacobson [a,c,d,e]

[a] Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States
[b] Mathematical Data Science Program, Dartmouth College, Hanover, NH, United States
[c] Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH, United States
[d] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States
[e] Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States

ARTICLE INFO

Keywords:
Generalized anxiety disorder
Worry dynamics
Digital intervention
Personalized mental health care
Machine learning

ABSTRACT

Objective: Generalized anxiety disorder (GAD) is a prevalent mental health disorder that often goes untreated. A core aspect of GAD is worry, which is associated with negative health outcomes, accentuating a need for simple treatments for worry. The present study leveraged pretreatment individual differences to predict personalized treatment response to a digital intervention.

Methods: Linear mixed-effect models were used to model changes in daytime and nighttime worry duration and frequency for 163 participants who completed a six-day worry postponement intervention. Ensemble-based machine learning regression and classification models were implemented to predict changes in worry across the intervention. Model feature importance was derived using SHapley Additive exPlanation (SHAP).

Results: Moderate predictive performance was obtained for predicting changes in daytime worry duration (test $r^2 = 0.221$, AUC = 0.77) and nighttime worry frequency (test $r^2 = 0.164$, AUC = 0.72), while poor predictive performance was obtained for nighttime worry duration and daytime worry frequency. Baseline levels of worry and subjective health complaints were most important in driving model predictions.

Limitations: A complete-case analysis was leveraged to analyze the present data, which was collected from participants that were Dutch and majority female.

Conclusions: This study suggests that treatment response to a digital intervention for GAD can be accurately predicted using baseline characteristics. Particularly, this worry postponement intervention may be most beneficial for individuals with high baseline worry but fewer subjective health complaints. The present findings highlight the complexities of and need for further research into daily worry dynamics and the personalizable utility of digital interventions.

1. Introduction

Generalized anxiety disorder (GAD) is characterized by chronic, persistent, and excessive worry (Stein and Sareen, 2015) experienced over an extended period of time (e.g., at least six months; Ruscio et al., 2017). Being one of the most common anxiety disorders, GAD has a lifetime prevalence of nearly 8 % in the US (Ruscio et al., 2017). In particular, women have a higher likelihood of experiencing GAD (Grenier et al., 2019; Haller et al., 2014), with estimates suggesting that women have twice the lifetime prevalence of men (Merikangas et al., 2010). Furthermore, the occurrence of GAD progressively increases throughout adolescence, with roughly 25 % of lifetime cases beginning before age 20 (Tiirikainen et al., 2019). A large contributing factor to the high lifetime prevalence is that patients with GAD often do not seek clinical treatment until many years after the onset of their symptoms (Stein and Sareen, 2015; Thompson et al., 2008). This hesitance to seek treatment poses a major health concern, as the development of GAD increases the likelihood of subsequently developing a comorbid disorder (Jacobson and Newman, 2017), with estimates suggesting that over 80 % of people with GAD proceed to develop a comorbid disorder (Ruscio et al., 2017). Given worsened health outcomes and the risk of heightened resistance to treatment resulting from comorbidity (Coplan et al.,

* Corresponding author at: Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, 46 Centerra Parkway, Suite 300, Lebanon, NH 03766, United States.
E-mail address: joseph.a.gyorda.22@dartmouth.edu (J.A. Gyorda).

https://doi.org/10.1016/j.jad.2022.09.112
Received 9 March 2022; Received in revised form 2 September 2022; Accepted 20 September 2022
Available online 24 September 2022

2015; Pelletier et al., 2017), it is paramount that research be done to augment the availability of simple, accessible treatments for conditions such as GAD.

The hallmark feature of GAD is pathological worry (Olatunji et al., 2010). Worry is often defined as a repetitive chain of negative thoughts and images that may be uncontrollable and are typically attributed to future events (Querstret and Cropley, 2013). People with high amounts of worry may experience negative health outcomes (Goodwin et al., 2017) and have a restricted working memory capacity when they worry, leading to difficulties redirecting their attention away from worrying thoughts (Hayes et al., 2008). This attention bias to threatening stimuli likely leads to and maintains more severe anxiety symptoms as well as perpetuates worrisome thoughts in many of these individuals (Goodwin et al., 2017; Olatunji et al., 2010). Repetitive negative, stress-related thoughts (e.g., worrying) are associated with myriad somatic health outcomes, such as increased blood pressure and heart rate, decreased antibody production, and changes in health behaviors including sleep, diet, and substance consumption (Brosschot and van der Doef, 2006; McCarrick et al., 2021). Intense and prolonged worry therefore potentiates negative mental and physiological health outcomes.

Given the prevalence of worry and GAD more broadly, along with the austerity of health outcomes for high-worrying individuals, it is imperative that research be done into efficacious, easily accessible treatments for worry. Unfortunately, many of the current evidence based treatments, such as cognitive behavioral therapy (Querstret and Cropley, 2013) and acceptance and commitment therapy (McCarrick et al., 2021) require interaction with a trained health care provider (Paxling et al., 2011). This is problematic because there is at most only one psychologist or psychiatrist for every 3800+ people in the US (Ku et al., 2021). Adding to the burden on individual providers, research also suggests that the population of mental health care professionals is both aging and declining (Butryn et al., 2017). These statistics emphasize the utility of personalizable interventions that can be delivered via a technological medium (i.e., smartphone, computer, tablet). Examples of such treatments include mindfulness (Delgado et al., 2010), psychological detachment (McCarrick et al., 2021), and worry postponement interventions (Brosschot and van der Doef, 2006; Versluis et al., 2016), all of which can be performed on one's own without a mental health professional. Furthermore, remote interventions of this type are beneficial to those unable or unwilling to attend clinical facilities, an important consideration in light of the COVID-19 pandemic (Hirsch et al., 2021). Given the reduced provider burden facilitated by digital administration, such interventions are therefore low cost, low commitment (e.g., they may only require 30 minutes each day), easily-accessible, and offer flexibility in scheduling (Wilhelm et al., 2020).

The present study specifically examines the personalized efficacy of a worry postponement intervention method. In particular, Versluis et al. (2016) implemented a randomized controlled trial in which participants received online instruction to log their worry over a six-day period. Members of the experimental group received additional instruction to complete a worry postponement intervention, whereas members of the control group simply logged their worry. The original analysis concluded, among other things, that while worry duration and frequency during both the day and night improved across the intervention period for the experimental group, linear mixed models revealed that the intervention did not yield significant improvements in comparison with the control group (Versluis et al., 2016). While the original study did not identify potential subgroups of individuals (e.g., those with high worry) that would be most likely to benefit from the intervention, previous works have experienced success implementing worry postponement interventions (Brosschot and van der Doef, 2006), and given the low-cost, low-burden design of this intervention, it is worthwhile further inspecting what factors, if any, may predispose individuals to benefitting from this intervention. The goal of this investigation is therefore to build a machine learning pipeline capable of predicting changes in both worry duration and frequency during the day and night for the participants in the study originally published by Versluis et al. (2016) and identifying what factors may have predisposed participants to a successful outcome. The significance of this endeavor lies in whether the intervention outcomes can be successfully predicted, as this will triage appropriate levels of care for individuals (i.e., suggesting individuals who are likely to benefit from a digital intervention of this type engage in the digital intervention, while referring individuals who are less likely to benefit to in-person care). Taken all together, the present study attempts to address the following questions:

(1) For those who completed the Internet-administered worry postponement intervention, how accurately can the changes in both daily/nightly worry duration and frequency be predicted using data collected prior to intervention administration?
(2) What individual patient characteristics available at baseline are most helpful in predicting the intervention's efficacy in ameliorating worry?

2. Methods

2.1. Participants

The present study leverages data from a worry postponement intervention implemented by Versluis et al. (2016), who conducted a non-stratified, randomized, parallel-group trial from 2005 to 2012. The Institutional Review Board at Leiden University granted approval for this study (Versluis et al., 2016). All participants were Dutch and were recruited via advertisements in electronic and print publications. Advertisements directed prospective participants to a website explaining that participants would log the daily amount of worry they experienced for six days. The website also asked that anyone who registered for the study complete the whole study, which included completing questionnaires and logging worry duration and frequency over six days. In total, 1035 people registered on the study's website. The only exclusion criteria was that participants were required to be 18 years of age or older, leaving 996 people eligible for the study (Versluis et al., 2016).

