Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Smith DL, Held P (2023). Moving toward precision PTSD treatment: predicting veterans' intensive PTSD treatment response using continuously updating machine learning models. *Psychological Medicine* **53**, 5500–5509. https://doi.org/10.1017/S0033291722002689

Received: 15 January 2022
Revised: 28 July 2022
Accepted: 2 August 2022
First published online: 19 October 2022

Key words:
Cognitive processing therapy; longitudinal analysis; machine learning; precision medicine; PTSD; veterans

Author for correspondence:
Dale L. Smith, E-mail: Dale_Smith@rush.edu

Moving toward precision PTSD treatment: predicting veterans' intensive PTSD treatment response using continuously updating machine learning models

Dale L. Smith[1,2] and Philip Held[1]

[1]Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, 325 S. Paulina St., Suite 200, Chicago, IL 60612, USA and [2]Behavioral Sciences, Olivet Nazarene University, 1 University Ave., Bourbonnais, Illinois 60914, USA

Abstract

Background. Considerable heterogeneity exists in treatment response to first-line post-traumatic stress disorder (PTSD) treatments, such as Cognitive Processing Therapy (CPT). Relatively little is known about the timing of when during a course of care the treatment response becomes apparent. Novel machine learning methods, especially continuously updating prediction models, have the potential to address these gaps in our understanding of response and optimize PTSD treatment.

Methods. Using data from a 3-week ($n = 362$) CPT-based intensive PTSD treatment program (ITP), we explored three methods for generating continuously updating prediction models to predict endpoint PTSD severity. These included Mixed Effects Bayesian Additive Regression Trees (MixedBART), Mixed Effects Random Forest (MERF) machine learning models, and Linear Mixed Effects models (LMM). Models used baseline and self-reported PTSD symptom severity data collected every other day during treatment. We then validated our findings by examining model performances in a separate, equally established, 2-week CPT-based ITP ($n = 108$).

Results. Results across approaches were very similar and indicated modest prediction accuracy at baseline ($R^2 \sim 0.18$), with increasing accuracy of predictions of final PTSD severity across program timepoints (e.g. mid-program $R^2 \sim 0.62$). Similar findings were obtained when the models were applied to the 2-week ITP. Neither the MERF nor the MixedBART machine learning approach outperformed LMM prediction, though benefits of each may differ based on the application.

Conclusions. Utilizing continuously updating models in PTSD treatments may be beneficial for clinicians in determining whether an individual is responding, and when this determination can be made.

Introduction

Mounting evidence supports the efficacy of Cognitive Processing Therapy (CPT; Resick, Monson, & Chard, 2017a), which is considered a first line intervention for treating post-traumatic stress disorder (PTSD; APA, 2017; ISTSS, 2017; VA/DoD, 2017). Support comes from randomized controlled trials (Monson et al., 2006; Resick, Nishith, Weaver, Astin, & Feuer, 2002; Resick et al., 2008, 2015, 2017b) as well as clinical research (Asmundson et al., 2019; Held, Smith, Pridgen, Coleman, & Klassen, 2022c; Lloyd et al., 2015). CPT has been successfully delivered in different formats such as the traditional 12 sessions delivered on a weekly basis (Monson et al., 2006; Resick et al., 2002, 2008, 2015, 2017b) and massed/intensive treatments which deliver a full course of treatment in as little as one to three weeks (Galovski et al., 2021; Held et al., 2022a, 2022c). Effect sizes for PTSD severity reduction in CPT are generally large and meaningful when delivered weekly or in massed format (e.g. $d > 1.0$; Asmundson *et al.* 2019; Held, Bagley, Klassen, & Pollack, 2019) and have been demonstrated to persist after treatment for up to ten years following treatment completion (Held et al., 2020b; Resick, Williams, Suvak, Monson, & Gradus, 2012). However, not all participants benefit to the same extent (Dewar, Paradis, & Fortin, 2020). Recent research on massed CPT delivered as part of an intensive PTSD treatment program (ITP) identified four separate PTSD response trajectories (Held et al., 2021). In line with other research examining response trajectories in weekly CPT (Galovski et al., 2016; Schumm, Walter, & Chard, 2013), approximately 15% reached treatment goals within a small number of sessions and 14% didn't respond to treatment in any meaningful way (Held et al., 2021). Given this variability in treatment response across treatment programs for psychiatric conditions, development of prediction models for determining who is, or is likely to be, benefitting from treatment is paramount.

The emerging emphasis on machine learning in developing prediction models in psychological medicine, as well as the increase in the types and amount of data collected in the field, has led to increased use of these methods for various applications, including tracking treatment response (Shatte et al., 2019). Such approaches often differ from traditional statistical approaches in their emphasis on prediction accuracy rather than probabilistic emphasis on specific predictors and aspects of their relationships with outcomes (e.g. slopes or odds ratios). Machine learning models are able to accommodate a larger number of variables as predictors than generally found in traditional statistical approaches. Although some baseline predictors, such as baseline PTSD severity or negative posttraumatic cognitions, have been shown to be useful in predicting such non-responders, the amount of variability in post-treatment PTSD and depression severity that can be accounted for solely via baseline assessment is usually limited (Held et al., 2021, 2022b; Hilbert et al., 2020; Nixon et al., 2021).

Primarily focusing on baseline predictors may be important for initial determination of the appropriateness of a treatment program for an individual (Held et al., 2021; Hilbert et al., 2020; Nixon et al., 2021), however such models also involve considerable uncertainty given the dynamic nature of treatment response over time. The recent emphasis on implementation of precision medicine approaches (Aafjes-van Doorn, Kamsteeg, Bate, & Aafjes, 2021; Chekroud et al., 2021; Delgadillo, 2021; Hilbert et al., 2020) necessitates identification of participants who may or may not be responding to treatment as early as possible. Recently developed machine learning approaches that account for the longitudinal structure of repeated assessments hold promise for improved accuracy in predicting participants' treatment response by continuously updating models with newly acquired information about a patient's treatment response (e.g. repeatedly measured symptom severity scores). The ability to assess individual progress during treatment and update predictions of patient's response is likely a necessary precursor to treatment adjustments in any precision medicine approach.

