Article:

Tait, J. orcid.org/0000-0002-5336-0363, Kellett, S. and Delgadillo, J. orcid.org/0000-0001-5349-230X (2025) Using machine learning methods to predict the outcome of psychological therapies for post-traumatic stress disorder: A systematic review. Journal of Anxiety Disorders, 112. 103003. ISSN 0887-6185

https://doi.org/10.1016/j.janxdis.2025.103003

Using machine learning methods to predict the outcome of psychological therapies for post-traumatic stress disorder: A systematic review

James Tait [a,b,*], Stephen Kellett [c], Jaime Delgadillo [d]

[a] *School of Psychology, University of Sheffield, ICOSS Building, 219 Portobello, Sheffield, S1 4DP, United Kingdom*

[b] *Department of Health Sciences, University of York, Seebohm Rowntree Building, Heslington, York, YO10 5DD, United Kingdom*

[c] *Grounded Research, RDaSH NHS Foundation trust, 2 St Catherine's Close, Tickhill Road Hospital, Balby, Doncaster, DN4 8QN, United Kingdom*

[d] *Clinical and Applied Psychology Unit, School of Psychology, University of Sheffield, Cathedral Court Floor F, 1 Vicar Lane, Sheffield, S1 2LT, United Kingdom*

* Corresponding author.
Email address: james.tait@york.ac.uk

Funding source declaration

The first author (JT) was funded by a University of Sheffield ESRC White Rose DTP Pathway Award Scholarship, awarded by the Economic and Social Research Council. The funding agency had no involvement in this study.

Acknowledgements

For the purpose of open access, the author has applied a Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) licence to any Author Accepted Manuscript version arising.

Declaration of interest

The authors declare that they have no known potential competing interests that could affect the objectivity of the work presented in this paper.

Citation

Tait, J., Kellett, S., & Delgadillo, J. (2025). Using machine learning methods to predict the outcome of psychological therapies for post-traumatic stress disorder: A systematic review. *Journal of Anxiety Disorders*, 103003. https://doi.org/10.1016/j.janxdis.2025.103003

Abstract

Background: A number of treatments are available for post-traumatic stress disorder (PTSD), however, there is currently a lack of data-driven treatment selection and adaptation methods for this condition. Machine learning (ML) could potentially help to improve the prediction of treatment outcomes and enable precision mental healthcare in practice.

Objectives: To systematically review studies that applied ML methods to predict outcomes of psychological therapy for PTSD in adults (e.g., change in symptoms, dropout rate), and evaluate their methodological rigour.

Methods: This was a pre-registered systematic review (CRD42022325021), which synthesised eligible clinical prediction studies found across four research databases. Risk of bias was assessed using the PROBAST tool. Study methods and findings were narratively synthesised, and adherence to ML best practice evaluated.

Results: Seventeen studies met the inclusion criteria, including samples derived from experimental and observational study designs. All studies were assessed as having a high risk of bias, notably due to inadequately powered samples and a lack of sample size calculations. Training sample size ranged from $N < 36 – 397$. The studies applied a diverse range of ML methods such as decision trees, ensembling and boosting techniques. Five studies used unsupervised ML methods, while others used supervised ML. There was an inconsistency in the reporting of hyperparameter tuning and cross-validation methods. Only one study performed external validation.

Conclusions: ML has the potential to advance precision psychotherapy for PTSD, but to enable this, ML methods must be applied with greater adherence to best practice guidelines.

Key words: Systematic Review; Posttraumatic Stress Disorder; Psychotherapy; Machine Learning.

Highlights

- All studies were rated high risk of bias, primarily due to inadequate sample size.
- None of the studies reported every step of the machine learning pipeline.
- Just one reported external validation, in randomly partitioned hold-out sample.
- Two studies compared machine learning to traditional methods, with mixed results.
- ML may advance precision treatment for PTSD but methods must be applied rigorously.