2.2. Measures

Three health measures were utilized by the original study to assess participant wellbeing at baseline and once the study was completed. One health measure was the Penn State Worry Questionnaire (PSWQ), which contains 16 items assessing worry measured on a 1–5 Likert scale; thus, scores from 16 to 39 indicate low worry, 40–59 indicate moderate worry, and 60–80 indicate high worry (Meyer et al., 1990). Another health measure administered was the Subjective Health Complaints (SHC) inventory, which had items assessing the severity (1–5 scale) and frequency (within the past three days) of 29 somatic and psychological complaints (Eriksen et al., 1999). The third health measure administered was the Positive and Negative Affect Schedule (PANAS), a 20-item assessment with 10 questions (1–5 scale) assessing positive affect and 10 questions (1–5 scale) assessing negative affect (Crawford and Henry, 2004; Watson et al., 1988).

2.3. Intervention protocol

Once registered, participants were randomly assigned to either a control (N = 498) or experimental group (N = 498). In both cases, participants were then instructed to complete a demographic questionnaire (e.g., age, gender, education, sleep, substance consumption), along with the three aforementioned health measures. A large number of participants in both the control group (N = 246; 49.4 %) and the experimental group (N = 262; 52.6 %) dropped out of the study during the baseline assessments. It is unknown why these individuals dropped out of the study; however, high dropout rates are common in Internet-administered interventions extended to entire communities

(Christensen et al., 2009). With this in consideration and given that the intervention had not yet begun, there is no reason to believe that the dropout rates had an impact on study outcomes.

Following baseline completion, the remaining participants were instructed to log, at the end of each day and each morning, their estimated number of worry episodes and duration of these episodes. Participants in the experimental group received additional instruction to set a special 30-minute period at the end of each day devoted to worrying (Versluis et al., 2016). The main idea was that if the participants realized they were worrying, then they should actively try to postpone their worry to this pre-defined 30-minute block of time at the end of the day. All participants were instructed to log their worry for six days. Of the 252 participants in the control group and 236 participants in the experimental group that completed baseline assessments, 60 (23.8 %) and 67 (28.4 %) participants, respectively, dropped out of the study partway through the six-day period (Versluis et al., 2016). These participants discontinued the study for unclear reasons, but no worsening of symptoms among these dropouts was noted by the original authors (Versluis et al., 2016). Once the intervention was finished, participants completed the same three health measures from baseline (PSWQ, SHC, PANAS) in addition to two questions assessing sleep quality and a question assessing the participant's ability to record their worry. Subjects in the intervention group were also asked how well they felt they postponed their worry. The final population of participants that completed the study was 361 (84.8 % female, $M_{age} = 36.36$; Versluis et al., 2016).

2.4. Data preprocessing

Participant data and a corresponding codebook were obtained from publicly available data published by the original investigators (Versluis et al., 2016). All data were read into Python version 3.8.5 and preprocessed for analysis using the *pandas* (Reback et al., 2021) and *NumPy* (Harris et al., 2020) packages. Given that the present goal is predicting individual treatment response, the data was filtered to only include participants from the intervention group. Following this exclusion, 169 participants were eligible for analysis. The authors of the original study elected to exclude participants in each group based on a few criteria that removed participants who recorded worry values that exceeded plausible limitations for daily or nightly worry. Two participants were excluded due to extreme daily worry (e.g., >14 hours); two participants were excluded due to extreme nightly worry (e.g., >6 hours); and six participants were excluded due to having worry frequency implausibly greater than worry duration (e.g., one participant registered 240 worry episodes and only 30 minutes of duration; Versluis et al., 2016). The same exclusion criteria were utilized for the current analysis; no participants were excluded on the basis of any other health concerns. The remaining individuals ($N = 163$, 82.8 % female, $M_{age} = 35.94$) were both part of the intervention group and had plausible worry data and thus represented the final data for analysis. This sample size is similar to other studies leveraging machine learning in the mental health domain (Cho et al., 2019). Descriptive statistics for age, gender, and individual questionnaire results for worry, health complaints, affect, reporting worry, and postponing worry were compiled for the final sample.

The present study leveraged only the information available at baseline for each of the 163 participants, meaning that any data collected after the intervention was excluded from analysis. This baseline data included demographic information and results from the three health measures (PSWQ, SHC, PANAS). In addition to reporting this baseline information, Versluis et al. (2016) recorded other metrics for each of the original 1035 registered participants. These variables included participant ID, whether a participant was included in the study, why the participant was excluded (if applicable), participant condition, whether the intervention was finished, date of last contact with the study (either through completion or dropping out), and at what point in the study the participant dropped out (if applicable). Given that the present study only considers participants in the experimental group who completed the intervention, these variables were all excluded from the present predictive analyses.

Using the available baseline data, additional features were engineered such that they provided unique information intuitively related to the worry outcomes of interest (see Outcome calculation subsection below) with the overarching goal of increasing downstream model performance as well as model interpretability (Lekkas et al., 2021b). Some features examined the interactions of general background information: for instance, the interactions of sleep duration and sleep quality, as well as alcohol and cigarette consumption, were both considered. Along with this, dummy coded variables for the year and season during which each participant completed the intervention were created. Furthermore, other engineered features combined the results of the baseline questionnaires. Many features were created using related specific items within a given questionnaire, since these items addressed specific health concerns such that taking an interaction of related items may distinguish participants more clearly. For instance, one feature was created as the interaction of two PANAS questionnaire items assessing guilt and shame. Other features examined between-survey interactions, including one feature examining the interaction between the overall scores of the PSWQ and SHC questionnaires. Individual questionnaire items, in addition to summative scores, have been leveraged in prior machine learning analyses (Gonzalez, 2021) and were included as predictors in this analysis as they address specific symptom-level information, thus improving downstream model interpretability. All new features were appended to the baseline data frame. The resulting number of features (157) was high relative to the sample size ($N = 163$); however, empirical evidence suggests that the optimal feature size is N-1 for uncorrelated features (Hua et al., 2005). The features in the present study exhibit little correlation (median = 0.13 after taking absolute value); therefore, the sample size was deemed sufficient to support the number of predictors included.

2.5. Outcome calculation

The overarching goal of the present study is to predict participant response to the worry postponement intervention. Although the original study examined the effects of the intervention on the scores of three health measures, these questionnaires are less able to capture the within- and between-day variability in worry than worry duration and frequency (Verkuil et al., 2021). Along with this, the original study measured worry on a daily basis, as opposed to the PSWQ and other questionnaires, which were measured only at baseline and following study completion; evidence suggests that repeated measures may be a better outcome measure for soft data (e.g., self-report affect/worry; Kraemer and Thiemann, 1989). Furthermore, uncontrollable and prolonged worry about a number of topics is the central defining feature of generalized anxiety disorder and thus is principally relevant as a primary outcome of the present analyses. Each participant recorded the duration and frequency of their worry episodes during both the day and night over the six-day intervention period. Hence, the four outcomes of interest were changes in daytime worry duration, nighttime worry duration, daytime worry frequency, and nighttime worry frequency. Given the temporal nature of the intervention, the present data has a multilevel structure, where observations across the six days are nested within each individual. Because an intervention may vary in efficacy from one individual to another, and considering the risk for harm from undergoing an unnecessary intervention (Suresh et al., 2017), it is important to account for individual effects and correlations when modeling the responses to the present intervention. Thus, in line with another machine learning study with nested longitudinal data (Zilcha-Mano et al., 2018), as well as a work suggesting random slopes may be better outcome measures for soft, repeated measure data (Kraemer and Thiemann, 1989), a mixed-effects model framework was implemented to model the four outcomes of interest, where trends in worry outcomes

can vary across participants over time.