Although others have attempted clinical prediction models in PTSD outcomes during the course of treatment (Held et al., 2022b; Nixon et al., 2021), these studies have not utilized approaches designed to accommodate the correlated structure inherent to longitudinal data, in which observations are nested within individuals, or have predicted variants of categorized non-response rather than overall PTSD severity. Given the lack of a generally agreed-upon standards for what may constitute non-response to PTSD treatment (Varker et al., 2020), and in the interest of modeling the full spectrum of variability in treatment response, predicting continuous PTSD severity may be a preferred solution.

The current study aimed to examine the ability for machine learning and statistical prediction models to utilize both baseline data and updated PTSD symptom severity information throughout the program to generate increasingly accurate and informative predictions of post-treatment PTSD severity for participants in a 3-week CPT-based ITP. This was evaluated using three approaches; Mixed Effect Random Forest (MERF; Hajjem, Bellavance, & Larocque, 2011, 2014) and Mixed Effects Bayesian Additive Regression Trees (MixedBART; Spanbauer & Sparapani, 2021), which both appropriately model random effects, and gold-standard statistical linear mixed-effects longitudinal models (LMMs) were used to generate these updating predictions. As shown previously (Held et al., 2022b), we expected that models would provide acceptable performance with baseline predictors, but that accuracy would improve throughout the program with the incorporation of updated PTSD severity information as treatment progressed and change trajectories became more apparent. Testing continuously improving models could provide foundational information in implementing a precision medicine-based approach in PTSD treatment. We were generally agnostic regarding the ability for machine learning to outperform mixed-effects regression predictions, given prior research demonstrating that machine learning approaches may not necessary outperform standard statistical approaches in making clinical predictions (Cho et al., 2021; Christodoulou et al., 2019; Li et al., 2021).

Methods

Participants

Data utilized in this study were from 361 veterans with PTSD who completed a 3-week CPT-based ITP at Rush University Medical Center's Road Home Program: Center for Veterans and Their Families. Participants were included if they had complete data[†1]. On average, veterans in the sample were 41.46 years old (S.D. = 9.43). The majority identified as male (63.71%) and White, (67.87%). Additional sample characteristics can be found in Table 1.

Program description

During the 3-week ITP, veterans received 14 individual CPT sessions, 13 group CPT sessions, 13 group mindfulness sessions, and 12 group yoga sessions in addition to psychoeducation classes on various topics, such as sleep hygiene. A more detailed description of the ITP and its outcomes can be found in elsewhere (Held et al., 2020a; Zalta et al., 2018). Veterans were eligible for the ITP if they met the diagnostic criteria for PTSD, which was verified using the Clinician-Administered PTSD Scale for DSM-5 (CAPS-5; Blevins, Weathers, Davis, Witte, & Domino, 2015; Bovin et al., 2016; Weathers et al., 2013). Exclusionary criteria were unstable housing, inability to independently complete activities of daily living, a suicide attempt in the previous 30 days, untreated psychosis or mania, or severe alcohol or drug use that would require continuous medical observation. The study procedures were approved by the Institutional Review Board at Rush University Medical Center with a waiver of consent as all assessments were collected as a part of routine care.

Measures

Veterans were asked to provide demographic information and complete several self-report measures before and during the ITP. A complete list of all features that were used in the different analytic models as well as when they were assessed in ITP can be found in Table 2.

Clinician administered PTSD scale for DSM-5(CAPS-5)

The CAPS-5 is a structured diagnostic PTSD assessment based on the DSM-5 criteria, administered at baseline (Weathers et al., 2018). It assesses the severity of PTSD symptoms across the four different clusters from 0 (absent) to 4 (extreme): intrusions, avoidance, alterations in cognition and mood, and hyperarousal. PTSD symptom severity was based on the past month. Cronbach's alpha within the current sample was 0.780.

[†] The notes appear after the main text.

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

Table 1. Demographic characteristics

Variable[a]	3-Week ITP ($N = 361$) N	3-Week ITP ($N = 361$) %	2-Week ITP[a] ($N = 108$) N	2-Week ITP[a] ($N = 108$) %
Sex				
Male	230	63.71	51	47.22
Female	131	36.29	57	52.78
Ethnicity				
Latinx	70	19.39	19	17.59
Not Latinx	291	80.61	89	82.41
Race				
American Indian/Alaskan Native	5	1.39	1	0.93
Asian	5	1.39	2	1.87
Black or African American	74	20.50	29	27.10
Native Hawaiian/Pacific Islander	3	0.83	0	0.00
Other	27	7.48	8	7.48
Refusal	1	0.28	0	0.00
Unknown	1	0.28	0	0.00
White	245	67.87	67	62.62
Income category				
\$0–\$20 000	36	10.56	22	9.24
\$20 000–\$49 999	116	34.02	73	30.67
\$50 000–\$99 999	109	31.96	90	37.82
>\$100 000	80	23.46	53	22.27
Military service branch				
Air force	21	5.82	10	9.26
Army	245	67.87	59	54.63
Coast guard	2	0.55	2	1.85
Marines	54	14.96	19	17.59
Navy	39	10.80	18	16.67
Service era				
Post 11 September 2001	331	91.61	92	85.19
Deployed				
Yes	286	79.22	77	71.30
Military sexual trauma				
Yes	154	42.66	28	25.93
	M [range]	S.D.	*M [range]*	S.D.
Age	41.46 (24–74)	9.43	42.93 (26–71)	9.27
PCL-5 Baseline[b]	55.63 (19–80)	12.33	50.72 (16–80)	15.15
PCL-5 Post-Treatment	33.79 (0–80)	19.38	34.13 (0–74)	16.35

[a] χ^2 or t test comparisons indicated that significant differences exist between the two programs in sex, race, service era, MST status, and PCL-5 at baseline ($ps < 0.05$).
[b] PCL-5 = PTSD Checklist for DSM-5.

