EUROPEAN JOURNAL OF PSYCHOTRAUMATOLOGY
2025, VOL. 16, NO. 1, 2472485
https://doi.org/10.1080/20008066.2025.2472485

CINICAL RESEARCH ARTICLE

OPEN ACCESS Check for updates

Predicting high anger intensity using ecological momentary assessment and wearable-derived physiological data in a trauma-affected sample

Olivia Metcalf [a,b], Karen E. Lamb[c,d], David Forbes [a], Meaghan L. O'Donnell[a], Tianchen Qian[e], Tracey Varker[a], Sean Cowlishaw[f] and Sophie Zaloumis[c,d]

[a]Phoenix Australia – Centre for Posttraumatic Mental Health, Department of Psychiatry, University of Melbourne, Carlton, Australia; [b]Centre for Digital Transformation of Health, University of Melbourne, Carlton, Australia; [c]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Carlton, Australia; [d]MISCH (Methods and Implementation Support for Clinical Health) Research Hub, Faculty of Medicine, Dentistry, and Health Sciences, University of Melbourne, Carlton, Australia; [e]Department of Statistics, University of California, Irvine, Irvine, CA, USA; [f]Turner Institute for Brain and Mental Health, Monash School of Psychological Sciences, Monash University, Clayton, Australia

ABSTRACT

Background: Digital technologies offer tremendous potential to predict dysregulated mood and behavior within an individual's environment, and in doing so can support the development of new digital health interventions. However, no prediction models have been built in trauma-exposed populations that leverage real-world data.

Objective: This project aimed to determine if wearable-derived physiological data can predict anger intensity in trauma-exposed adults.

Method: Heart rate variability (i.e. a commercial wearable stress score) was combined with ecological momentary assessment (EMA) data collected over 10 days ($n = 84$). Five summary measures from stress scores collected 10 min prior to each EMA were selected using factor analysis of 24 candidates.

Results: A high area under the receiver operating curve (AUC) was found for a logistic mixed effects model including these measures as predictors, ranging 0.761 (95% CI:0.569–0.921) to 0.899 (95% CI:0.784–0.980) across cross-validation methods.

Conclusions: While the predictive performance may be overly optimistic due to the outcome prevalence (13.8%) and requires replication with larger datasets, our promising findings have significant methodological and clinical implications for researchers looking to build novel prediction and treatment approaches to respond to posttraumatic mental health.

Predicción de ira de alta intensidad mediante la evaluación momentánea ecológica y datos fisiológicos derivados de dispositivos portátiles en una muestra afectada por trauma

Antecedentes: Las tecnologías digitales ofrecen un enorme potencial para predecir el estado de ánimo y comportamiento desregulado dentro del entorno de un individuo y, al hacerlo, pueden apoyar el desarrollo de nuevas intervenciones de salud digital. Sin embargo, no se han construido modelos de predicción en poblaciones expuestas a trauma que aprovechen datos del mundo real.

Objetivo: Este proyecto tuvo como objetivo determinar si los datos fisiológicos derivados de dispositivos portátiles pueden predecir la intensidad de la ira en adultos expuestos a trauma.

Método: La variabilidad de la frecuencia cardíaca (es decir, una puntuación de estrés portátil comercial) se combinó con datos de evaluación momentánea ecológica (EMA, por sus siglas en inglés) recopilados durante 10 días ($n = 84$). Se seleccionaron cinco medidas de resumen de las puntuaciones de estrés recopiladas 10 minutos antes de cada EMA mediante el análisis factorial de 24 candidatos.

Resultados: Se encontró un área bajo la curva operativa del receptor (AUC, por sus siglas en inglés) alta para un modelo logístico de efectos mixtos que incluye estas medidas como predictores, con un rango de 0.761 (IC del 95%: 0.569–0.921) a 0.899 (IC del 95%: 0.784–0.980) en los métodos de validación cruzada.

Conclusiones: Si bien el desempeño predictivo puede ser demasiado optimista debido a la prevalencia resultante (13.8%) y requiere replicación con conjuntos de datos más grandes, nuestros hallazgos prometedores tienen implicaciones metodológicas y clínicas significativas para los investigadores que buscan desarrollar nuevos enfoques de predicción y tratamiento para responder a la salud mental postraumática.

ARTICLE HISTORY

Received 18 December 2024
Revised 31 January 2025
Accepted 18 February 2025

KEYWORDS

Anger; trauma; HRV; physiological data; digital phenotyping; machine learning; prediction modeling

PALABRAS CLAVE

Ira; trauma; predicción; datos fisiológicos; fenotipado digital

HIGHLIGHTS

- Using smartphone and wearable data derived from 98 adults with trauma exposure, this study was able to predict high anger intensity using heart rate variability-based data from the 10 min prior.
- High anger intensity episodes occurred 14%, and this prevalence rate may affect the performance of the prediction model.
- Real world prediction tools remain challenging due to data quality issues.

CONTACT Olivia Metcalf olivia.metcalf@unimelb.edu.au Phoenix Australia – Centre for Posttraumatic Mental Health, Department of Psychiatry, University of Melbourne, Carlton, Victoria, Australia; Centre for Digital Transformation of Health, University of Melbourne, Level 3 Alan Gilbert Building, 161 Barry Street, Carlton, Victoria, Australia

Supplemental data for this article can be accessed online at https://doi.org/10.1080/20008066.2025.2472485

Globally, as climate change intensifies occurrence of natural disasters, health systems adjust to the aftermath of the pandemic, and international conflicts continue, it is increasingly recognized how critical it is to respond effectively to the aftermaths of trauma. Emotion-focused models of posttraumatic mental health posit that emotions that were adaptive during the traumatic event, such as fear and anger, can become dysregulated and lead to the development of mood disorders (Wastell, 2020). Dysregulated anger, termed problem anger, is characterized by rapid, frequent, and disproportionate escalations in anger that harms individuals and their relationships (Forbes et al., 2014). A growing body of evidence indicates that post-trauma anger is a common mental health issue, affecting up to 31% of veteran populations (Varker, Cowlishaw, et al., 2022), and in some studies, occurs more frequently in women, affecting up to 1 in 5 new mothers, and affecting more women than men after natural disasters (Adler et al., 2020; Cowlishaw et al., 2021; Plummer Lee et al., 2024; Varker, Cowlishaw, et al., 2022). Problem anger is linked to severe harms, including suicidality (Varker, Cowlishaw, et al., 2022), cardiovascular disease (Chida & Steptoe, 2009), dangerous and illegal driving (Zhang & Chan, 2016), and some types of relationship violence (Eckhardt et al., 2008). Despite the frequency and consequences of problem anger after trauma, there is significantly less research into problem anger compared to internalizing disorders, such as depression and anxiety.