All preprocessed data saved as a CSV file was read into R version 4.0.3 and prepared for linear modeling. The *lme4* package (Bates et al., 2015) was used to implement four linear mixed-effects models with identical predictors, where daytime worry duration, nighttime worry duration, daytime worry frequency, and nighttime worry frequency were the four response variables. Time was included as both a fixed effect and a random effect in all models. To generate the four outcomes of interest, the slope of the random effects of time for each participant was extracted from each mixed model. In sum, changes in worry duration and frequency at day and night for each participant were modeled with a random slope of time generated from a linear mixed model. These four outcomes were appended to the preprocessed data frame, which was then read back into Python for use in subsequent machine learning modeling. The distributions of the four outcomes were plotted using the *matplotlib* (Hunter, 2007) and *seaborn* (Waskom, 2021) Python libraries and are shown in Fig. 1 below. Furthermore, four more outcomes were created by binarizing each of the four continuous outcomes obtained from the mixed models to frame the present task as a classification problem. Binarization was performed by coding negative random slopes (e.g., improvement in worry) as 1 and non-negative random slopes (e.g., no change or deterioration in worry) as 0.

2.6. Machine learning methodology

Following outcome calculation, the following machine learning pipeline was implemented for each outcome. First, the baseline data was split into train and test sets, with 70 % being allocated to training and 30 % being allocated to testing. Data from 114 participants were thus allocated to the train set and data from 49 were allocated to the test set. A random state was set for reproducibility. Five-fold cross validation was used to iteratively split the train set into training and validation sets. Machine learning models were trained and evaluated on the validation set to generate predicted outcomes. Ensemble-based machine learning approaches have shown to give better predictive performance than individual models (Rokach, 2010) and have been commonly employed to good effect within the mental health literature (Lekkas et al., 2022; Lekkas et al., 2021a; Nemesure et al., 2021); therefore, an ensemble-based approach was used in which the predictions from multiple machine learning models were averaged together to create the final prediction. Hyperparameters for each model were tuned across the five-fold validation. Once the ensemble models were tuned, final predictions for the entire test set were generated by evaluating the ensemble models' predictions on the entire test set for each fold, and predictions were averaged together across the five folds to obtain the final predictions. Importantly, the hyperparameters for the ensemble approach (including both within-model hyperparameters and model selection) were never tuned to optimize test set performance. For the four continuous outcomes, the models implemented were regressors including linear, tree-based, support vector, ensemble, neural network, and neighbor models from Python's *sklearn* (Pedregosa et al., 2018), *xgboost* (Chen and Guestrin, 2016), and *lightgbm* (Ke et al., 2017) libraries. Many of these models penalize complexity such that only features that provide information are used in generating the final prediction for a given set of input data, helping to reduce the large feature space of the present study. For the four binary outcomes, the models implemented were classifiers

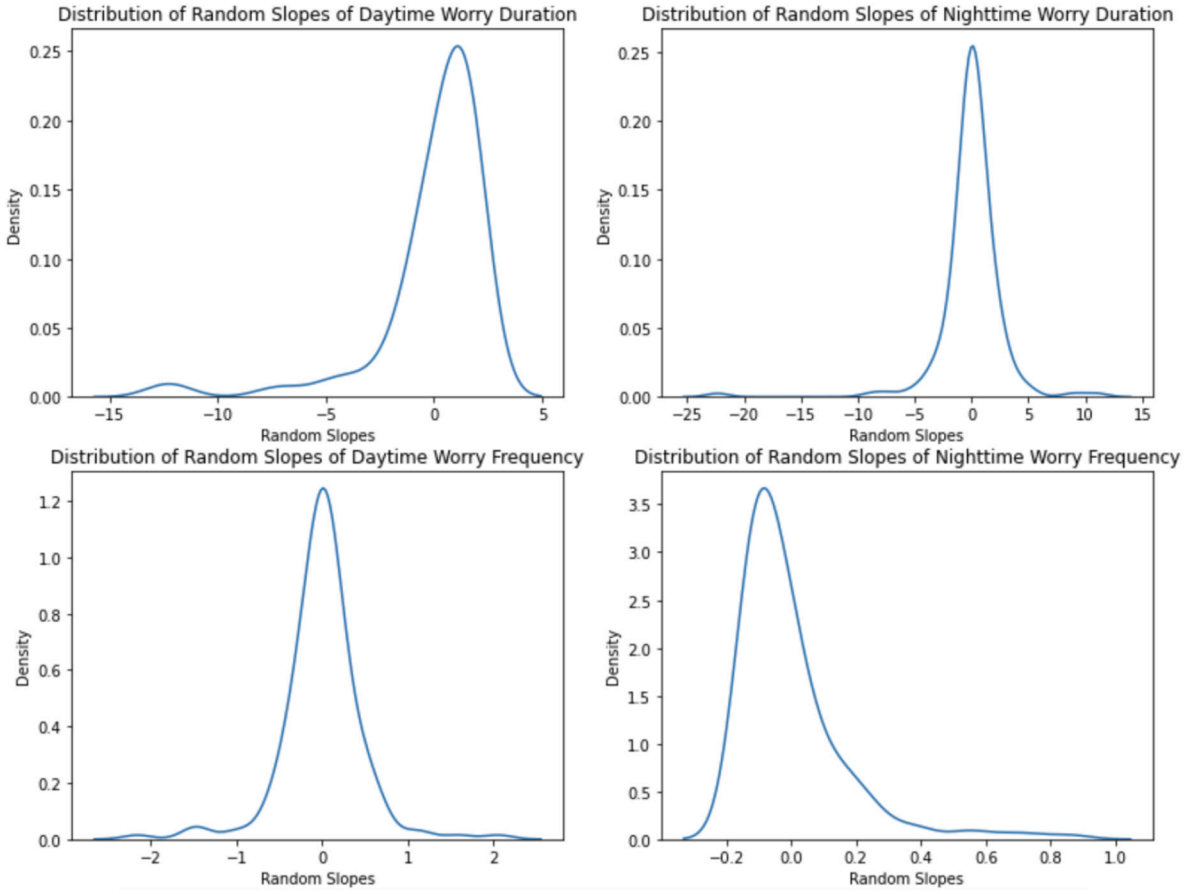

Fig. 1. *Distributions of the Continuous Outcomes of the Worry Postponement Intervention.*
Note: Each graph corresponds to one of the four continuous outcomes of interest from the present study. Top Left: Random slopes of daytime worry duration. Top Right: Random slopes of nighttime worry duration. Bottom Left: Random slopes of daytime worry frequency. Bottom Right: Random slopes of nighttime worry frequency.

including linear, tree-based, support vector, ensemble, neural network, naive Bayes, and neighbor models from Python's *sklearn*, *xgboost*, and *lightgbm* libraries.

2.7. Model evaluation

Various metrics were used to assess model fit. The regression models used to fit the four continuous outcomes were evaluated by calculating the R-squared values between the model predictions and actual outcomes. Metrics for both the validation and test sets were obtained. Along with this, the percent improvement in mean absolute error (MAE) over a naive approach, or simply predicting the mean, was calculated. Using the percent improvement makes the MAE agnostic to the scale of the data. Both R-squared and MAE are metrics commonly employed for machine learning model evaluation (Lekkas et al., 2021b). Furthermore, two metrics were used to account for the presence of outliers in the data. Spearman rank correlation was implemented with the *SciPy* Python package (Virtanen et al., 2020) to assess the association non-parametrically to account for outliers. Additionally, all metrics aside from spearman rank correlation were calculated on the outcomes with winsorization (Hoo et al., 2002) performed to replace outliers with values corresponding to the 5th and 95th percentiles of the data. Accuracy, specificity, sensitivity, and area under the curve (AUC)—all of which are commonly employed classification evaluation metrics (Lekkas et al., 2021b)—were calculated from the classification models used to fit the four binary outcomes. Sensitivity, or recall, may be of particular importance as it captures the proportion of positive cases that were predicted correctly. The positive class was defined as participants with improvement in their worry across the intervention period; therefore, sensitivity gives an assessment of how well the model is able to accurately predict individuals who respond well to the intervention. Lastly, additional analyses were completed assessing the model's prediction of worry in controls to ascertain how many people in the control condition would have benefitted from the intervention had they been a part of it (see Supplementary analysis), an idea adapted from the personalized advantage index as described by DeRubeis et al. (2014).