PTSD checklist for DSM-5 (PCL-5)

The PCL-5 is a self-report measure that assess PTSD severity (Weathers et al., 2013). Individuals were asked to rate how much they were bothered by each of the 20 PTSD symptoms from 0 (not at all) to 4 (extremely). PTSD symptom severity was rated based on the past month during the intake and past week at every other timepoint after that. In the 3-week program, the PCL-5 was assessed at baseline and on days 2, 3, 5, 6, 8, 10, 11,

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

Table 2. List of features used in machine learning models

Measure	Description	Scoring	Assessment timepoint
Age		1-item measure	Baseline
Sex		1-item measure recoded as Male or Female	Baseline
Race		1-iteam measure recoded as White or All Other Races	Baseline
Ethnicity		1-item measure recoded as non-Hispanic or Hispanic or Latino	Baseline
Education		1-item measure dummy coded from No high school diploma (0) to Master's or Doctorate degree (7)	Baseline
Marital status		1-item measure dummy coded from Married or Domestic Partnership (0) or Not Married (1)	Baseline
Branch		1-item measure dummy coded from U.S. Army & US Army National Guard (0), U.S. Air Force (1) U.S. Navy (2), U.S. Marines (3), or All Reserves and Coast Guard (4)	Baseline
Deployed	Screens lifetime deployment	1-item measured Yes/No	Baseline
Served after 11 September 2001		1-item measured Yes/No	Baseline
Military sexual trauma	Screens endorsement of military sexual trauma	2-items measured Yes/No	Baseline
Referral source	Origin of referral	1-item measure dummy coded from Warrior Care Network/Wounded Warrior Project (0), Other (1), Employee (2), Veterans Affairs (3), Road Home Program Outreach (4), Community Provider (5)	Baseline
Clinician administered PTSD Scale (CAPS-5; Weathers et al., 2018)	Past month PTSD severity	20-items measured Absent (0) – Extreme/Incapacitating (4)	Baseline
PTSD Checklist for DSM-5(PCL-5; Bovin et al., 2016; Weathers et al., 2013)	Past month PTSD severity at baseline; Past week PTSD severity at all other timepoints	20-items measured Not At All (0) – Extremely (4)	Baseline, Days 2, 3, 5, 6, 8, 10, 11, 13, Post
Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001)	Past two weeks depression symptoms	9-items measured Not At All (0) – Nearly Every Day (4)	Baseline
Alcohol Use Disorder Identification Test (AUDIT-C; Bush et al., 1998)	Past year alcohol use	3-items measured Never (0) – Four or More Times a Week (4)	Baseline
Neurobehavioral Symptoms Inventory (NSI) Validity-10 (Cicerone & Kalmar, 1995; Vanderploeg et al., 2014)	Screens possible symptom exaggeration/overreporting on the NSI	10-items measured None (0) – Very Severe (4)	Baseline
Posttraumatic Cognitions Inventory (PTCI; Foa et al., 1999)	Assesses trauma related cognitions	33-items measured Totally Disagree (1) – Totally Agree (7)	Baseline

13, and post-treatment. A total score of 33 is considered the threshold for 'probable PTSD.' Cronbach's alphas ranged from 0.897-0.962 across study timepoints.

Patient health questionnaire (PHQ-9)

The PHQ-9 is a 9-item self-report measure of depressive symptoms (Kroenke, Spitzer, & Williams, 2001). Individuals were asked to rate how much they were bothered by their depression symptoms from 0 (not at all) to 3 (nearly every day). For the present study, depression symptoms were assessed for the past two weeks at baseline. Cronbach's alpha within the current sample was 0.810.

Posttrauma cognition inventory (PTCI)

The PTCI is a 33-item self-report measure of negative posttrauma cognitions was administered at baseline (Foa, Ehlers, Clark, Tolin, & Orsillo, 1999). Individuals were asked to rate how much they agreed or disagreed with a range of beliefs from 1 (totally disagree) to 7 (totally agree). Cronbach's alpha among study participants was 0.951.

Alcohol use disorder identification test – consumption (AUDIT-C)

The AUDIT-C is a 3-item self-report measure of alcohol consumption (Bush, Kivlahan, McDonell, Fihn, & Bradley, 1998). Individuals were asked to rate how often they drank, how many drinks they had when they were drinking, and how often they had six or more drinks on one occasion. The AUDIT-C assessed alcohol consumption over the past year and was administered at baseline. Cronbach's alpha in this study was 0.866.

Neurobehavioral symptom inventory – 10-item validity scale (VAL-10)

(Vanderploeg et al., 2014). The VAL-10 is a 10-item self-report scale made up of items from the Neurobehavioral Symptom Inventory, assessed at baseline (Vanderploeg et al., 2014). The items were selected to identify individuals who may be over-reporting neurobehavioral symptoms. Cronbach's alpha among study participants was 0.907.

Analytic strategy

We employed three mixed-effects-based prediction models designed to accommodate the longitudinal structure inherent to assessment of symptom severity during and at the end of the treatment program[2]. The first, Mixed Bayesian Additive Regression Trees (MixedBART) is a recently developed non-parametric Bayesian approach which accommodates random effects within machine learning. This approach utilizes an ensemble of decision trees to predict response. Priors, which are utilized in Bayesian analyses and represent existing beliefs regarding quantities or distributions in Bayesian analysis, are placed on program parameters, including variable selection probabilities. MixedBART and BART default parameters regarding priors and number of trees, without extensive cross-validation, are generally adequate and outperform other machine learning and statistical methods under many conditions. Based on insight from previous work (Held et al., 2022b), we used Dirichlet, rather than uniform, priors for variable selection probabilities. This allows models to adapt to the existence of more useful predictors in the dataset, thus accommodating the expectation that clinical features and updated PTSD severity values are likely to be more useful in prediction than demographic features (Held et al., 2021, 2022b). As a Bayesian analytic method, MixBART approaches inference by sampling from the posterior distribution generated computationally utilizing existing data and relevant priors. We used 10 000 posterior draws with 5000 burn-in draws, which was a conservative approach compared to other applications of MixBART (Spanbauer & Sparapani, 2021), but aligns with common practices and recommendations in Bayesian analysis (e.g. Raftery & Lewis, 1991) and resulted in good overall model convergence. Based on prior recommendations using BART approaches we employed 200 trees (Chipman, George, & McCulloch, 2010), though we explored reduced numbers of trees to assess importance of individual features due to the tendency for BART models to potentially incorporate more irrelevant features when the number of trees is large. However, due to overall consistency across models with differing numbers of trees we report results of the primary models utilizing 200 trees here[3].

The second approach utilized mixed-effects random forest (MERF; Hajjem et al., 2011, 2014). This tree-based random forest approach accommodates random effects for longitudinal or otherwise clustered data utilizing the expectation-maximization (EM) algorithm, a maximum likelihood estimation method that progresses through stages of estimating latent variables and optimizing the model until convergence is reached. Five-folds cross validation on the training set was applied. We also progressively increased numbers of trees and iterations in training set model development, though asymptotes for the utility of such increases in both appeared to exist at beyond approximately 150 iterations and 200–300 trees.