An emerging body of work has focused on how problem anger manifests, leveraging ecological momentary assessment (EMA) to map anger symptoms over time (Arjmand et al., 2023; Metcalf et al., 2021). EMA is an increasingly common method used in psychological research, that involves regular prompting of an individual to self-report their cognitive, behavior, and affective states, to understand the dynamic nature of these psychological sequelae at moment-to-moment timescales (Trull & Ebner-Priemer, 2009). Advances in digital tools such as wearables have created opportunities to pair EMA with objective measures of physiological data. Such an approach enables the investigation of the prediction of affective and behavior disturbances post-trauma, with the ultimate goal of developing technology-based interventions that leverage hardware such as smartphones and wearables with evidence-based components, to provide just-in-time support. While leveraging smartphone and sensor data have tremendous promise in psychiatry (Hickey et al., 2021; Huhn et al., 2022; Seppälä et al., 2019), very few studies have built real-world prediction models of affective states and behavior. This nascent field has some preliminary research showing potential. In 2019, a seminal study investigated the potential for physiological data derived via a researcher-grade (as opposed to consumer-grade) wearable devices in an inpatient psychiatry setting in predominately non-verbal youths with autism (Goodwin et al., 2019). The subsequent prediction models found that aggression could be reliably predicted one minute before it occurs, using the three minutes prior biosensor data, with an accuracy of 0.71 for a between-persons model and 0.84 for within-person. A second study leveraged heart rate data derived via a consumer-grade wearable in veterans with posttraumatic stress disorder during physical activity, with the aim of investigating heart rate patterns during PTSD hyperarousal events (Sadeghi et al., 2022). Veterans were completing intense physical activity, cycling for several hours per day, over multiple days, and self-reporting hyperarousal events while continuously collecting heart rate data. The study modeled heart rate patterns and found unique heart rate patterns during PTSD hyperarousal events as distinguishable from non-events. Emotional state during the hyperarousal event was not measured, and may indicate fear or anger, but the findings indicated that real-world heart rate detection and modeling is feasible.

There remains a gap in whether such novel approaches can be applied to real-world data in adults who have experienced trauma to predict affective states. Reliable and feasible prediction of affective states and behavior occurring within a timescale capable of delivering an effective intervention would result in a paradigm shift in psychiatry, as digital tools could be leveraged to deliver evidence-based, just-in-time support, within an individual's environment. Anger very commonly (albeit not always) precedes aggressive behavior (Douglas & Martinko, 2001), meaning that detecting high anger may have potential in reducing harmful behaviors such as workplace aggression, family violence, and community violence. With mental health workforces negatively impacted by the pandemic, and persistent challenges with engaging individuals with problem anger in traditional care (Hyatt et al., 2023), there is a need for digital mental health solutions, particularly to reach individuals who have experienced trauma and are unable to access traditional care for financial or logistical reasons. The aim of this research was to explore if continuously measured stress scores based on heart rate variability (HRV) derived via a consumer-grade wearable can be leveraged to predict high anger intensity measured using ecological momentary assessment (EMA).

1. Methods

1.1. Study design

Full details of the study design are published elsewhere Metcalf et al. (2022). Trauma-exposed adults aged 18+

years with problem anger (i.e. ≥ 12 points on the Dimensions of Anger Reactions Scale (Forbes et al., 2014)); who owned an internet-connected smartphone were recruited between August 2022 and March 2023 to participate in this prospective cohort study. Adults were excluded if they had a recent history of significant violence, were active smokers, had physical health conditions affecting the cardiovascular system, or were unwilling to use a wearable for 10 days. Potential participants were screened online, and then telephone screened by experienced researchers to ensure they were suitable for participation. Potential participants were asked about whether they were experiencing current violence in their relationship. Participants recruited from across Australia from either a database of trauma-exposed individuals held by the researchers or targeted social media adverts were asked to complete a pre-study online self-report survey, ambulatory assessment for 10 days (i.e. including continuous data from a vívosmart® 4 Garmin wearable and EMAs delivered four times per day), and post-study online self-report survey. The EMA included questions about anger-related rumination, anger intensity and anger expression and were delivered to the participant's smartphone via the mEMA-sense platform (ilumivu, Inc), designed to support researchers conducting EMA studies with consumer-grade wearables. The Garmin data was transferred via Bluetooth to the participant's smartphone and then pushed to the cloud-based mEMA-sense platform for storage. Alphanumeric participant codes were used to de-identify data. All participants provided written informed consent, and when requested by participants or suggested by researchers, were followed up with further psychological support for PTSD symptoms and/or anger. The total feasible sample size was 100 participants due to practical and cost constraints. Each would provide up to 40 EMAs, resulting in 4000 possible outcome observations.

1.2. High anger intensity episodes

High anger intensity episodes were captured via self-reported anger intensity from EMAs. At each EMA, participants were asked to rate the overall intensity of the anger or irritability that they felt right now, with scores ranging from 1 (lowest intensity) to 10 (highest intensity). High anger intensity was defined as intensity scores of 7 + .

2. Garmin stress scores

Continuously measured stress scores captured by the Garmin wearable sensors were used as a predictor in this study. Garmin devices use HRV to calculate stress level based on an algorithm from Firstbeat Analytics (Firstbeat Technologies Ltd, 2014). The device takes a baseline HRV measurement when the participant is inactive and compares subsequent values to this resting value to determine stress scores on a range from 0 to 100 (higher values indicate greater stress). HRV is the most widely used physiological indicator in mental health research (Hickey et al., 2021). Garmin stress scores were considered as continuous variables in the analysis and were considered useable for the prediction modeling if captured within 10 min prior to a completed EMA.