1. Introduction

Post-traumatic stress disorder (PTSD) is a severe and often chronic mental health problem that can develop following exposure to one or more traumatic events, and is associated with significantly impaired quality of life, increased incidence of physical health problems, co-occurring mental health problems, and suicide (Karatzias et al., 2019; Pacella et al., 2013; Shalev et al., 2017; Yehuda et al., 2015). PTSD affects around 4% of adults worldwide, with higher prevalence rates associated with low income and social deprivation (Fear et al., 2016; Koenen et al., 2017; Ravi et al., 2023). Clinical practice guidelines (CPG) recommend trauma-focussed psychological therapies such as cognitive processing therapy (CPT), prolonged exposure (PE), and eye movement desensitisation and reprocessing (EMDR), as first-line treatments for PTSD (APA, 2017; VA/DoD, 2023; NICE, 2018). Meta-analyses of randomised controlled trials (RCTs) have evidenced that these are currently the most effective forms of psychological therapy for PTSD, and when compared to waitlist controls, pooled effect sizes were large (Jericho et al., 2021; Lewis, Roberts, Andrew, et al., 2020; Mavranezouli et al., 2020). Further, a network meta-analysis (Merz et al., 2019) found that trauma-focussed psychological therapies were equivalent to pharmacological therapy in the short-term and were more effective long-term (Merz et al., 2019), and there is evidence that a majority of patients prefer psychological therapy (including trauma-focussed therapy) to pharmacological therapy (Simiola et al., 2015; Swift et al., 2017).

Despite the availability of efficacious psychological therapies for PTSD, many patients do not respond well to treatment. In a systematic review of RCTs, Schottenbauer et al. (2008) found that non-response rates ranged from 20% – 67% for PE, 3.6% – 48% for CPT, and 7.3% – 92% for EMDR. In a smaller but more recent review of treatment for combat-related PTSD, Steenkamp et al. (2015) found that 60% – 72% of patients still met diagnostic criteria for PTSD after receiving CPT or PE. Response rates may be even lower in routine clinical practice; an analysis of 2,493 patient records from the English National Health Service (NHS) *Talking Therapies* programme found that only 32% of patients accessing trauma-focussed cognitive behavioural therapy (Tf-CBT) achieved reliable and clinically significant improvement in symptoms (Robinson et al., 2020). A contributing factor to nonresponse is poor acceptability of the psychological therapy and associated dropout. Lewis et al. (2020) systematically reviewed dropout from RCTs of psychological therapies for PTSD and found that the pooled dropout rate was 16% (95% CI [14, 18%]), suggesting that around one in six patients dropout. Furthermore, dropout rates were higher for trauma-focussed therapies. The pooled dropout rate for EMDR was 18% (95% CI [12, 24%]), for PE was 22% (95% CI [16%, 28%]), and for CPT was 30% (95% CI [22%, 39%]). This highlights a dilemma, which is that patients with PTSD appear most likely to drop out from the treatments that are the most efficacious. As with treatment response, dropout rates may be even higher in routine clinical practice than in RCTs (Najavits, 2015).

One way that PTSD treatment outcomes might be improved is through personalised mental healthcare. This entails identifying the optimal treatment approach, length, or intensity, based on patients' individual characteristics (Cohen et al., 2021). There is evidence for heterogeneity in response to psychological therapy for PTSD (Herzog & Kaiser, 2022), and studies have found that patients with specific demographic and clinical characteristics may be more likely to respond to a specific trauma-focussed therapy (Deisenhofer et al., 2018; Keefe et al., 2018). For example, Deisenhofer et al. (2018) developed a statistical algorithm to identify patients who were more likely to respond to Tf-CBT than EMDR, and vice versa, based on pre-treatment demographic and clinical data. Implementing a treatment selection algorithm such as this in clinical practice has the potential to improve treatment outcomes by allocating individual patients to the treatment that is most likely to benefit them. Further, PTSD is a complex and heterogeneous condition (Galatzer-Levy & Bryant, 2013), and a number of studies have found evidence for subtypes of PTSD. For example, a “threat reactivity” subtype, high in intrusions, hyperarousal and avoidance, and a “dysphoric” subtype, high in anhedonia and negative affect (Campbell et al., 2020; Campbell-Sills et al., 2022; Horn et al., 2016; Pietrzak et al., 2014). Recent studies have found that patients with certain subtypes of depression respond differentially to CBT (Catarino et al., 2022; Simmonds-Buckley et al., 2021), and it is possible that this is also the case for PTSD (Forbes et al., 2003).