2.8. Model interpretation

One major issue with complex machine learning pipelines for prediction is the concept of the "black box". Essentially, the idea is that even if models are making good predictions, there is no way of knowing if they are learning true signals instead of bias or noise in the data. To circumvent this issue, SHapley Additive exPlanation (SHAP; Lundberg and Lee, 2017) was implemented as a wrapper for each model pipeline. This method allows for introspection into how each independent variable affects the model's prediction for each individual. This method works by iteratively adjusting model inputs and examining how the changes affect model predictions and thus builds a relationship between each variable and the predicted outcome for each person. This idea can be equated to creating "digital twins": copies of a participant's data with slight changes to assess how the model predicts differently. For each outcome's machine learning pipeline, the five most important features calculated by SHAP were reported.

3. Results

3.1. Participant demographics

Table 1 provides a summary of the baseline information of the participants considered in the present analysis. Individual questionnaire results for worry, health complaints, and affect can also be found in Table 1, where values represent the overall score on each questionnaire. The majority of participants receiving the intervention were female and typically in their 30s. Based on survey responses, participants also seemed to typically have moderate to high levels of worry, a moderate to low number of subjective health complaints, moderate to high levels of positive affect, and moderate to low levels of negative affect. Additionally, participants believed on average that they had a moderate level of success in registering their worry, whereas they were slightly less confident in their ability to postpone their worry. Table 2 displays the mean, standard deviation, five number summaries, and percent negative—or percent of outcomes below zero—for each of the four continuous worry outcomes of interest (random slopes). The only outcome with the majority of participants having a negative random slope—meaning their worry decreased over time—was nighttime worry frequency, where roughly two thirds of participants experienced symptom improvement. Further inspection of the intervention outcomes reveals that just under half of participants experienced decreases in worry duration during both the day and during the night, and well over half of participants experienced decreases in worry frequency during both the day and night. However, only 6 % experienced decreases in all four outcomes (both daytime/nighttime worry duration/frequency; Versluis et al., 2016). Fig. 1 displays the distributions of the outcomes of interest. Both nighttime worry duration and daytime worry frequency have unimodal and symmetric distributions, while daytime worry duration and nighttime worry frequency were strongly skewed to the left and to the right, respectively.

Table 1
Descriptive statistics of final population of participants considered in analysis.

	Mean/Percent (SD)
Gender	82.82 % female
Age	35.94 (12.90)
Penn State Worry Questionnaire (16–80)	56.13 (11.45)
Subjective Health Complaints (0–29)	9.28 (4.70)
Negative Affect (1–50)	23.71 (8.18)
Positive Affect (1–50)	31.20 (8.16)
Registration (1−10)	6.63 (1.82)
Postponement (1–10)	4.09 (2.52)

Note: Table 1 describes all questionnaire total scores for participants ($N = 163$) included in the machine learning pipeline. Values for each of the three health measures correspond to the total score on each questionnaire (e.g., total worry, total health complaints, total negative/positive affect). Registration = the extent to which participants succeeded in registering worry (1 = 'very bad', 10 = 'very good'); Postponement = the extent to which participants succeeded in postponing worry (1 = 'very bad', 10 = 'very good').

3.2. Daytime worry duration

3.2.1. Predicting linear trend

The results of predictions for measuring change in daytime worry duration over the course of the intervention can be found in Table 3. The

Table 2
Summary statistics of worry outcomes for participants considered in analysis.

	Daytime worry duration	Nighttime worry duration	Daytime worry frequency	Nighttime worry frequency
Mean	~0	~0	~0	~0
Standard deviation	2.64	2.74	0.46	0.17
Minimum	−12.85	−22.34	−2.16	−0.15
First quartile	−0.43	−0.57	−0.17	−0.11
Median	0.73	0.17	~0	−0.05
Third Quartile	1.54	0.66	0.16	0.03
Maximum	2.08	11	2.06	0.86
Percent negative	33.74 %	49.08 %	39.88 %	66.26 %

Note: Outcomes correspond to random slopes of time calculated from linear mixed models. Means for all outcomes were approximately 0. Percent Negative = percent of values below zero for each outcome.

Table 3
Results from continuous outcome prediction.

	Daytime worry duration (validation/test)	Nighttime worry duration (validation/test)	Daytime worry frequency (validation/test)	Nighttime worry frequency (validation/test)
r^2	0.224/0.221	0.025/0.024	0.014/0.041	0.212/0.164
MAE	1.262/1.274	2.008/2.088	0.328/0.219	0.096/0.102
MAE percent improvement	21.13 %/24.87 %	-49.11 %/−28.28 %	−1.16 %/−1.75 %	13.29 %/13.71 %
Spearman ρ	0.567/0.654	0.128/0.109	0.169/0.231	0.512/0.515
Winsorized r^2	0.179/0.292	_	_	0.159/0.207
Winsorized MAE	0.900/0.756	_	_	0.073/0.080
Winsorized MAE percent improvement	16.40 %/31.44 %	_	_	5.21 %/10.36 %

Note: Metrics are reported from ensemble model evaluation on the validation and test sets. MAE = Mean Absolute Error. Winsorized performance metrics for nighttime worry duration and daytime worry frequency are not reported.

ensemble machine learning models for both the validation and test sets yielded r^2 values of approximately 0.22, indicating a moderately predictive relationship (Rice and Harris, 2005). The MAE of both the validation and test set predictions were similar with values of 1.262 (21.13 % improvement vs. naive model) and 1.274 (24.87 % improvement), respectively, implying that the trained model greatly outperformed the naive model for both the validation and test set. The spearman rank correlations for the validation and test sets were 0.567 and 0.654, respectively. The outcomes for daytime worry duration were then winsorized, and each metric was recalculated on the winsorized outcomes. The r^2 values yielded on the validation and test sets were 0.179 and 0.292, respectively; given these values modestly deviated from the non-winsorized r^2 values, this suggests outliers did not heavily influence model performance. The MAE of the winsorized validation set was 0.900 (16.40 % improvement vs. naive model) and 0.756 (31.44 % improvement) for the test set, implying that the trained model greatly outperformed the naive model for both the validation and test set.

Fig. 2A displays the five most important features used to create predictions. Ordered in descending importance, the features were the interaction between the total score on the PSWQ and total number of subjective health complaints ("PSWQ * SHC"), the interaction between alcohol and cigarette consumption ("Alcohol * Smoke"), the total score on the PSWQ ("PSWQ"), the item in the PSWQ measuring the extent to which participants worry under pressure ("Worrying under pressure"), and the interaction of the PANAS items corresponding to inspiration and determination ("Inspired * Determined").

3.2.2. *Predicting improvement or deterioration*

The results for predicting whether participants experienced overall improvement or deterioration in their daytime worry duration over the course of the intervention can be found in Table 4. The ensemble machine learning model for both the validation and test sets yielded overall