Models involving only stress scores were considered, as well as models with both stress scores and the following covariates thought to be predictive of anger intensity: gender (1 = Female, 0 = Male), age (years), sleep rating (response to EMA sleep quality item '*During the past night, how would you rate your sleep quality overall?*: 0 = Very good, 1 = Fairly good, 2 = Fairly bad, 3 = Very bad) and problem anger status at previous EMA survey (1 = Yes, 0 = No).

2.1. Statistical methods

A two-stage approach was selected for the prediction modeling because it can handle predictors that have been continuously measured prior to longitudinal outcomes (Matthews et al., 1990). The two-stage approach involves: (1) calculating summary measures (maximum score, variance, etc.) from the continuously measured stress scores for each participant collected 10 min before each EMA survey; and (2) including them as predictors in a generalized linear mixed effects model (GLMM) of the longitudinally measured binary outcome, problem anger. A logistic link function was selected for the GLMM, and two random effect structures were examined: (1) a random effect for the participant (examines between participant variability in log-odds of a problem anger episode); and (2) random effects for study day nested in participant (examines variation in log-odds of a problem anger episode on the same day within a participant, as well as the between participant variability). Summary measures were selected using the approach proposed in Leffondré et al. (2004). This approach involves calculating 24 summary measures capable of discriminating between stable–unstable, increasing–decreasing, linear–nonlinear, monotonic–non-monotonic patterns of change in longitudinal data. Factor analysis is then used to select a subset of non-redundant summary measures. An advantage of this approach is that it has the potential to objectively identify several summary measures that capture the patterns of change in longitudinal data, which may perform better than trying to identify a single summary measure (e.g. mean) that captures the patterns of change in longitudinal data. Each participant had to have a minimum of four stress scores recorded 10 min prior to an EMA survey to calculate the

summary measures. Each predictor/summary measure was standardized by subtracting its mean from each value and dividing by its standard deviation. The variable importance of the summary measures was assessed by comparing the magnitude of the standardized regression coefficients from the GLMMs.

Only participants with usable EMA survey data (both anger intensity captured and stress scores available 10 min prior) were included in the analysis. The distribution of demographic variables and EMA survey responses between usable and non-usable EMA survey data were compared to assess whether the characteristics of participants used in the prediction modeling differed to those excluded. Train/test splits, forward chaining cross-validation (CV), and population informed forward chaining were used to assess whether the predictive performance (calibration and discrimination) of each model will generalize to an independent dataset (Cochrane et al., 2021). Train and test sets for each approach were created by splitting the data by usable EMA surveys and by days with one or more usable surveys. Each of the methods used evaluates a model's predictive performance in a way that respects the temporal structure of the data. Further details about the train/test splits constructed for each cross-validation approach are provided in the eMethods and eTable 1.

Calibration performance is the degree of agreement between the estimated probability of a problem anger episode produced by the model and the actual observed proportion. Due to the low outcome frequency in the train and test sets, calibration performance could only be assessed using the calibration intercept (intercept estimate (logit-scale) from a logistic regression of the outcome on the predicted probabilities included as an offset or with regression coefficient fixed at 1) and slope (slope estimate (logit-scale) from a logistic regression of the outcome on the predicted probabilities) (Van Calster et al., 2016). To produce a precise calibration curve, a minimum of 200 events and 200 non-events has been recommended (Van Calster et al., 2016). A perfectly calibrated model has a calibration intercept and slope of 0 and 1, respectively. A calibration slope <1 indicates the model is overfitted, i.e. predicted probabilities overestimate high observed proportions (or those at high risk) and underestimate low observed proportions (or those at low risk). A calibration slope >1 suggests the model is underfitted, i.e. predicted probabilities underestimate high observed proportions (or those at high risk) and overestimate low observed proportions (or those at low risk). A negative calibration intercept suggests the model overestimates the observed proportions, on average, whereas positive values suggest underestimation of the observed proportions, on average. To examine how training and test set size influence the calibration intercept and slope estimates, these values were calculated for the smallest and largest forward chaining train and test sets (eTable 1).

Discrimination performance is the model's ability to separate EMA data for each participant into high anger intensity or low anger intensity. The discrimination performance of the models was assessed using the area under the receiver operating curve (AUCs) derived for each test set and the balanced accuracy. AUCs close to 1 indicate good discrimination performance (able to discriminate/separate high anger intensity from low anger intensity); AUCs around 0.5 indicate no/poor ability to separate high anger intensity from low anger intensity episodes; and AUCs close to 0 indicate the model is predicting the incorrect event (e.g. predicting high anger intensity as low anger intensity and vice versa). Balanced accuracy is the average of the sensitivity (proportion of high anger intensity correctly predicted) and specificity (proportion of low anger intensity correctly predicted) and was derived by determining a single cutpoint. Predicted probabilities derived from the test sets that were greater than or equal to this cutpoint were classified as high anger intensity episodes. Cutpoint optimization was performed by maximizing the Youden-Index on the training sets. Percentile bootstrap 95% confidence intervals (CIs) were derived for the AUCs from 1000 bootstrap samples stratified by outcome, problem anger, to keep the number of Yes's/No's constant in each sample. Performance measures could not be calculated for the population informed forward chaining splits because each test set contains data for a single participant. As a result, an 'aggregate' AUC and balanced accuracy were derived by aggregating/combining the predicted responses for each split across participants.

GLMMs assume the relationship between the predictors and the outcome (or more accurately the log-odds) is linear. In sensitivity analyses, this assumption was relaxed by exploring the predictive performance of generalized additive mixed models (GAMMs). GAMMs allow complex nonlinear relationships between the predictors (e.g. summary measures) and outcome (Wood, 2004). The two random effect structures specified for the GLMM were also explored for the GAMM. Further details on GLMMs and GAMMs fitted are provided in the eMethods. Additional sensitivity analyses were undertaken to assess the model performance when (i) the anger intensity threshold was lowered to 5 to indicate problem anger, (ii) the summary measures were calculated from stress scores measured 15 min prior to an EMA survey, (iii) gender (pre-survey measure), age (pre-survey measure), sleep rating (EMA) and problem anger status at previous EMA included as predictors (see Table 1 for full list of sensitivity analyses).

Table 1. Description of the primary and sensitivity analyses performed.