Finally, linear mixed effects regression models (LMMs) were also explored for machine learning model comparison to traditional statistical model accuracy using the same data. This is an accepted approach to modeling longitudinal data due to its accommodation of random effects and missing data, and less restrictive assumptions (Hedeker & Gibbons, 2006). For this analysis, we both examined models with all predictors and models utilizing only the top five predictors as defined by both the MixedBART and MERF machine learning programs, which both identified the same five predictors. Since the use of the top five predictors resulted in models that were as accurate as those including more, or all, covariates at every timepoint we examined, we present only these results of the LMM approach. The same strategy of creating and testing a model on the training and test sets, respectively, was utilized in order to remain comparable to the machine learning models.. Cross validation was not used in linear mixed model analyses to best approximate typical applied statistical use of this longitudinal approach.

We randomly split the data from the 3-week ITP approximately 60:40 into training ($n = 232$) and test ($n = 130$) datasets. This random split was implemented at the participant-level due to the nesting of timepoint measurements within individuals. Training and test sets did not differ on any demographic or clinical variable ($ps > 0.10$). The training set was then used to train machine learning and LMM models with all baseline demographic and clinical data (see Table 2) as well as lagged PCL-5 scores predicting post-treatment PCL-5. Following training, we examined prediction accuracy on the test set at baseline as well as at each assessment timepoint (see online Supplementary Table S4 for accuracy using training data). Thus, when examining accuracy on the test set at baseline only baseline predictors were used to predict post-treatment PTSD severity. On program days 3, 5, 6, 8, 10, 11, and 13 all baseline features as well as PCL-5 scores for all days up to, and including, that day's PCL-5

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

measurement were used. Only PTSD severity score was continuously updated throughout the program. Accuracy of predictions was assessed via R^2 and RMSE. Each analytic approach models change longitudinally, though our primary emphasis here is on prediction of post-treatment PTSD severity measurement. For MixedBART these values were obtained via the mean of each participant's predicted values against actual post-treatment PCL-5 scores.

Due to the importance of external validation of prediction models, we examined the predictive accuracy of these three models in a sample of 108 participants who had completed a separate, equally established, 2-week CPT-based ITP with similar programming combining individual CPT with adjunctive services, which has previously been demonstrated to be non-inferior to the 3-week program (Held et al., 2022c). Due to the differences in timeline between the two ITPs, assessment timepoints were mapped onto the existing time points based on proportion of the program that had been completed at each measurement timepoint. The three longitudinal prediction models that were generated with 3-week training data were then used to predict post-treatment PTSD severity in the 2-week ITP using the same updating-prediction model approach. In the 2-week ITP we focused on baseline and mid-program (beginning of week 2) predictions of post-treatment PTSD severity. MixedBART and LMM analyses were conducted using the MxBART and LMER4 packages in R version 4.1.1, and MERF analyses were conducted using the MERF package in Python version 3.6. Figures were created using R.

Results

Veterans in the 3-week ITP improved in PTSD severity by an average of 21.57 points (S.D. = 18.80). Approximately 70% ($n = 263$) improved by at least 10 points, with 51% ($n = 185$) finishing treatment below the PCL-5 cutoff of 33. As illustrated in Fig. 1, this constituted meaningful overall change across program timepoints, though considerable variability existed in the amount of change, particularly as treatment progressed. This increase in variability across time is generally expected and illustrates the effect of participants' differential improvement during treatment. The demographic and clinical variables in the models other than PCL-5 accounted for approximately 6% of the variability in treatment response throughout the program beyond what PTSD severity accounted for, indicating that both initial accuracy and improvements in predictions were largely driven by PTSD severity and updated PTSD severity measurements.

Both machine learning approaches identified PCL-5, time, baseline PTCI, baseline PHQ-9, and CAPS-5 Intrusions as the most important or utilized features in predicting PTSD severity. Thus, these were used in subsequent LMMs for comparison (see online Supplementary Table S5 for comparison of LMM with all features and only these features). The three analytic approaches to predicting post-treatment PTSD severity closely aligned with regards to accuracy. Baseline predictions of final PCL-5 score on the test sample yielded an overall R^2 of 0.18 for final PCL-5 severity score prediction across all three models (see Table 3). As expected, as updated PTSD severity scores became available during treatment, the accuracy of final timepoint predictions increased substantially (see Fig. 2). At the start of the second week of treatment, (Day 6), all models were able to account for roughly half of the variability in post-treatment PTSD severity. This could potentially represent a milestone at which current treatment progress could be reliably determined in the 3-week ITP. By mid-program (Day 8) R^2 exceeded 0.60 for all analytic methods.

Results of external validation with the 2-week ITP suggest model predictions were similarly accurate as in the 3-week ITP