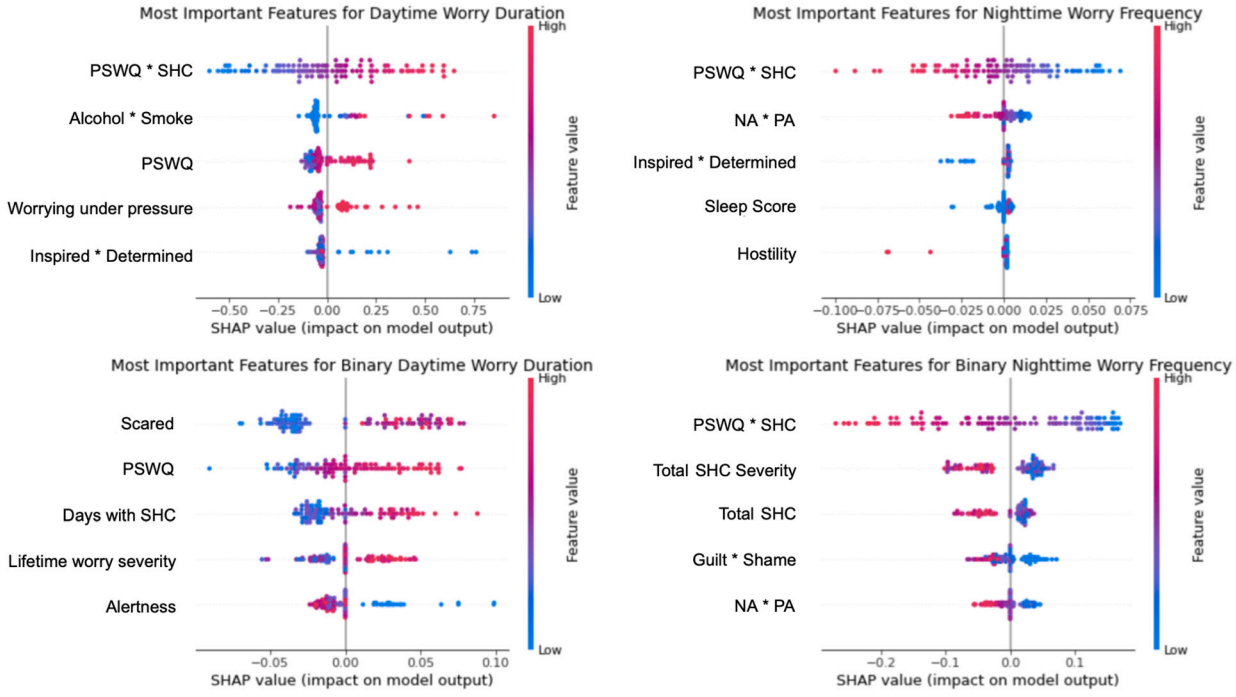

Fig. 2. SHAP Feature Importance.
Note: A positive SHAP value (horizontal axis) indicates that the model predicted a decline (improvement) in worry for the given Feature value, whereas a negative SHAP value indicates that the model predicted an increase (deterioration) in worry. Panel A (top left) shows the relative contributions for each of the top five features in predicting changes in daytime worry duration over time. For this plot, a positive Feature value (red point) for PSWQ * SHC corresponded to positive SHAP values, indicating to the model that the participant should be predicted to have a greater decline in daytime worry duration over the course of the study. Panel B (top right) shows the relative contributions for each of the top five features in predicting changes in nighttime worry frequency over time. Panel C (bottom left) shows the relative contributions for each of the top five features in predicting improvement or deterioration in daytime worry frequency. Panel D (bottom right) shows the relative contributions for each of the top five features in predicting improvement or deterioration in nighttime worry frequency. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Results from binary outcome prediction.

	Daytime worry duration (validation/test)	Nighttime worry duration (validation/test)	Daytime worry frequency (validation/test)	Nighttime worry frequency (validation/test)
Accuracy	0.74/0.76	0.53/0.55	0.55/0.53	0.74/0.69
Specificity	0.75/0.78	0.53/0.62	0.33/0.35	0.58/0.42
Sensitivity	0.71/0.69	0.52/0.48	0.77/0.74	0.83/0.78
AUC	0.77/0.77	0.59/0.54	0.54/0.55	0.79/0.72

Note: Metrics are reported from ensemble model evaluation on the validation and test sets. Sensitivity = Recall. AUC = Area Under the Curve.

accuracy scores of 0.74 and 0.76, respectively. The specificity on the validation and test sets was 0.75 and 0.78, respectively, and the sensitivity on the validation and test sets was 0.71 and 0.69, respectively. Additionally, the AUC on the validation and test sets were both 0.77, indicating a moderately predictive relationship (Rice and Harris, 2005).

Fig. 2C displays the five most important features used to create predictions. Ordered in descending importance, the features are the PANAS item corresponding to how scared one feels ("Scared"), the total score on the PSWQ ("PSWQ"), the sum of the number of days each subjective health complaint listed on the SHC inventory was experienced ("Days with SHC"), the item on the PSWQ measuring the extent to which participants consider themselves lifetime worriers ("Lifetime worry severity"), and the PANAS item assessing alertness ("Alertness").

3.3. Nighttime worry duration and daytime worry frequency

3.3.1. Predicting linear trend

The results of predictions for measuring change in nighttime worry duration and daytime worry frequency over the course of the intervention can be found in Table 3. For both outcomes, the models performed poorly on both the validation and test sets (e.g., r^2 values indicate a small predictive relationship; Rice and Harris, 2005), and winsorized performance metrics and feature importance are not reported as a result. This indicates there was likely no signal in the data to inform accurate predictions for these items.

3.3.2. Predicting improvement or deterioration

The results for predicting whether participants experienced overall improvement or deterioration in their nighttime worry duration and daytime worry frequency over the course of the intervention can be found in Table 4. Similar to the corresponding continuous outcomes, for both of these outcomes, the models performed poorly on both the validation and test sets (e.g., AUC values indicate a small predictive relationship; Rice and Harris, 2005), and feature importance is not reported as a result.

3.4. Nighttime worry frequency

3.4.1. Predicting linear trend

The results of predictions for measuring change in nighttime worry frequency over the course of the intervention can be found in Table 3. The ensemble machine learning models for the validation and test sets yielded r^2 values of approximately 0.212 and 0.164, respectively, indicating a moderately predictive relationship (Rice and Harris, 2005). The MAE of both the validation and test set predictions were similar with values of 0.096 (13.29 % improvement vs. naive model) and 0.102 (13.71 % improvement), respectively, implying that the trained model outperformed the naive model for both the validation and test set. The spearman rank correlations for the validation and test sets are 0.512 and 0.515, respectively. The outcomes for nighttime worry frequency were then winsorized, and each metric was recalculated on the winsorized outcomes. The r^2 values yielded on the validation and test sets were 0.159 and 0.207, respectively; given these values modestly deviated from the non-winsorized r^2 values, this suggests outliers did not heavily influence model performance. The MAE of the winsorized validation set was 0.073 (5.21 % improvement from naive model) and 0.080 (10.36 % improvement) for the test set, implying that the trained model outperformed the naive model for both the validation and test set.

Fig. 2B displays the five most important features used to create predictions. Ordered in descending importance, the features are the interaction between the total score for the PSWQ and total number of subjective health complaints ("PSWQ * SHC"), the interaction between total negative affect and total positive affect as assessed by the PANAS ("NA * PA"), the product of the PANAS items corresponding to inspiration and determination ("Inspired * Determined"), the interaction of sleep duration and sleep quality divided by the sum of the SHC inventory items corresponding to sleep problems and tiredness ("Sleep Score"), and the PANAS item corresponding to hostility ("Hostility").

3.4.2. Predicting improvement or deterioration

The results for predicting whether participants experienced overall improvement or deterioration in nighttime worry frequency over the course of the intervention can be found in Table 4. The ensemble machine learning models for both the validation and test sets yielded overall accuracy scores of 0.74 and 0.69, respectively. The specificity on the validation and test sets was 0.58 and 0.42, respectively, and the sensitivity on the validation and test sets was 0.83 and 0.78, respectively. Additionally, the AUC on the validation and test sets were 0.79 and 0.72, respectively, indicating a moderately predictive relationship (Rice and Harris, 2005).

Fig. 2D displays the five most important features used to create predictions. Ordered in descending importance, the features are the interaction between the total score for the PSWQ and total number of subjective health complaints ("PSWQ * SHC"), the total severity of all subjective health complaints assessed by the SHC inventory ("Total SHC Severity"), the total number of subjective health complaints experienced ("SHC Total"), the interaction of the PANAS items assessing guilt and shame ("Guilt * Shame"), and the interaction between total negative affect and total positive affect as assessed by the PANAS ("NA * PA").

4. Discussion

The goal of the present study was to determine how well the outcomes of a worry postponement intervention could be predicted with machine learning methodology, as well as to ascertain which baseline features offer the most predictive importance. This modeling approach would elucidate whether this simple, non-intrusive intervention could be effective in decreasing the worry of a specific individual. Predictions were also evaluated using an independent test set of unseen participants to avoid model overfitting and obtain performance estimates more realistic to a real-world setting (Hornstein et al., 2021). The present study expounds on the work of a prior study (Versluis et al., 2016), one of many recent investigations conducted to assess the efficacy of various methods of personalized mental health interventions (Delgado et al., 2010; McCarrick et al., 2021; Versluis et al., 2016). In an attempt to promote personalizable mental healthcare, numerous recent studies have leveraged machine learning models to predict outcomes (e.g., treatment response) for a given individual in response to a treatment (Chekroud et al., 2016; Hahn et al., 2015), particularly ones that are digitally-administered (Hornstein et al., 2021; Jacobson and Nemesure, 2021; Meinlschmidt et al., 2020). However, an issue with supervised machine learning models used in these contexts is the "black box" concept—it is unclear how features are driving model predictions (Perna et al., 2018). Understanding how varying levels of a specific baseline feature can influence predicted treatment response is paramount to appropriately prescribing an intervention for specific individuals. The present study thereby contributes to the growing literature on personalizable mental health by (1) further analyzing a specific form of

personalized mental health care via a unique ensemble-based machine learning approach and (2) leveraging individualized data and examining feature importance via SHAP to reveal what baseline characteristics may predispose individuals to benefitting from interventions of this type.