Analysis	Summary measures[a]	Anger intensity threshold	Additional covariates[b]
Primary	Garmin stress scores 10 mins prior	7	No
Sensitivity 1	Garmin stress scores 10 mins prior	5	No
Sensitivity 2	Garmin stress scores 10 mins prior	7	Yes
Sensitivity 3	Garmin stress scores 10 mins prior	5	Yes
Sensitivity 4	Garmin stress scores 15 mins prior	7	No
Sensitivity 5	Garmin stress scores 15 mins prior	5	No
Sensitivity 6	Garmin stress scores 15 mins prior	7	Yes
Sensitivity 7	Garmin stress scores 15 mins prior	5	Yes

[a]Summary measures derived from Garmin stress scores available during this time period.
[b]The additional covariates that could be included were gender (1 = Female, 0 = Male) and age (years).

All analyses were performed in R version 4.3.1. The Leffondré et al. (2004) approach was implemented in R's traj package (Sylvestre & Vatnik, 2014). GLMMs were fitted using the glmer function in the lme4 package (Bates et al., 2008). AMMs were fitted using the gam function in the mgcv package (Wood, 2004). The default regression spline in the gam function, thin plate regression splines, was selected (Wood, 2003). In the mgcv package generalized CV is implemented by default to choose the number of knots for the regression splines. Calibration intercepts and slopes were estimated using valProbggplot function from R's CalibrationCurve package (De Cock et al., 2023; Van Calster et al., 2016). Cutpoint optimization was performed using the cutpointr function from R's cutpointr package (Thiele & Hirschfeld, 2021). The AUCs and corresponding percentile bootstrap 95% CIs were also derived using the cutpointr function.

3. Results

3.1. Sample size

In total, 98 participants were recruited. Seven participants had no EMA survey and Garmin device data (only pre-study survey data) and three had no Garmin device data. Among the 88 remaining participants, a total of 3842 EMAs were delivered. In total, 1166 EMA surveys from 84 participants were usable (i.e. had both anger intensity recorded and Garmin stress scores available 10 min prior). Problem anger was observed in 13.8% (161/1166) of surveys from 35 participants. A study flowchart showing how the sample size of the analysis dataset was arrived at is provided in Figure 1. A comparison of the distribution of demographic and EMA survey responses between usable and non-usable EMA survey data is provided in eTable 2. Demographics and survey responses look reasonably similar for those included and excluded from the analysis. In total, 81.6% of respondents had probable PTSD (Price et al., 2016), and experienced on average four traumatic life events according to the Life Events Checklist for DSM-5 (Weathers, 2013). The majority of the sample experienced civilian traumatic events; 43.3% reported at least one sexual assault and 3.1% reported military trauma.

3.2. Summary measure selection

Five non-redundant summary measures were selected using Leffondré et al. (2004) approach (Table 2). The distribution of the summary measures in eFigures 1–5 appears to vary with anger intensity, although those derived prior to high anger intensity report do not appear to be consistently higher or lower than those derived prior to a low anger intensity report.

3.3. Primary analysis

Calibration intercept estimates for the GLMM including a random effect for participant are 0.07 (95% CI: −0.16, 0.31) and 0.06 (95% CI: −0.15, 0.26) for the smallest and largest forward chaining training sets, which suggest that the predicted probabilities slightly underestimate the observed proportions, on average. The calibration slopes for this model (1.23 (95% CI:1.03, 1.43) for smallest forward chaining training set and 1.21 (95% CI:1.03, 1.38) for largest forward chaining training set) suggest the model is underfitted to the training data (although the lower limits of the 95% CIs are close to 1). The calibration slope for the model including a random effect for study day nested within participant was larger than that for the model including only the participant random effect. The calibration slopes were similar for the training sets created by splitting by usable EMA surveys and for those created by splitting by days with usable EMA surveys (eTable 3). The 95% CIs for the calibration intercept and slope estimates derived for the smallest and largest forward chaining test datasets were wide and indicate that the calibration performance of the GLMM could not be accurately estimated for the test sets (eTable 4).

The AUCs derived from each CV method for the GLMM are summarized in Table 3. The AUCs for the test sets appear to stay quite high irrespective of how the train/test splits are constructed, with estimates ranging from 0.76 (95% CI: 0.57, 0.92) to 0.90 (95% CI: 0.83, 0.97), suggesting that the five selected summary measures of the stress scores are predictive of problem anger status (Table 3). The percentile bootstrap 95% CIs also exclude an AUC of 0.5 (a value indicating no predictive ability). However, this high predictive performance might be an artifact of the low outcome frequency. In the train/tests splits the