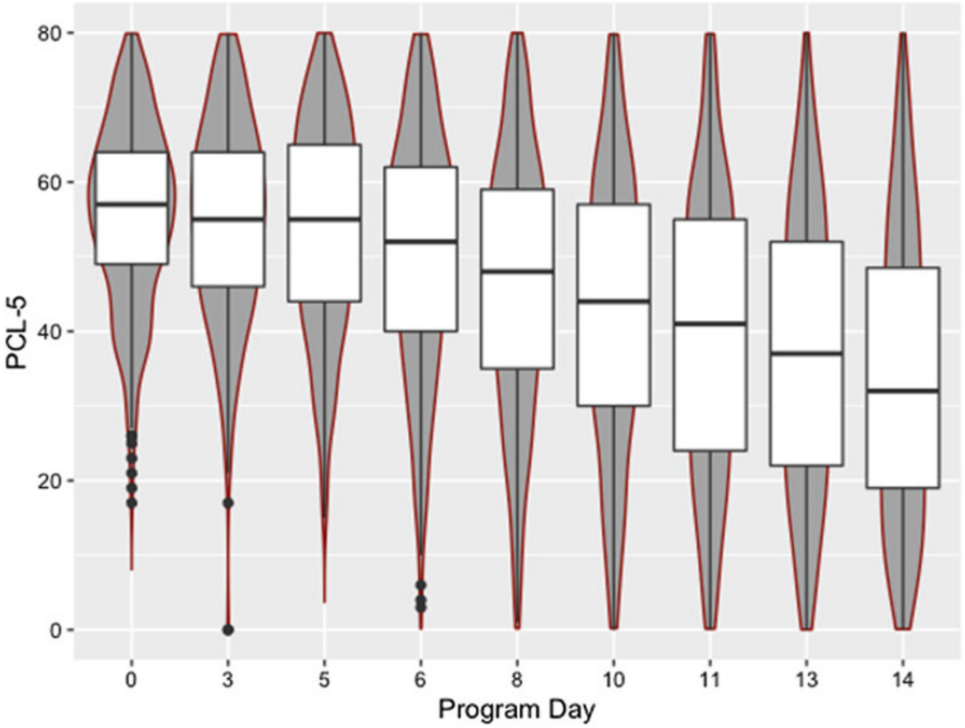

Fig. 1. Distributions of PTSD severity across all participants over time.
Note. Violin plot illustrates distribution at each timepoint, with internal box plots representing median and interquartile range.

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

Table 3. Longitudinal updating models comparison

	MixedBART		MERF		LMM top 5 features[a]	
Features	R^2	RMSE	R^2	RMSE	R^2	RMSE
Baseline predictors only[b]	0.18	17.66	0.18	23.12	0.18	21.97
Baseline + PCL to day 3	0.29	21.37	0.27	21.54	0.29	16.35
Baseline + PCL to day 5	0.48	18.79	0.46	18.50	0.45	14.36
Baseline + PCL to day 6	0.51	16.41	0.49	16.29	0.50	13.71
Baseline + PCL to day 8	0.62	13.38	0.62	13.17	0.62	11.98
Baseline + PCL to day 10	0.74	10.42	0.71	10.68	0.73	10.16
Baseline + PCL to day 11	0.79	9.19	0.80	8.75	0.80	8.57
Baseline + PCL to day 13	0.86	7.70	0.83	8.52	0.85	7.59

[a] LMMs including more predictors were examined but did not outperform the five-predictor model.
[b] Baseline model contained all baseline data, including intake PCL-5 score.

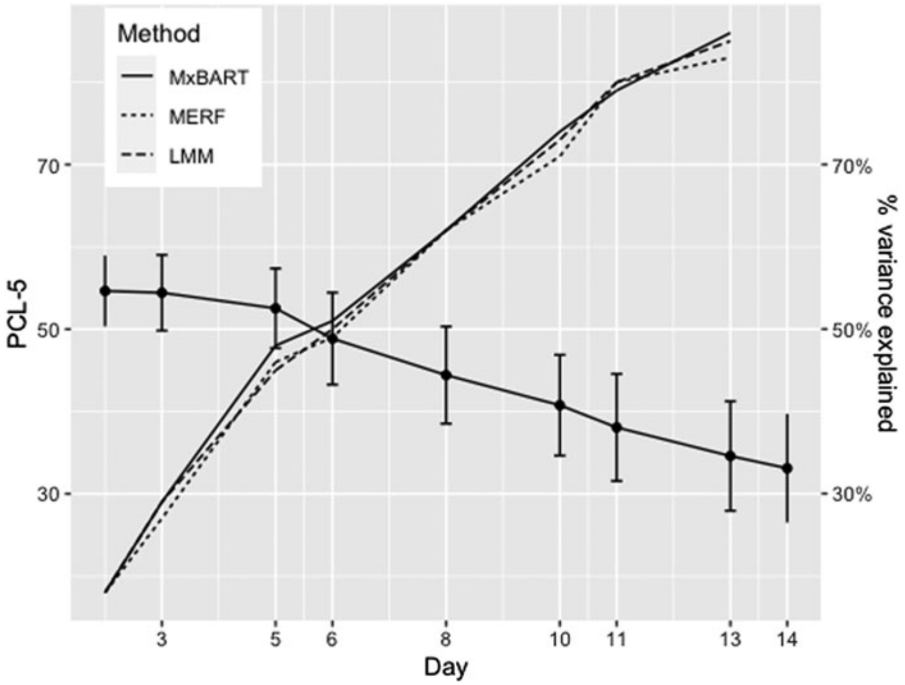

Fig. 2. Test set PTSD severity and predictive accuracy over time.
Note. Error bars represent 95% confidence intervals.

Table 4. External validation results

	MixedBART		MERF		LMM top 5	
Features	R^2	RMSE	R^2	RMSE	R^2	RMSE
Baseline predictors only[a]	0.20	27.47	0.18	18.26	0.23	14.94
Baseline + PCL-5 to mid-program	0.60	11.82	0.58	11.56	0.55	11.33

[a] Baseline model contained all baseline data, including intake PCL-5 score, mid-program predictions included baseline data plus PCL-5 scores to mid-program.

despite not training the models on these data (see Table 4). Baseline predictions generally accounted for about 20% of the variability in post-treatment PTSD severity. Including PCL-5 data up to mid-program led to being able to account for over half of the variability in final PTSD severity by that point. This supports the generalizability of model predictions to similar, but external, clinical data.

Discussion

Our results support the utility of updating prediction models of PTSD severity as a potential clinical tool for assessing PTSD treatment progress and to help identify timepoints for altering a participant's treatment approach. Before the 3-week ITP's midpoint, each model was able to account for a large proportion of the variability in post-treatment PTSD severity. This remained true even in an external 2-week ITP sample. These models can provide valuable clinical information that support a precision-medicine approach to PTSD treatment, as the majority of those identified as likely non-responders with some certainty at mid-program were found to be non-responders at the end of treatment (see online Supplementary Fig. S1). Thus, by deploying such relatively low-cost models in clinical practice, a clinician would be able to obtain acceptable near real-time estimates about their patient's likely endpoint PTSD severity. As such, continuously updating prediction models may be helpful in PTSD treatment in general and may be particularly useful for intensive treatments given the rapid nature of this treatment approach and the limited time clinicians have to evaluate data before needing to make treatment decisions.