The first takeaway from the present modeling approach was that changes in daytime worry duration and nighttime worry frequency were able to be predicted with moderate accuracy, whereas changes in daytime worry frequency and nighttime worry duration were unable to be predicted with any confidence. The model performance on daytime worry duration and nighttime worry frequency was in line with the performance of other works applying machine learning models to predict digital treatment response (Jacobson and Nemesure, 2021). The fact that accurate predictions could not be obtained for nighttime worry duration and daytime worry frequency is a finding in itself, as this potentially comments on the complexity of how much worry occurs at night and how often worry occurs during the day. This is an interesting finding given the nature of worry and how it manifests at different points in the day. Research has shown that worse sleep leads to increased worry the following day; however, increased daytime worry has a weaker association with nighttime sleep. Essentially, this indicates a differential bidirectional relationship where daytime worry affects sleep differently than sleep impacts daytime worry (Narmandakh et al., 2021). With this in mind, it makes sense that the model was differentially predicting aspects of worry for the different parts of the day.

The second takeaway from the present analyses is the baseline features that were particularly important in driving model predictions for changes in worry. One of the most important features across all models was the interaction between total worry severity and subjective health complaints ("PSWQ * SHC"). Of the other most important features, many included some quantity measured by the PSWQ or SHC inventory, such as the total PSWQ and SHC scores on their own and the number of days where SHC were experienced. Other features with strong importance in multiple models were the product of negative affect and positive affect and the product of the PANAS items measuring inspiration and determination. It is interesting to consider how features impacted predictions differently across models. For instance, high values of the interaction of total worry severity and SHC corresponded to a positive impact on daytime worry duration model output (shown in Fig. 2A), implying that people with higher baseline levels of worry and/or SHC were predicted to see improvements in daytime worry duration over the course of the intervention, whereas people with lower values of worry and/or SHC were predicted to see deteriorations in daytime worry duration. Conversely, high values of the interaction of total worry severity and SHC corresponded to a negative impact on nighttime worry frequency model output (shown in Fig. 2B), implying that people with higher baseline levels of worry and/or SHC were predicted to see increases in nightly worry frequency over the course of the intervention. These trends were seen for other features related to PSWQ or SHC across all models, where higher baseline values of worry and/or SHC were attributed to predicted decreases in daytime worry duration and predicted increases in nighttime worry frequency.

Why is it that (1) higher levels of worry and/or SHC corresponded to a predicted reduction in daytime worry duration but a predicted increase in nighttime worry frequency, and that (2) lower levels of worry and/or SHC corresponded to a predicted increase in daytime worry duration but a predicted decrease in nighttime worry frequency? The associations in (1) may be partly explained by the relationship between sleep and worry. The design of the present worry postponement intervention is for participants to postpone their worry to the end of the day; thus, people with initially high worry and/or SHC will attempt to postpone their worry and will likely experience an influx at the end of the day. Worry experienced prior to sleeping is strongly predictive of the number of nighttime awakenings (McGowan et al., 2016; Thielsch et al., 2015). Additionally, subjective health complaints have shown to be associated with poor sleep quality (Thormar et al., 2014). Importantly, 50–70 % of people with GAD also have insomnia (McGowan et al., 2016); therefore, excessive worry before bed may exacerbate existing worry-induced sleep issues related to worry. People that wake up more at night are likely to have a greater frequency of worry given that they will have greater opportunities for worrying. Thus, because the present study instructs participants to postpone worry until the end of the day, people with initially high worry and SHC will see declines in daytime worry duration (because they are postponing their worry during the day) but increases in nighttime worry frequency (because of increased worry before bedtime). Data from the present study supports this conclusion: of participants with the highest 25 % of values for the interaction of the PSWQ and SHC inventory scores, 63 % experienced deterioration in nighttime worry frequency, whereas 68 % saw improvements in daytime worry duration (Versluis et al., 2016).

On the other hand, the observed association described in (2) may have occurred because participants with low worry and/or SHC were forced to attend to their worry more than usual, in that they had to consciously postpone their minor worries, driving up their overall worry duration throughout the day. Conversely, these individuals potentially experienced improvements in nighttime worry frequency because they were likely waking up less at night to worry given their initial low levels of worry and SHC. Even though these individuals were postponing their worry until the evening, because they had such low worry to begin with, it is possible that there was little to no impact on nighttime awakenings and thus nighttime worry frequency. Data from the present study supports this conclusion: of the participants in the lowest 25 % of values for the interaction of the PSWQ and SHC inventory scores, 95 % experienced improvements in nighttime worry frequency, whereas 93 % experienced deterioration in daytime worry duration (Versluis et al., 2016).

Beyond the implications derived from understanding the driving factors in model prediction, the predictive model itself may be a useful framework in clinical applications. High sensitivity was obtained for predicting change in daytime worry duration and nighttime worry frequency (see Table 4), indicating that improvements in these types of worry were predicted with confidence. Along with this, while the feature importance plots in Fig. 2 do not directly identify subsets of individuals that would respond well to this intervention, the feature importance values in Fig. 2 suggest that individuals with higher baseline worry but lower baseline subjective health complaints may be most likely to benefit from a worry postponement intervention of this type. Thus, given a scenario where participant baseline information used in model training is easily obtainable (e.g., via individual characteristics or publicly available questionnaires), the present results would allow for people to ascertain whether a worry postponement intervention of this type could ameliorate their worry prior to beginning the treatment. All in all, the present study further lays the foundation for personalized medicine, identifying individuals which may benefit the most from this low-cost, low-burden intervention designed to reduce worry and providing an analytical framework that can be applied to other interventions.

When considering the present results, it is also worth considering their limitations. First, there may have been unmeasured confounding participant-level factors that impacted—in positive and negative ways—participants' abilities to address their worry at different times of day throughout the worry postponement intervention. Worry is a very unique experience to the individual, and the present analyses could not capture everything related to an individual's worry and their ability to complete the intervention. Along with unmeasured confounding factors, the present results may not be generalizable for the broader population because the majority of study participants were both Dutch and female, the latter being a subgroup that is more likely to live with conditions such as GAD (Grenier et al., 2019; Haller et al., 2014). Furthermore, the present study was a complete case analysis, and only participants who completed the worry postponement intervention were considered. Sixty seven participants discontinued the intervention at some point over its course (Versluis et al., 2016), and by dropping these participants some

important information regarding the efficacy of the worry postponement intervention may have been lost. The present study could have instead been run as an intent-to-treat analysis, and methods such as multiple imputation could have been implemented to include the participants with missing data. However, due to the excessive missingness in the present study, it was deemed infeasible to adequately perform multiple imputation on all participants who did not complete the study. A related limitation of the current study was the small sample size ($N = 163$) relative to the number of features, which may mean that the current study was underpowered and generalizability to other samples may be limited. Lastly, duration and frequency of worry episodes were recorded at the end of the day throughout the study period (Versluis et al., 2016), and therefore there is a chance that recorded values were vulnerable to recall bias. The original study also did not record any functional outcome measures corresponding to changes in worry.

The present analyses build a supervised machine learning framework for personalized prediction of treatment response to a simple worry postponement intervention. While this endeavor provides great clinical merit as aforementioned, future research may seek to develop an unsupervised machine learning approach (e.g., clustering) in an effort to identify subsets of individuals within the population that responded well to this intervention as opposed to predicting individual outcomes. Along with this, while the present study considers only short-term changes in worry duration and frequency, it is important to note that worry is a dynamic process that lasts beyond the duration of this study. Given this, future studies of worry postponement interventions may 1) want to consider implementing a longer intervention and/or 2) track participant levels of worry after completing the intervention (Hirsch et al., 2021). Either approach would augment the understanding of the long-term dynamics of worry. Furthermore, given that worry may have a varied impact throughout the day (McGowan et al., 2016), future research may want to investigate implementing a worry postponement intervention at different times of the day. Additionally, the present analyses did not examine day-to-day changes in worry across the intervention, but only overall changes. Worry is volatile and influenced by many factors: for instance, worry before sleep increases sleep disturbances, which in turn increases nighttime worry and is predictive of increased worry the following day (Thielsch et al., 2015). Thus, examining short-term changes in worry may reveal more about worry dynamics on a daily basis. Each of these future investigations would build on the framework developed in this study and further our understanding of worry dynamics and personalized medicine.