In psychotherapy outcome research, a large number of variables each explain a small proportion of variance in treatment outcome (Barawi et al., 2020; Dewar et al., 2020; Malejko et al., 2017), and it is likely that many of these variables covary, interact, or are non-linear. To account for this, researchers have begun to utilise machine learning (ML) methods (Aafjes-van Doorn et al., 2021), which are particularly well suited to analyse data of this nature (Chekroud et al., 2021). For example, penalised regression methods such as elastic net (Zou & Hastie, 2005) can perform predictor selection by shrinking coefficients for variables with little predictive value or high multicollinearity. Decision tree methods such as random forest (Breiman, 2001) can also implicitly handle complex non-linear relationships and interactions by sequentially dividing the data at the most informative threshold on important predictor variables.

ML is a data-driven approach that uses algorithms to detect patterns in data, with the goal of making accurate predictions in new data (Delgadillo, 2021). In this way ML methods are distinct from classical statistical methods, which predominantly aim to test hypotheses, make inferences, and explain variance within a particular sample (Bi et al., 2019; Yarkoni & Westfall, 2017). When applied optimally, the ML approach follows a sequence of six steps referred to as the *ML pipeline* (Delgadillo & Atzil-Slonim, 2022). These are [1] sample size calculation, [2] data pre-processing, [3] hyperparameter selection, [4] training the model, [5] testing the model with internal cross-validation, and [6] external validation of the model in independent data. Neglecting or inadequately performing any of the first five steps leads to *overfitting* (i.e., capitalising on the idiosyncrasies of the training data to the detriment of generalisability). Without step six, the extent of overfitting is unknown. The strength of evidence provided by ML studies can be categorised into three levels of increasing robustness: In *level 1* evidence, model performance is only evaluated within the training dataset without internal-cross validation; In *level 2*, internal cross-validation is applied; In *level 3*, the model is externally validated by applying the predictors and parameters selected during internal cross-validation to predict outcomes in independent data (Delgadillo & Atzil-Slonim, 2022).

Thus far, much of the research applying ML methods to predict psychological therapy outcomes has focussed on the treatment of depression and anxiety, and relatively little has focussed on treatment for PTSD (Aafjes-van Doorn et al., 2021; Lee et al., 2018; Sajjadian et al., 2021; Vieira et al., 2022). Ramos-Lima et al. (2020) systematically reviewed the use of ML methods in PTSD research but focussed primarily on studies that sought to predict the presence or onset of PTSD and did not include any studies that sought to predict the outcome of CPG recommended psychological therapies. Vieira et al. (2022) systematically reviewed studies that applied ML methods to predict outcomes for CBT, but this review excluded studies that predicted continuous outcomes (e.g., Deisenhofer et al., 2018), excluded other trauma-focussed psychological therapies (e.g., PE, EMDR), and only included one study that sought to predict outcomes in adults with PTSD (Zhutovsky et al., 2019). Further, the above reviews noted frequent methodological issues such as inadequate sample size and validation methods. If ML methods are not applied robustly then prediction models will not generalise and will be of little clinical utility. None of the previous reviews used a quality benchmark of the stages of a ML study provided by the *ML pipeline*.

Therefore, the present study aimed to conduct the first systematic review of studies that used ML methods to predict psychological therapy outcomes for PTSD. For the reasons outlined above, the focus of this review is on the application and reporting of each study's methods, benchmarked against the ML pipeline. The review question was framed following the recommendations of Moons et al. (2014) and Palazón-Bru et al. (2020) for framing systematic reviews of prognostic modelling studies and was reported following PRISMA guidelines (Page et al., 2021). After assessing risk of bias, study methods and results were synthesised, and the adherence to each step of the ML pipeline was evaluated.

2. Method

2.1. Pre-registration

The systematic review protocol was pre-registered with the PROSPERO database prior to conducting searches (Reference: CRD42022325021). The pre-registration can be accessed here: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022325021

2.2. Eligibility criteria

Inclusion and exclusion criteria are described in Table 1. To be included any study must have applied ML methods to pre-treatment data to predict the outcome of a psychological therapy recommended by clinical practice guidelines (CPG) as a first line treatment for PTSD in adults. CPG are intended to bridge the gap between evidence and practice by recommending treatments based on systematic reviews of empirical evidence and/or consensus in expert opinion (Hamblen et al., 2019). The inclusion criteria for this systematic review were guided by CPGs grounded in well conducted systematic reviews, which been appraised to meet an acceptable quality standard using a standardised measure (Martin et al., 2021) and were published in the previous 5-years to ensure that they were contemporaneous (Shekelle et al., 2001). This included the following CPGs: American Psychological Association (2017), International Society for Traumatic Stress Studies (2018), National Institute for Health and Care Excellence (2018), Phoenix Australia Centre for Posttraumatic Mental Health (2021), and Veterans Affairs/Department of Defence (2017). The psychological therapies recommended in these CPGs (see Supplementary Table 1) were predominantly trauma-focussed cognitive behavioural therapies, exposure-based therapies, and EMDR.