As illustrated, and commonly seen in treatment, improvement was far from uniform, with the amount of variability in reported PTSD severity increasing across time. Though generally expected in longitudinal studies, this highlights the need for increased attention to individual change, and the utility of assessing such change during treatment. Indeed, change in PTSD severity during the program was clearly the most effective predictor of PTSD severity at endpoint. Other clinical and demographic predictors accounted for approximately 6% of the variability in endpoint PTSD severity, with baseline PTSD severity accounting for both the remaining 14% at baseline and the improvements in these predictions as additional severity measurements became available. Thus, the best predictor of heterogeneity in total treatment response is clearly the amount of improvement that the individual is making during the program. This highlights the importance of models that can effectively accommodate this and the additional assumptions inherent to longitudinal modeling rather than basing treatment decisions entirely on baseline predictors or a predetermined amount of change that needs to have been reached by mid-treatment without accounting for change trajectories.

Results obtained here do not support the superiority of any specific analytic method utilized, though all models performed at least as well as machine learning models that ignore the longitudinal structure of these data, without the potential bias that can arise when ignoring the lack of independence of observations over time (see online Supplementary Table S2). Linear mixed effects regression models were capable of predicting PTSD outcome severity with the same degree of accuracy as machine learning models. This result joins a wealth of evidence that traditional statistical approaches can perform similarly to machine learning models (Cho et al., 2021; Christodoulou et al., 2019; Li et al., 2021), though, to our knowledge, this study represents the first such application in a continuously updating prediction model for psychiatric treatment response.

Despite similarities in prediction accuracy, unique benefits to each longitudinal approach used exist. LMMs provide easily interpretable slope coefficients and metrics regarding significance of individual predictors. Assumptions, as well as aspects of longitudinal structure such as covariance structure or autocorrelation, are easily assessed with this approach, and missing data is easily accommodated. Conversely, both machine learning approaches may more readily accommodate more predictors in applications involving high dimensional datasets or multiple correlated predictors. An additional well-known benefit of Bayesian approaches is the ability to quantify and visualize uncertainty in estimates. Although the mean predicted value for each participant from the posterior is reported in model output and was utilized above to obtain model accuracy metrics, the credible intervals can also be easily obtained to assess the degree of uncertainty in predictions. However, we found that MixedBART yielded overly optimistic estimates of variability around prediction means, so we would caution against use of program generated credible intervals blindly[4].

A number of limitations need to be acknowledged. The use of self-report assessments may have increased variability in reporting and it being the only continuously updated variable may be viewed as a limitation, though our prior work has suggested that updates in other variables did not improve predictions in any meaningful way when including lagged PCL-5 scores. Additionally, the fact that PTSD severity measurements over time explained most of the variability in treatment response may have obscured potential roles of other contributing factors. However, this also highlights the importance of utilizing such updated severity information. Also, sample size considerations for some demographic variables, such as race, reduced power for intensive examination of demographic moderators of treatment response, though our prior work has indicated that such demographic variables generally did not impact treatment response for either the 3- or 2-week ITP (Held et al., 2022c). Only ITP completers were examined, although a completer bias is unlikely since completion rates were quite high (>90%), and completers and non-completers did not differ on any baseline demographic or clinical variables for either sample, except for a difference in race in the 3-week sample. Although use of a 2-week ITP validation sample is a strength of the current analysis, it may be similar in many ways to the original sample that an external sample may not. For example, although the treatment schedule differed between the 3- and 2-week ITPs with the later drastically reducing group treatment and adjunctive service components, both centered around CPT (Held et al., 2022c). However, demographics breakdown between the two programs indicated that significant differences existed in sex, MST status, race, service era, and baseline PTSD severity (see Table 1). Finally, although many of the exclusion criteria resemble those used in other PTSD treatment, some were specific to ITPs (e.g. stable housing, ability to travel) and may limit the generalizability of the findings presented here.

Conclusion

Considerable additional research is warranted to better understand specific individual factors that could interact with the chosen treatment approach to individualized treatment. However, our demonstration of the use of continuously updating machine learning or predictive modeling using standard longitudinal statistical approaches to assess progress and predict PTSD treatment outcomes shows promise for precision medicine in the field of PTSD. Such models can provide clinicians with information about which patients may progress through treatment as expected or benefit from treatment alterations based on their predicted response. Using the models presented here, such decisions can relatively reliably be made by mid-treatment in the two ITPs we examined. Future research should examine the feasibility of integrating these models into clinical care and systematically testing whether treatment modifications for individuals predicted to have less favorable treatment responses can improve their outcomes, as well as whether findings generalize to more traditional weekly treatment and/or evidence-based PTSD treatments.

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291722002689.

Financial support. This work was supported by the Road Home Program at Rush's partnership with Wounded Warrior Project®.

Conflict of interest. Philip Held receives grant support from Wounded Warrior Project® and RTI International. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, Wounded Warrior Project®, or any other funding agency.

All other authors declare that they have no competing interests.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Notes

[1] Some participants (i.e., 27% of the original sample) were excluded due to missing covariate data. Data were missing due to a number of reasons, including participants missing assessment sessions due to sickness. However, as this is a sample of program completers, covariate and outcome missingness was not associated with any relevant predictors or PTSD severity and can likely be considered missing at random (MAR). Because of our interest in approximating clinical applications that may require complete cases in machine learning approaches, only complete-case results using listwise deletion are reported here. As a sensitivity analysis we examined robustness of presented results to results using imputation (MICE) of baseline covariates. See online Supplementary Table S1 for results of this sensitivity analysis.

[2] Traditional machine learning models using updated PCL-5 scores at each timepoint were also explored with the same training and test sets predicting PCL-5 at final measurement for comparison. These included Random Forest with 5-folds cross validation on the training set. We provide a sample of these results in online supplementary Table S2. Performance was similar, though R^2 values were generally slightly lower, in these models ignoring the longitudinal structure of data.

[3] Variations on priors for random effects error estimates were also explored as recommended, including degrees of freedom for the inverse chi-squared distribution between three and ten, as well as priors on probabilistic structure of regression trees from cross-validation, allowing for variable selection probabilities to be equal across predictors, as well as differing specification of probabilistic structures of regression trees. Test-set performance metrics were largely unaffected, except that models that assumed equal importance of priors were always poorer predictors than this using Dirichlet priors (see online supplementary Table S3).

[4] Although default settings were used for performance metrics, calibration resulted in more accurate test-set predictions. Estimates provided with regards to credible interval estimation are those obtained following tuning of priors for variability estimates and calibration via linear transformation of predicted values. Regardless of calibration status, we found credible intervals were often too narrow, and that between 74 and 86% of the MixedBART-generated 94% credible intervals contained actual final PCL-5 scores.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729.