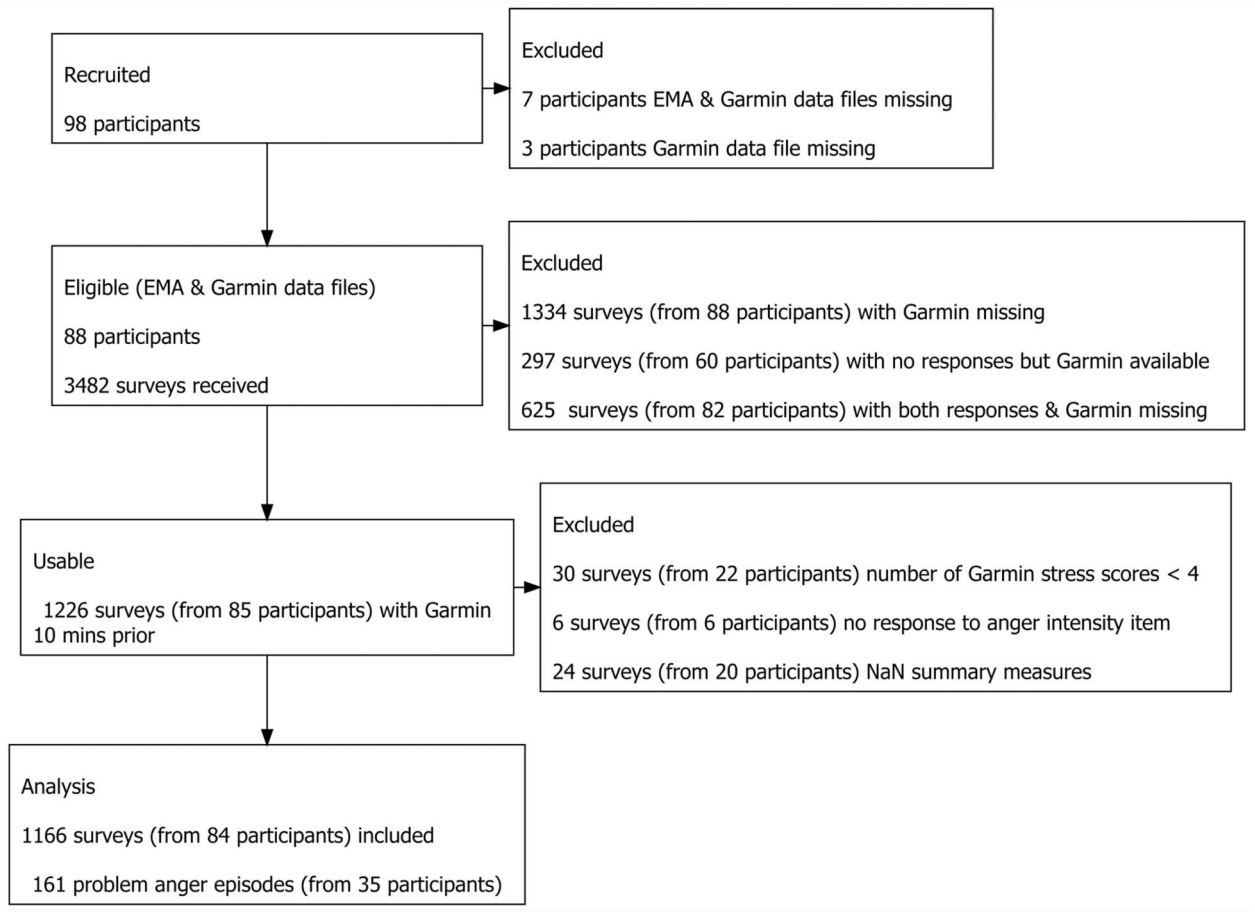

Figure 1. Flow diagram of study participants.

Table 2. Description of selected summary measures.

Summary measure ID	Name	Description
m2	Mean-over-time	Average of the Garmin stress scores.
m5	Change	Difference between the last and first stress score.
m13	SD of the first differences per time unit	Higher values indicate nonlinearity. Smaller (or zero) values indicate approximate linearity, i.e. first differences (across the first differences for the timescale) are relatively constant.
m20	Mean of the absolute second differences	Measures inconsistency of changes. Higher values indicate many peaks and valleys or short-term fluctuations.
m21	Maximum of the absolute second differences	Higher values indicate whether there is a large peak or valley.

outcome frequency can range from 10.7% (8/75) to 16.3% (22/135) in the test set (eTables 5–8). The receiver operating characteristic (ROC) curves for the primary analysis in eFigures 6 (EMA splits) and 7 (days splits) have large steps, which indicate that the outcome prevalence is too low to accurately estimate the predictive performance of this model, i.e. difficult to estimate the sensitivity and 1-specificity at each threshold used to construct the curve.

For the GLMM, the variable importance of the summary measures was assessed by comparing the magnitude of the standardized regression coefficients. Table 4 contains the odds ratio estimates for the summary measures averaged [min, max] across the CV methods and splits. Most of the summary measures were consistent with a reduction in the odds of a problem anger episode for a standard deviation increase in the summary measures, except for the 'SD of the first differences per time unit', which was consistent with no change in the odds of a problem anger episode, on average. The summary measure with the largest reduction was the maximum of the absolute differences (an 8% reduction, on average, in the odds of a problem anger episode was observed for a standard deviation increase in this measure). eFigures 8 and 9 include the 95% CIs for the odds ratio estimates for the re_id and re_study:day models, respectively, derived by fitting these models to the EMA splits. eFigures 10 and 11 are the corresponding figures for the day splits. The 95% CIs, corresponding to the summary measure odds ratio estimates, are quite wide and indicate that these associations are not accurately estimated.

The balanced accuracy varies from 62% to 81% across the CV methods, splits and random effects models (eTable 9). The confusion matrices show that

Table 3. Area under the receiver operating characteristic curves (AUCs) (percentile bootstrap 95% confidence intervals) derived from the test set of each cross-validation method for the primary analysis.

		EMA[a]			Days[b]		
Model	Split no.	T/T	FC	PIFC	T/T	FC	PIFC
re_id	1	0.87 (0.74, 0.97)	0.85 (0.71, 0.97)	0.82 (0.67, 0.95)	0.81 (0.68, 0.93)	0.90 (0.83, 0.97)	0.90 (0.83, 0.96)
re_id	2		0.90 (0.78, 0.98)	0.86 (0.73, 0.97)		0.86 (0.77, 0.93)	0.85 (0.76, 0.93)
re_id	3		0.76 (0.57, 0.92)	0.76 (0.57, 0.92)		0.83 (0.68, 0.96)	0.84 (0.68, 0.96)
re_id	4		0.86 (0.73, 0.97)	0.86 (0.72, 0.97)			
re_id:study_day	1	0.86 (0.72, 0.97)	0.86 (0.71, 0.97)	0.82 (0.68, 0.94)	0.81 (0.64, 0.93)	0.90 (0.82, 0.96)	0.90 (0.82, 0.96)
re_id:study_day	2		0.89 (0.77, 0.98)	0.86 (0.71, 0.96)		0.86 (0.77, 0.93)	0.85 (0.76, 0.93)
re_id:study_day	3		0.77 (0.58, 0.95)	0.77 (0.58, 0.92)		0.83 (0.70, 0.95)	0.84 (0.70, 0.95)
re_id:study_day	4		0.85 (0.70, 0.97)	0.84 (0.69, 0.97)			

Note: re_id = GLMM including a random effect for participant; re_id:study_day = GLMM including random effects for study day nested in participant. T/T – Train/test; FC – Forward chaining; PIFC – Population informed forward chaining.
[a]EMA – Test sets created by splitting at usable EMA surveys.
[b]Days – Test sets created by splitting at days with usable EMA surveys available.

at the optimal threshold most of the high anger intensity reports are predicted accurately, but at a cost of predicting a high number of false positives, e.g. for the EMA train/test split and re_id model (eTable 10) the: true negative rate is 47/67 (correctly predicted as negative / observed negatives); false positive rate is 20/67 (incorrectly predicted as positive / observed negatives); false negative rate is 2/8 (incorrectly predicted as negative / observed positives); true positive rate is 6/8 (correctly predicted as positive / observed positives). Similar findings were observed for the forward chaining and population informed forward chaining EMA splits and for the day splits (eTables 11 and 12). Plots of the observed and predicted problem anger responses for a subset of participants are presented in eFigures 12 and 13 for the train/test splits, eFigures 14 and 15 for the forward chaining splits and eFigures 16 and 17 for the population informed forward chaining CV methods, respectively.