Data statement

All data analyzed during the present analyses were obtained from Versluis et al. (2016). The variables and relationships examined in the present article have not been examined in any previous or current articles, or to the best of our knowledge in any papers that will be under review soon. The datasets generated during the current study are available from the corresponding author on reasonable request.

Author statement

Funding for this article was provided by the Kaminsky Undergraduate Research Award. Additionally, this research is supported in part by T32 (T32DA037202-07) and P30 (P30DA029926) grants provided by the National Institute of Drug Abuse. All data and code are made available upon author request. Acknowledgements: none.

Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

The authors would like to acknlowledge the support of the AIM HIGH Lab at the Geisel School of Medicine at Dartmouth College, their families, and all others who provided support in the preparation of this work. The authors also thank the authors of the original study (Versluis et al., 2015) for their permissions and invaluable contributions to the literature.

Appendix A. Supplementary analysis

A supplementary analysis to this article can be found online at https://doi.org/10.1016/j.jad.2022.09.112.

References

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1) https://doi.org/10.18637/jss.v067.i01.

Brosschot, J.F., van der Doef, M., 2006. Daily worrying and somatic health complaints: testing the effectiveness of a simple worry reduction intervention. Psychol. Health 21 (1), 19–31. https://doi.org/10.1080/14768320500105346.

Butryn, T., Bryant, L., Marchionni, C., Sholevar, F., 2017. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. International journal ofAcad. Med. 3 (1) https://doi.org/10.4103/IJAM.IJAM_49_17.

Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R., 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 3 (3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Cho, G., Yim, J., Choi, Y., Ko, J., Lee, S.-H., 2019. Review of machine learning algorithms for diagnosing mental illness. Psychiatry Investig. 16 (4), 262–269. https://doi.org/10.30773/pi.2018.12.21.2.

Christensen, H., Griffiths, K.M., Farrer, L., 2009. Adherence in internet interventions for anxiety and depression. J. Med. Internet Res. 11 (2), e13 https://doi.org/10.2196/jmir.1194.

Coplan, J.D., Aaronson, C.J., Panthangi, V., Kim, Y., 2015. Treating comorbid anxiety and depression: psychosocial and pharmacological approaches. World J. Psychiatry 5 (4), 366–378. https://doi.org/10.5498/wjp.v5.i4.366.

Crawford, J.R., Henry, J.D., 2004. The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. Br. J. Clin. Psychol. 43 (3), 245–265. https://doi.org/10.1348/0144665031752934.

Delgado, L.C., Guerra, P., Perakakis, P., Vera, M.N., Reyes del Paso, G., Vila, J., 2010. Treating chronic worry: psychological and physiological effects of a training programme based on mindfulness. Behav. Res. Ther. 48 (9), 873–882. https://doi.org/10.1016/j.brat.2010.05.012.

DeRubeis, R.J., Cohen, Z.D., Forand, N.R., Fournier, J.C., Gelfand, L.A., Lorenzo-Luaces, L., 2014. The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. PLoS ONE 9 (1), e83875. https://doi.org/10.1371/journal.pone.0083875.

Eriksen, H.R., Ihlebæk, C., Ursin, H., 1999. A scoring system for subjective health complaints (SHC). Scand. J. Public Health 27 (1), 63–72. https://doi.org/10.1177/14034948990270010401.

Gonzalez, O., 2021. Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. Psychol. Methods 26 (2), 236–254. https://doi.org/10.1037/met0000317.

Goodwin, H., Yiend, J., Hirsch, C.R., 2017. Generalized anxiety disorder, worry and attention to threat: a systematic review. Clin. Psychol. Rev. 54, 107–122. https://doi.org/10.1016/j.cpr.2017.03.006.

Grenier, S., Payette, M., Gunther, B., Askari, S., Desjardins, F.F., Raymond, B., Berbiche, D., 2019. Association of age and gender with anxiety disorders in older adults: a systematic review and meta-analysis. Int. J. Geriatr. Psychiatry 34 (3), 397–407. https://doi.org/10.1002/gps.5035.

Hahn, T., Kircher, T., Straube, B., Wittchen, H.-U., Konrad, C., Ströhle, A., Wittmann, A., Pfleiderer, B., Reif, A., Arolt, V., Lueken, U., 2015. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiatry 72 (1), 68. https://doi.org/10.1001/jamapsychiatry.2014.1741.

Haller, H., Cramer, H., Lauche, R., Gass, F., Dobos, G.J., 2014. The prevalence and burden of subthreshold generalized anxiety disorder: a systematic review. BMC Psychiatry 14 (1), 128. https://doi.org/10.1186/1471-244X-14-128.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hayes, S., Hirsch, C., Mathews, A., 2008. Restriction of working memory capacity during worry. J. Abnorm. Psychol. 117 (3), 712–717. https://doi.org/10.1037/a0012908.