APA. (2017). *Clinical practice guideline for the treatment of posttraumatic stress disorder (PTSD) in adults*. American Psychological Association. https://www.apa.org/ptsd-guideline.

Asmundson, G. J. G., Thorisdottir, A. S., Roden-Foreman, J. W., Baird, S. O., Witcraft, S. M., Stein, A. T., … Powers, M. B. (2019). A meta-analytic review of cognitive processing therapy for adults with posttraumatic stress disorder. *Cognitive Behaviour Therapy*, *48*(1), 1–14. https://doi.org/10.1080/16506073.2018.1522371.

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The posttraumatic stress disorder checklist for *DSM*-5 (PCL-5): Development and initial psychometric evaluation: Posttraumatic stress disorder checklist for *DSM*-5. *Journal of Traumatic Stress*, *28*(6), 489–498. https://doi.org/10.1002/jts.22059.

Bovin, M. J., Marx, B. P., Weathers, F. W., Gallagher, M. W., Rodriguez, P., Schnurr, P. P., & Keane, T. M. (2016). Psychometric properties of the PTSD checklist for diagnostic and statistical manual of mental disorders–fifth edition (PCL-5) in veterans. *Psychological Assessment*, *28*(11), 1379–1391. https://doi.org/10.1037/pas0000254.

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., & Bradley, K. A. (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. *Archives of Internal Medicine*, *158*(16), 1789–1795. https://doi.org/10.1001/archinte.158.16.1789.

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., … Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-AOAS285.

Cho, S. M., Austin, P. C., Ross, H. J., Abdel-Qadir, H., Chicco, D., Tomlinson, G., … Lee, D. S. (2021). Machine learning compared with conventional statistical models for predicting myocardial infarction readmission and mortality: A systematic review. *Canadian Journal of Cardiology*, *37*(8), 1207–1214. https://doi.org/10.1016/j.cjca.2021.02.020.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

Cicerone, K. D., & Kalmar, K. (1995). Persistent postconcussion syndrome: The structure of subjective complaints after mild traumatic brain injury. *Journal of Head Trauma Rehabilitation*, *10*(3), 1–17. https://doi.org/10.1097/00001199-199510030-00002.

Delgadillo, J. (2021). Machine learning: A primer for psychotherapy researchers. *Psychotherapy Research*, *31*(1), 1–4. https://doi.org/10.1080/10503307.2020.1859638.

Dewar, M., Paradis, A., & Fortin, C. A. (2020). Identifying trajectories and predictors of response to psychotherapy for post-traumatic stress disorder in adults: A systematic review of literature. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *65*(2), 71–86. https://doi.org/10.1177/0706743719875602.

Foa, E. B., Ehlers, A., Clark, D. M., Tolin, D. F., & Orsillo, S. M. (1999). The posttraumatic cognitions inventory (PTCI): Development and validation. *Psychological Assessment*, *11*(3), 303–314. https://doi.org/10.1037/1040-3590.11.3.303.

Galovski, T. E., Harik, J. M., Blain, L. M., Farmer, C., Turner, D., & Houle, T. (2016). Identifying patterns and predictors of PTSD and depressive symptom change during cognitive processing therapy. *Cognitive Therapy and Research*, *40*(5), 617–626. https://doi.org/10.1007/s10608-016-9770-4.

Galovski, T. E., Werner, K. B., Weaver, T. L., Morris, K. L., Dondanville, K. A., Nanney, J., … Iverson, K. M. (2021). Massed cognitive processing therapy for posttraumatic stress disorder in women survivors of intimate partner violence. *Psychological Trauma: Theory, Research, Practice and Policy.*, 769–779. https://doi.org/10.1037/tra0001100.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451–459. https://doi.org/10.1016/j.spl.2010.12.003.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328. https://doi.org/10.1080/00949655.2012.741599.

Hedeker, D. R., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley-Interscience.

Held, P., Bagley, J. M., Klassen, B. J., & Pollack, M. H. (2019). Intensively delivered cognitive-behavioral therapies: An overview of a promising treatment delivery format for PTSD and other mental health disorders. *Psychiatric Annals*, *49*(8), 339–342. https://doi.org/10.3928/00485713-20190711-01.

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press

Held, P., Klassen, B. J., Boley, R. A., Wiltsey Stirman, S., Smith, D. L., Brennan, M. B., … Zalta, A. K. (2020a). Feasibility of a 3-week intensive treatment program for service members and veterans with PTSD. *Psychological Trauma: Theory, Research, Practice and Policy*, *12*(4), 422–430. https://doi.org/10.1037/tra0000485.

Held, P., Kovacevic, M., Petrey, K., Meade, E. A., Pridgen, S., Montes, M., … Karnik, N. S. (2022a). Treating posttraumatic stress disorder at home in a single week using 1-week virtual massed cognitive processing therapy. *Journal of Traumatic Stress*, 35, 1215–1225.. https://doi.org/10.1002/jts.22831.

Held, P., Schubert, R. A., Pridgen, S., Kovacevic, M., Montes, M., Christ, N. M., … Smith, D. L. (2022b). Who will respond to intensive PTSD treatment? A machine learning approach to predicting response prior to starting treatment. *Journal of Psychiatric Research*, *151*, 78–85. https://doi.org/10.1016/j.jpsychires.2022.03.066.

Held, P., Smith, D. L., Bagley, J. M., Kovacevic, M., Steigerwald, V. L., Van Horn, R., & Karnik, N. S. (2021). Treatment response trajectories in a three-week CPT-based intensive treatment for veterans with PTSD. *Journal of Psychiatric Research*, *141*, 226–232. https://doi.org/10.1016/j.jpsychires.2021.07.004.

Held, P., Smith, D. L., Pridgen, S., Coleman, J. A., & Klassen, B. J. (2022c). More is not always better: 2 weeks of intensive cognitive processing therapy-based treatment are noninferior to 3 weeks. *Psychological Trauma: Theory, Research, Practice, and Policy*. https://doi.org/10.1037/tra0001257.

Held, P., Zalta, A. K., Smith, D. L., Bagley, J. M., Steigerwald, V. L., Boley, R. A., … Pollack, M. H. (2020b). Maintenance of treatment gains up to 12-months following a three-week cognitive processing therapy-based intensive PTSD treatment programme for veterans. *European Journal of Psychotraumatology*, *11*(1), 1789324. https://doi.org/10.1080/20008198.2020.1789324.