3.4. Sensitivity analyses

The results of the GAMM modeling indicated very little evidence of non-linear relationships between the selected summary measures and the log-odds of a problem anger episode (eFigures 18–21). Consequently, the predictive performance of the GAMM would closely resemble the GLMM, as illustrated by the very similar train/test AUCs for the GLMM (Table 3) and GAMM (eTable 13 and eFigure 22). Therefore, it was decided not to further evaluate the predictive performance of the GAMM using the remaining computationally intensive CV methods.

In addition to the primary analysis, seven sensitivity analyses were performed (Table 1). For the sensitivity analyses involving the additional covariates, large parameter estimates and very wide confidence intervals were observed for the covariates problem anger at previous EMA survey and sleep rating. A cross tab of problem anger (outcome), problem

Table 4. Odds ratio estimates for the summary measures averaged [min, max] across the cross-validation methods for the primary analysis.

		Average odds ratio estimate (min, max)	
Split	Summary measure	re_id	re_id:study_day
EMA[a]			
	Maximum of the absolute second differences	0.92 [0.89, 0.96]	0.93 [0.89, 0.97]
	Mean-over-time	0.93 [0.90, 0.96]	0.93 [0.90, 0.97]
	Mean of the absolute second differences	0.95 [0.83, 0.98]	0.93 [0.80, 0.96]
	Change	0.99 [0.92, 1.01]	0.99 [0.92, 1.01]
	SD of the first differences per time unit	1.01 [0.95, 1.42]	1.02 [0.95, 1.50]
Days[b]			
	Maximum of the absolute second differences	0.92 [0.88, 0.99]	0.93 [0.89, 1.01]
	Mean-over-time	0.93 [0.91, 1.02]	0.93 [0.91, 1.02]
	Mean of the absolute second differences	0.95 [0.82, 0.98]	0.93 [0.78, 0.97]
	Change	0.98 [0.88, 1.01]	0.98 [0.89, 1.01]
	SD of the first differences per time unit	1.01 [0.95, 1.26]	1.02 [0.96, 1.30]

Notes: Estimates derived by fitting the GLMM to the training set of each cross-validation method. re_id = GLMM including a random effect for participant; re_id:study_day = GLMM including random effects for study day nested in participant.
[a]EMA – Training sets created by splitting at usable EMA surveys.
[b]Days – Training sets created by splitting at days with usable EMA surveys available.

anger previous, gender and sleep rating categories revealed this was due to low frequencies (<5) for some combinations. The only additional covariates that could be included in the GLMM were age and gender. Further details regarding GLMM convergence are provided in eTable 14.

The AUCs for the seven sensitivity analyses remained high and comparable to those derived from the primary analysis (eTable 15, eFigures 23 and 24), indicating that the predictive performance of the GLMM including the five summary measures as predictors appears robust to changes in the anger intensity threshold, time used for the Garmin data and inclusion of additional predictors. As for the primary analysis, this high predictive performance might be an artifact of the low outcome frequency. The ROC curves for the sensitivity analyses in eFigures 6 (EMA splits) and 7 (days splits) have large steps, which indicate that the outcome prevalence is too low to accurately estimate the predictive performance of GLMM model from these curves.

4. Discussion

Problem anger after trauma is common, destructive, and difficult to address using traditional psychological methods. New prediction approaches hold significant promise, but remain a nascent area of research, with limited research focusing on prediction in real-world settings. The aim of this project was to determine if physiological data in the form of a HRV-converted stress score, collected via a consumer-grade wearable, could predict high anger intensity episodes in individuals who have experienced trauma. The results showed that a prediction model with good performance could be built leveraging real-world collected stress scores (i.e. a converted measure of HRV) in the ten minutes prior to a high anger intensity moment. These novel findings add to a small body of research showing the potential for physiological derived data to revolutionize prediction approaches to dysregulated emotion and behavior (Hickey et al., 2021). These findings add to an emerging novel evidence base that shows that digital technology can be used to predict affective or cognitive states linked to harmful behavior (Kleiman et al., 2021). It needs to be noted, however, that the predictive performance demonstrated in this study could be optimistic due to the low frequency of high anger intensity and that replication with larger datasets recording higher incidents rates is required. Further research is also needed to build an evidence base around the selections made in this study, such as the prediction variables, the outcome measure of anger intensity, and the temporal window of prediction, to maximally optimize prediction model approaches.

While EMA missing data was consistent with comparable clinical studies (Kivelä et al., 2022; Varker, Arjmand, et al., 2022), the additional ~30% of missing physiological data was unanticipated. Previous qualitative research in this cohort has shown that while wearable data collection is highly feasible, missing data is frequent and is more commonly due to technical issues, such as signal processing, and software and hardware features to store and transfer the data between wearable, smartphone, and cloud, than participant-related factors such as failing to wear the device (Metcalf et al., 2022). While research-grade wearables may have higher data precision, they often cost in excess of USD$1600, and are visibly noticeable to other individuals (Peake et al., 2018). This price point, and the attention they draw, mean they are not feasible or acceptable to certain populations, including vulnerable individuals who have experienced trauma. Consumer-grade wearables are affordable, discreet, and can be leveraged at the scale needed for digital phenotyping research. The measures of HRV are available across all consumer-grade wearables, and although there are some variations (Bent et al., 2020), they require no pre-processing, meaning more feasible translation of the models to widespread use and adoption. However, while promising, further work is needed to prepare researchers for the expected usable data, and for prediction models to manage data when prevalence is relatively low, as is common in most affective and behavior disturbances. As technology continues to rapidly evolve and improve, so will reliability of data for research purposes (Hickey et al., 2021).