Hirsch, C.R., Krahé, C., Whyte, J., Krzyzanowski, H., Meeten, F., Norton, S., Mathews, A., 2021. Internet-delivered interpretation training reduces worry and anxiety in individuals with generalized anxiety disorder: a randomized controlled experiment. J. Consult. Clin. Psychol. 89 (7), 575–589. https://doi.org/10.1037/ccp0000660.
Hoo, K.A., Tvarlapati, K.J., Piovoso, M.J., Hajare, R., 2002. A method of robust multivariate outlier replacement. Comput. Chem. Eng. 26 (1), 17–39. https://doi.org/10.1016/S0098-1354(01)00734-7.
Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K., Hilbert, K., 2021. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. Dig. Health 7, 205520762110606. https://doi.org/10.1177/20552076211060659.
Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21 (8), 1509–1515. https://doi.org/10.1093/bioinformatics/bti171.
Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9 (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
Jacobson, N.C., Nemesure, M.D., 2021. Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: evidence from a transdiagnostic randomized controlled trial. Psychiatry Res. 295, 113618 https://doi.org/10.1016/j.psychres.2020.113618.
Jacobson, N.C., Newman, M.G., 2017. Anxiety and depression as bidirectional risk factors for one another: a meta-analysis of longitudinal studies. Psychol. Bull. 143 (11), 1155–1200. https://doi.org/10.1037/bul0000111.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Proces. Syst. 30, 3146–3154.
Kraemer, H.C., Thiemann, S., 1989. A strategy to use soft data effectively in randomized controlled clinical trials. J. Consult. Clin. Psychol. 57 (1), 148–154. https://doi.org/10.1037/0022-006X.57.1.148.
Ku, B.S., Li, J., Lally, Cathy, Compton, M.T., Druss, B.G., 2021. Associations between mental health shortage areas and county-level suicide rates among adults aged 25 and older in the USA, 2010 to 2018. Gen. Hosp. Psychiatry 70, 44–50. https://doi.org/10.1016/j.genhosppsych.2021.02.001.
Lekkas, D., Klein, R.J., Jacobson, N.C., 2021. Predicting acute suicidal ideation on instagram using ensemble machine learning models. Internet Interv. 25, 100424 https://doi.org/10.1016/j.invent.2021.100424.
Lekkas, D., Price, G., McFadden, J., Jacobson, N.C., 2021. The application of machine learning to online mindfulness intervention data: a primer and empirical example in compliance assessment. Mindfulness 12 (10), 2519–2534. https://doi.org/10.1007/s12671-021-01723-4.
Lekkas, D., Price, G.D., Jacobson, N.C., 2022. Using smartphone app use and lagged-ensemble machine learning for the prediction of work fatigue and boredom. Comput. Hum. Behav. 127, 107029 https://doi.org/10.1016/j.chb.2021.107029.
Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. ArXiv:1705.07874 [Cs, Stat]. http://arxiv.org/abs/1705.07874.
McCarrick, D., Prestwich, A., Prudenzi, A., O'Connor, D.B., 2021. Health effects of psychological interventions for worry and rumination: a meta-analysis. Health Psychol. https://doi.org/10.1037/hea0000985.
McGowan, S.K., Behar, E., Luhmann, M., 2016. Examining the relationship between worry and sleep: a daily process approach. Behav. Ther. 47 (4), 460–473. https://doi.org/10.1016/j.beth.2015.12.003.
Meinlschmidt, G., Tegethoff, M., Belardi, A., Stalujanis, E., Oh, M., Jung, E.K., Kim, H.-C., Yoo, S.-S., Lee, J.-H., 2020. Personalized prediction of smartphone-based psychotherapeutic micro-intervention success using machine learning. J. Affect. Disord. 264, 430–437. https://doi.org/10.1016/j.jad.2019.11.071.
Merikangas, K.R., He, J., Burstein, M., Swanson, S.A., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., Swendsen, J., 2010. Lifetime prevalence of mental disorders in U.S. Adolescents: results from the National Comorbidity Survey Replication-Adolescent Supplement (NCS-A). J. Am. Acad. Child Adolesc. Psychiatry 49 (10), 980–989. https://doi.org/10.1016/j.jaac.2010.05.017.
Meyer, T.J., Miller, M.L., Metzger, R.L., Borkovec, T.D., 1990. Development and validation of the penn state worry questionnaire. Behav. Res. Ther. 28 (6), 487–495. https://doi.org/10.1016/0005-7967(90)90135-6.
Narmandakh, A., Oldehinkel, A.J., Masselink, M., de Jonge, P., Roest, A.M., 2021. Affect, worry, and sleep: between- and within-subject associations in a diary study. J. Affect. Disord. Rep. 4, 100134 https://doi.org/10.1016/j.jadr.2021.100134.
Nemesure, M.D., Heinz, M.V., Huang, R., Jacobson, N.C., 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci. Rep. 11 (1), 1980. https://doi.org/10.1038/s41598-021-81368-4.
Olatunji, B.O., Wolitzky-Taylor, K.B., Sawchuk, C.N., Ciesielski, B.G., 2010. Worry and the anxiety disorders: a meta-analytic synthesis of specificity to GAD. Appl. Prev. Psychol. 14 (1–4), 1–24. https://doi.org/10.1016/j.appsy.2011.03.001.
Paxling, B., Almlöv, J., Dahlin, M., Carlbring, P., Breitholtz, E., Eriksson, T., Andersson, G., 2011. Guided internet-delivered cognitive behavior therapy for generalized anxiety disorder: a randomized controlled trial. Cogn. Behav. Ther. 40 (3), 159–173. https://doi.org/10.1080/16506073.2011.576699.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2018. Scikit-learn: Machine Learning in Python. ArXiv:1201.0490 [Cs]. http://arxiv.org/abs/1201.0490.
Pelletier, L., O'Donnell, S., McRae, L., Grenier, J., 2017. The burden of generalized anxiety disorder in Canada. Health Promot. Chronic Dis. Prev. Can. 37 (2), 54–62. https://doi.org/10.24095/hpcdp.37.2.04.
Perna, G., Grassi, M., Caldirola, D., Nemeroff, C.B., 2018. The revolution of personalized psychiatry: will technology make it happen sooner? Psychol. Med. 48 (5), 705–713. https://doi.org/10.1017/S0033291717002859.
Querstret, D., Cropley, M., 2013. Assessing treatments used to reduce rumination and/or worry: a systematic review. Clin. Psychol. Rev. 33 (8), 996–1009. https://doi.org/10.1016/j.cpr.2013.08.004.
Reback, J., Jbrockmendel, McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Hawkins, S., Gfyoung, Roeschke, M., Sinhrks, Klein, Petersen, A.Terji, Tratner, J., She, C., Ayd, W., Hoefler, P., Naveh, S., Garcia, M., Schendel, J., Seabold, Skipper, 2021. pandas-dev/pandas: Pandas 1.3.3 (v1.3.3). Zenodo. https://doi.org/10.5281/ZENODO.3509134.
Rice, M.E., Harris, G.T., 2005. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law Hum. Behav. 29 (5), 615–620. https://doi.org/10.1007/s10979-005-6832-7.
Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33 (1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7.
Ruscio, A.M., Hallion, L.S., Lim, C.C.W., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L.H., Borges, G., Bromet, E.J., Bunting, B., Caldas de Almeida, J.M., Demyttenaere, K., Florescu, S., de Girolamo, G., Gureje, O., Haro, J.M., He, Y., Hinkov, H., Hu, C., Scott, K.M., 2017. Cross-sectional comparison of the epidemiology of DSM-5 generalized anxiety disorder across the globe. JAMA Psychiatry 74 (5), 465. https://doi.org/10.1001/jamapsychiatry.2017.0056.
Stein, M.B., Sareen, J., 2015. Generalized anxiety disorder. N. Engl. J. Med. 373 (21), 2059–2068. https://doi.org/10.1056/NEJMcp1502514.
Suresh, H., Hunt, N., Johnson, A., Celi, L.A., Szolovits, P., Ghassemi, M., 2017. Clinical Intervention Prediction and Understanding Using Deep Networks. ArXiv:1705.08498 [Cs]. http://arxiv.org/abs/1705.08498.
Thielsch, C., Ehring, T., Nestler, S., Wolters, J., Kopei, I., Rist, F., Gerlach, A.L., Andor, T., 2015. Metacognitions, worry and sleep in everyday life: studying bidirectional pathways using ecological momentary assessment in GAD patients. J. Anxiety Disord. 33, 53–61. https://doi.org/10.1016/j.janxdis.2015.04.007.
Thompson, A., Issakidis, C., Hunt, C., 2008. Delay to seek treatment for anxiety and mood disorders in an Australian clinical sample. Behav. Chang. 25 (2), 71–84. https://doi.org/10.1375/bech.25.2.71.
Thormar, S.B., Gersons, B.P.R., Juen, B., Djakababa, M.N., Karlsson, T., Olff, M., 2014. The impact of disaster work on community volunteers: the role of peri-traumatic distress, level of personal affectedness, sleep quality and resource loss, on post-traumatic stress disorder symptoms and subjective health. J. Anxiety Disord. 28 (8), 971–977. https://doi.org/10.1016/j.janxdis.2014.10.006.
Tiirikainen, K., Haravuori, H., Ranta, K., Kaltiala-Heino, R., Marttunen, M., 2019. Psychometric properties of the 7-item generalized anxiety disorder scale (GAD-7) in a large representative sample of Finnish adolescents. Psychiatry Res. 272, 30–35. https://doi.org/10.1016/j.psychres.2018.12.004.
Verkuil, B., Brownlow, B.N., Vasey, M.W., Brosschot, J.F., Thayer, J.F., 2021. A brief scale of pathological worry that everyone already has. Curr. Psychol. https://doi.org/10.1007/s12144-021-01603-z.
Versluis, A., Verkuil, B., Brosschot, J.F., 2016. Reducing worry and subjective health complaints: a randomized trial of an internet-delivered worry postponement intervention. Br. J. Health Psychol. 21 (2), 318–335. https://doi.org/10.1111/bjhp.12170.
Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., van Mulbregt, P., 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods 17 (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
Waskom, M., 2021. Seaborn: statistical data visualization. J. Open Source Softw. 6 (60), 3021. https://doi.org/10.21105/joss.03021.
Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Soc. Psychol. 54 (6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063.
Wilhelm, S., Weingarden, H., Ladis, I., Braddick, V., Shin, J., Jacobson, N.C., 2020. Cognitive-behavioral therapy in the digital age: presidential address. Behav. Ther. 51 (1), 1–14. https://doi.org/10.1016/j.beth.2019.08.001.
Zilcha-Mano, S., Roose, S.P., Brown, P.J., Rutherford, B.R., 2018. A machine learning approach to identifying placebo responders in late-life depression trials. Am. J. Geriatr. Psychiatry 26 (6), 669–677. https://doi.org/10.1016/j.jagp.2018.01.001.