Hilbert, K., Kunas, S. L., Lueken, U., Kathmann, N., Fydrich, T., & Fehm, L. (2020). Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: A machine learning approach. *Behaviour Research and Therapy*, *124*, 103530. https://doi.org/10.1016/j.brat.2019.103530.

ISTSS. (2017). *Posttraumatic stress disorder prevention and treatment guidelines: Methodology and recommendations*. International Society of Traumatic Stress Studies. https://istss.org/clinical-resources/treating-trauma/new-istss-prevention-and-treatmentguidelines.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16* (9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS ONE*, *16*(4), e0250370. https://doi.org/10.1371/journal.pone.0250370.

Lloyd, D., Couineau, A.-L., Hawkins, K., Kartal, D., Nixon, R. D. V., Perry, D., & Forbes, D. (2015). Preliminary outcomes of implementing cognitive processing therapy for posttraumatic stress disorder across a national veterans' treatment service. *The Journal of Clinical Psychiatry*, *76*(11), e1405–e1409. https://doi.org/10.4088/JCP.14m09139.

Monson, C. M., Schnurr, P. P., Resick, P. A., Friedman, M. J., Young-Xu, Y., & Stevens, S. P. (2006). Cognitive processing therapy for veterans with military-related posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *74*(5), 898–907. https://doi.org/10.1037/0022-006X.74.5.898.

Nixon, R. D. V., King, M. W., Smith, B. N., Gradus, J. L., Resick, P. A., & Galovski, T. E. (2021). Predicting response to cognitive processing therapy for PTSD: A machine-learning approach. *Behaviour Research and Therapy*, *144*, 103920. https://doi.org/10.1016/j.brat.2021.103920.

Raftery, A. E., & Lewis, S. (1991). *How many iterations in the Gibbs sampler?*. Fort Belvoir, VA: Defense Technical Information Center. https://doi.org/10.21236/ADA640705.

Resick, P. A., Monson, C. M., & Chard, K. M. (2017a). *Cognitive processing therapy for PTSD: A comprehensive manual*. New York: Guilford Press.

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology*, *70*(4), 867–879.

Resick, P. A., Uhlmansiek, M. O., Clum, G. A., Galovski, T. E., Scher, C. D., & Young-Xu, Y. (2008). A randomized clinical trial to dismantle components of cognitive processing therapy for posttraumatic stress disorder in female victims of interpersonal violence. *Journal of Consulting and Clinical Psychology*, *76*(2), 243–258. https://doi.org/10.1037/0022-006X.76.2.243.

Resick, P. A., Wachen, J. S., Dondanville, K. A., Pruiksma, K. E., Yarvis, J. S., Peterson, A. L., … Young-McCaughan, S. (2017b). Effect of group vs individual cognitive processing therapy in active-duty military seeking treatment for posttraumatic stress disorder: A randomized clinical trial. *JAMA Psychiatry*, *74*(1), 28. https://doi.org/10.1001/jamapsychiatry.2016.2729.

Resick, P. A., Wachen, J. S., Mintz, J., Young-McCaughan, S., Roache, J. D., Borah, A. M., … Peterson, A. L. (2015). A randomized clinical trial of group cognitive processing therapy compared with group present-centered therapy for PTSD among active duty military personnel. *Journal of Consulting and Clinical Psychology*, *83*(6), 1058–1068. https://doi.org/10.1037/ccp0000016.

Resick, P. A., Williams, L. F., Suvak, M. K., Monson, C. M., & Gradus, J. L. (2012). Long-term outcomes of cognitive-behavioral treatments for posttraumatic stress disorder among female rape survivors. *Journal of Consulting and Clinical Psychology*, *80*(2), 201–210. https://doi.org/10.1037/a0026602.

Schumm, J. A., Walter, K. H., & Chard, K. M. (2013). Latent class differences explain variability in PTSD symptom changes during cognitive processing therapy for veterans. *Psychological Trauma: Theory, Research, Practice, and Policy*, *5*(6), 536–544. https://doi.org/10.1037/a0030359.

Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, *49*(9), 1426–1448.

Spanbauer, C., & Sparapani, R. (2021). Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. *Statistics in Medicine*, *40*(11), 2665–2691. https://doi.org/10.1002/sim.8924.

VA/DoD. (2017). *VA/DOD clinical practice guideline for the management of posttraumatic stress disorder and acute stress disorder*. Department of Veterans Affairs, Department of Defense. https://www.healthquality.va.gov/guidelines/MH/ptsd/VADoDPTSDCPGFinal.pdf.

Vanderploeg, R. D., Cooper, D. B., Belanger, H. G., Donnell, A. J., Kennedy, J. E., Hopewell, C. A., & Scott, S. G. (2014). Screening for postdeployment conditions: Development and cross-validation of an embedded validity scale in the neurobehavioral symptom inventory. *Journal of Head Trauma Rehabilitation*, *29*(1), 1–10. https://doi.org/10.1097/HTR.0b013e318281966e.

Varker, T., Kartal, D., Watson, L., Freijah, I., O'Donnell, M., Forbes, D., … Hinton, M. (2020). Defining response and nonresponse to posttraumatic stress disorder treatments: A systematic review. *Clinical Psychology: Science and Practice*, *27*(4), e12355. https://doi.org/10.1037/h0101781.

Weathers, F. W., Bovin, M. J., Lee, D. J., Sloan, D. M., Schnurr, P. P., Kaloupek, D. G., … Marx, B. P. (2018). The clinician-administered PTSD scale for DSM–5 (CAPS-5): Development and initial psychometric evaluation in military veterans. *Psychological Assessment*, *30*(3), 383–395. https://doi.org/10.1037/pas0000486.

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The PTSD Checklist for DSM-5 (PCL-5). Scale available from the National Center for PTSD at www.ptsd.va.gov.

Zalta, A. K., Held, P., Smith, D. L., Klassen, B. J., Lofgreen, A. M., Normand, P. S., … Karnik, N. S. (2018). Evaluating patterns and predictors of symptom change during a three-week intensive outpatient treatment for veterans with PTSD. *BMC Psychiatry*, *18*(1), 242. https://doi.org/10.1186/s12888-018-1816-6.

https://doi.org/10.1017/S0033291722002689 Published online by Cambridge University Press