The findings of this study have significant implications for future development of digital mental health tools that look to deliver evidence-based interventions in real-time. These approaches, known as just-in-time-adaptive interventions (JITAIs), have potential to leverage both EMA and wearable data to determine optimal and non-optimal times to deliver intervention components. Our findings indicate that while wearable data can be leveraged to detect anger intensity, and thus potentially be incorporated into a JITAI to deliver intervention components at optimal moments, there remains challenges around the feasibility of such an approach when missing data is high (i.e. how many anger intensity moments would a JITAI miss?) and questions remain around tolerance of false positives (i.e. how will users react to being prompted to address their anger intensity, when the model has incorrectly detected it?). Further work around the technical side to improve data collection, and the ethical side of such interventions is needed.

4.1. Limitations

This study is limited by high rates of missing data, and the focus on one feature of the consumer wearable (i.e. HRV). Future studies, with larger datasets and

multiple features might leverage other forms of machine learning. Moreover, the sample size is primarily women, which lowers generalizability to other gender trauma populations with problem anger. External validation samples are needed to examine how well the model performs in other populations. Nevertheless, problem anger in women is grossly understudied and is a recognized area of research need.

4.2. Conclusions

Digital mental health approaches for trauma-affected populations hold tremendous unexplored potential and these novel findings provide highly novel evidence that sensors can be leveraged to predict affective states. However, real-world limitations around reliable data extraction from commercial devices remain challenging to the development of subsequent digital interventions. Overall, these results present a promising outcome for further predictions of abnormal mood and behavior using continuously measured data in individuals who have experienced trauma. With an ongoing crisis in mental health care globally, digital mental health tools that can leverage prediction models will result in safe, affordable, easily accessible solutions.

Acknowledgements

This study was funded by the National Health and Medical Research Council (NHMRC) of Australia (APP2001218).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request. The data are not publicly available due to privacy considerations and ethical approval.

Contributions

OM: funding acquisition, conceptualization, methodology, investigation, writing – original draft preparation. KL: conceptualization, supervision, writing – reviewing and editing. DF: funding acquisition, supervison, writing – reviewing and editing. MOD: supervison, writing – reviewing and editing. TQ: funding acquisition, writing – reviewing and editing. TV: funding acquisition, writing – reviewing and editing. SC: funding acquisition, writing – reviewing and editing. SZ: conceptualization, methodology, data curation, formal analysis, writing – original draft preparation.

ORCID

Olivia Metcalf http://orcid.org/0000-0001-9570-8463
David Forbes http://orcid.org/0000-0001-9145-1605

References

Adler, A. B., LeardMann, C. A., Roenfeldt, K. A., Jacobson, I. G., Forbes, D., & Team, M. C. S. (2020). Magnitude of problematic anger and its predictors in the Millennium Cohort. *BMC Public Health*, *20*(1), 1–11. https://doi.org/10.1186/s12889-020-09206-2

Arjmand, H.-A., Forbes, D., Varker, T., O'Donnell, M. L., Finlayson-Short, L., & Metcalf, O. (2023). Understanding the temporal dynamics of problem anger using sequence analysis. *Emotion*, *23*(8), 2322–2330.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2008). *Fitting linear mixed-effects models using the lme4 package in R*.

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digital Medicine*, *3*(1), Article 18. https://doi.org/10.1038/s41746-020-0226-6

Chida, Y., & Steptoe, A. (2009). The association of anger and hostility with future coronary heart disease: A meta-analytic review of prospective evidence. *Journal of the American College of Cardiology*, *53*(11), 936–946. https://doi.org/10.1016/j.jacc.2008.11.044

Cochrane, C., Ba, D., Klerman, E. B., & Hilaire, M. A. S. (2021). An ensemble mixed effects model of sleep loss and performance. *Journal of Theoretical Biology*, *509*, Article 110497. https://doi.org/10.1016/j.jtbi.2020.110497

Cowlishaw, S., Metcalf, O., Varker, T., Stone, C., Molyneaux, R., Gibbs, L., Block, K., Harms, L., MacDougall, C., Gallagher, H. C., Bryant, R., Lawrence-Wood, E., Kellett, C., O'Donnell, M., & Forbes, D. (2021). Anger dimensions and mental health following a disaster: Distribution and implications after a major bushfire. *Journal of Traumatic Stress*, *34*(1), 46–55. https://doi.org/10.1002/jts.22616

De Cock, B., Nieboer, D., Van Calster, B., Steyerberg, E., & Vergouwe, Y. (2023). *The CalibrationCurves package: Validating predicted probabilities against binary events*. R package version 0.1. 5.

Douglas, S. C., & Martinko, M. J. (2001). Exploring the role of individual differences in the prediction of workplace aggression. *Journal of Applied Psychology*, *86*(4), 547–559. https://doi.org/10.1037/0021-9010.86.4.547

Eckhardt, C. I., Samper, R. E., & Murphy, C. M. (2008). Anger disturbances among perpetrators of intimate partner violence: Clinical characteristics and outcomes of court-mandated treatment. *Journal of Interpersonal Violence*, *23*(11), 1600–1617. https://doi.org/10.1177/0886260508314322

Firstbeat Technologies Ltd. (2014). *Stress & recovery analysis method based on 24-hour HRV*. https://www.firstbeat.com/wp-content/uploads/2015/10/Stress-and-recovery_white-paper_20145.pdf

Forbes, D., Alkemade, N., Mitchell, D., Elhai, J. D., McHugh, T., Bates, G., Novaco, R. W., Bryant, R., & Lewis, V. (2014). Utility of the Dimensions of Anger Reactions–5 (DAR-5) scale as a brief anger measure. *Depression and Anxiety*, *31*(2), 166–173. https://doi.org/10.1002/da.22148

Goodwin, M. S., Mazefsky, C. A., Ioannidis, S., Erdogmus, D., & Siegel, M. (2019). Predicting aggression to others in youth with autism using a wearable biosensor.

Autism Research, *12*(8), 1286–1296. https://doi.org/10.1002/aur.2151

Hickey, B. A., Chalmers, T., Newton, P., Lin, C.-T., Sibbritt, D., McLachlan, C. S., Clifton-Bligh, R., Morley, J., & Lal, S. (2021). Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors*, *21*(10), Article 3461. https://doi.org/10.3390/s21103461

Huhn, S., Axt, M., Gunga, H.-C., Maggioni, M. A., Munga, S., Obor, D., Sié, A., Boudo, V., Bunker, A., Sauerborn, R., Bärnighausen, T., & Barteit, S. (2022). The impact of wearable technologies in health research: Scoping review. *JMIR MHealth and UHealth*, *10*(1), Article e34384. https://doi.org/10.2196/34384

Hyatt, C. S., Sleep, C. E., Hemmy Asamsama, O., & Reger, M. A. (2023). Surveying veterans affairs mental health care providers on experiences working with veteran patients with antagonistic clinical presentations. *Psychological Services*, *21*(2), 379–387.

Kivelä, L., van der Does, W. A., Riese, H., & Antypa, N. (2022). Don't miss the moment: A systematic review of ecological momentary assessment in suicide research. *Frontiers in Digital Health*, *4*, 611–617. https://doi.org/10.3389/fdgth.2022.876595

Kleiman, E. M., Bentley, K. H., Maimone, J. S., Lee, H.-I. S., Kilbury, E. N., Fortgang, R. G., Zuromski, K. L., Huffman, J. C., & Nock, M. K. (2021). Can passive measurement of physiological distress help better predict suicidal thinking? *Translational Psychiatry*, *11*(1), 611–617. https://doi.org/10.1038/s41398-021-01730-y

Leffondré, K., Abrahamowicz, M., Regeasse, A., Hawker, G. A., Badley, E. M., McCusker, J., & Belzile, E. (2004). Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *Journal of Clinical Epidemiology*, *57*(10), 1049–1062. https://doi.org/10.1016/j.jclinepi.2004.02.012

Matthews, J., Altman, D. G., Campbell, M., & Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, *300*(6719), 230–235. https://doi.org/10.1136/bmj.300.6719.230

Metcalf, O., Finlayson-Short, L., Lamb, K. E., Zaloumis, S., O'Donnell, M. L., Qian, T., Varker, T., Cowlishaw, S., & Brotman, M. (2022). Ambulatory assessment to predict problem anger in trauma-affected adults: Study protocol. *PLoS One*, *17*(12), Article e0278926. https://doi.org/10.1371/journal.pone.0278926

Metcalf, O., Little, J., Cowlishaw, S., Varker, T., Arjmand, H.-A., O'Donnell, M., Phelps, A., Hinton, M., Bryant, R., Hopwood, M., McFarlane, A., & Forbes, D. (2021). Modelling the relationship between poor sleep and problem anger in veterans: A dynamic structural equation modelling approach. *Journal of Psychosomatic Research*, *150*, Article 110615. https://doi.org/10.1016/j.jpsychores.2021.110615

Peake, J. M., Kerr, G., & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology*, *9*, Article 329783. https://doi.org/10.3389/fphys.2018.00743

Plummer Lee, C., Mersky, J. P., & Liu, X. (2024). Postpartum anger among low-income women with high rates of trauma exposure. *Journal of Traumatic Stress*, *38*(1), 124–134.

Price, M., Szafranski, D. D., van Stolk-Cooke, K., & Gros, D. F. (2016). Investigation of abbreviated 4 and 8 item versions of the PTSD Checklist 5. *Psychiatry Research*, *239*, 124–130. https://doi.org/10.1016/j.psychres.2016.03.014

Sadeghi, M., Sasangohar, F., McDonald, A. D., & Hegde, S. (2022). Understanding heart rate reactions to post-traumatic stress disorder (PTSD) among veterans: A naturalistic study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *64*(1), 173–187. https://doi.org/10.1177/00187208211034024

Seppälä, J., De Vita, I., Jämsä, T., Miettunen, J., Isohanni, M., Rubinstein, K., Feldman, Y., Grasa, E., Corripio, I., Berdun, J., D'Amico, E., & Bulgheroni, M. (2019). Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: Systematic review. *JMIR Mental Health*, *6*(2), Article e9819. https://doi.org/10.2196/mental.9819

Sylvestre, M.-P., & Vatnik, D. (2014). Using traj package to identify clusters of longitudinal trajectories. *CRAN Packag Traj*, *593*, Article 15.

Thiele, C., & Hirschfeld, G. (2021). *Cutpointr*: Improved estimation and validation of optimal cutpoints in R. *Journal of Statistical Software*, *98*(11), 1–27. doi:doi:10.18637jss.v098.i11

Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment*, *21*(4), 457–462. https://doi.org/10.1037/a0017653

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, 167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005

Varker, T., Arjmand, H., Metcalf, O., Cowlishaw, S., O'Donnell, M., Forbes, D., McFarlane, A., Bryant, R. A., Hopwood, M., Phelps, A., & Hinton, M. (2022). Using an ecological momentary assessment protocol to understand problem anger in veterans. *Journal of Behavior Therapy and Experimental Psychiatry*, *76*, Article 101746. https://doi.org/10.1016/j.jbtep.2022.101746

Varker, T., Cowlishaw, S., Baur, J., McFarlane, A. C., Lawrence-Wood, E., Metcalf, O., Van Hooff, M., Sadler, N., O'Donnell, M. L., Hodson, S., Benassi, H., & Forbes, D. (2022). Problem anger in veterans and military personnel: Prevalence, predictors, and associated harms of suicide and violence. *Journal of Psychiatric Research*, *151*, 57–64. https://doi.org/10.1016/j.jpsychires.2022.04.004

Wastell, C. (2020). *Understanding trauma and emotion: Dealing with trauma using an emotion-focused approach*. Routledge.

Weathers, F. (2013). *The Life Events Checklist for DSM5 (LEC-5)*. National Center for PTSD.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. https://doi.org/10.1198/016214504000000980

Zhang, T., & Chan, A. H. (2016). The association between driving anger and driving outcomes: A meta-analysis of evidence from the past twenty years. *Accident Analysis & Prevention*, *90*, 50–62. https://doi.org/10.1016/j.aap.2016.02.009