2.3. Information sources, searching, and screening

Pre-defined search terms were used to search four databases: APA PsycInfo (via Ovid), PTSDpubs (via ProQuest), PubMed, and Scopus. The search terms were designed to return any studies that mentioned any form of psychological therapy, PTSD or trauma, and any form of machine learning in the title, abstract or key words. The full search strategy is presented in Supplementary Materials. No limits, restrictions, or filters were applied. Databases were searched on 27th April 2022. The following review articles were checked for potentially eligible studies: Aafjes-van Doorn et al. (2021), Chekroud et al. (2021), Chen et al. (2022), Dewar et al. (2020), Dwyer et al. (2018), Hahn et al. (2017), Glaz et al. (2021), Malgaroli and Schultebraucks (2021), Manchia et al. (2020), Meehan et al. (2022), Ramos-Lima et al. (2020). Forward and backward citation searches for all eligible studies were performed using the R package *citationchaser* (Haddaway, 2021). The authors of all eligible studies were contacted to request further studies. Article metadata and abstracts for all search results were imported into EndNote 20 (https://endnote.com/). Duplicates automatically identified by EndNote 20 were screened and removed manually. Further duplicates were identified manually and removed during title and abstract screening. All titles and abstracts were manually screened against the inclusion and exclusion criteria in EndNote 20 by the first author, and full text files of potentially eligible studies were retrieved and screened.

2.4. Data extraction and synthesis

Relevant data from all eligible studies was extracted by the first author using a standardised data extraction table in Microsoft Excel, based on the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS; Moons et al., 2014), and the ML pipeline domains described by Delgadillo and Atzil-Slonim (2022). This included sample characteristics; treatment details; measures (including outcome and candidate predictor variables); ML methods and their purpose (e.g., predictor selection, prediction, clustering); pre-processing details; hyperparameter setting methods, validation methods, model evaluation metrics including accuracy (e.g., R^2, balanced accuracy), error (e.g., root mean squared error, mean absolute error), calibration, and discrimination (e.g., sensitivity, specificity, area under the receiver operating characteristic curve); predictors included in final model; relevant findings; and authors' interpretation of findings. When necessary, study authors were contacted via email to clarify methods and results. Study characteristics, methods, and findings were tabulated and summarised using a narrative synthesis. The pre-registered intention was to quantitatively synthesise prediction model performance metrics using random effects meta-analysis, but this was not possible due to heterogeneity of study methods. Studies adherence to the ML pipeline and corresponding level of evidence (apparent validation, internal cross-validation, external validation) was evaluated.

2.5. Risk of bias assessments

Risk of bias was assessed using the Prediction model study Risk Of Bias Assessment Tool (PROBAST; Moons et al., 2019). A second researcher independently conducted risk of bias assessments for 50% of the included studies. Cohen's kappa was calculated as a measure of agreement, discrepancies were discussed, and a third researcher was consulted where necessary. After consulting with a third researcher a unanimous decision was reached on all ratings.

3. Results

3.1. Study selection

The study selection process is presented in the PRISMA diagram (see Figure 1). In total, 1,570 titles and abstracts were screened, 48 potentially eligible full texts were screened, and 17 studies met the inclusion criteria for the review. Full texts that were screened and excluded are presented in Supplementary Table 2 with reasons for exclusion. Frequent reasons for exclusion included: No ML methods used (*k* = 15), did not predict treatment outcome (*k* = 6), no CPG recommended therapy for PTSD (*k* = 6).

3.2. Study characteristics

Study characteristics are presented in Table 2. Most studies conducted a retrospective analysis of data (*k* = 12), either from clinical trials (*k* = 5), cohort studies (*k* = 1), or routine clinical practice (*k* = 6). Five studies prospectively collected data for analysis, either as a clinical trial (*k* = 1) or cohort study (*k* = 4). Five studies sampled any adults seeking treatment for PTSD; six sampled from military populations; five specified PTSD related to interpersonal-, childhood-, or sexual-abuse; and two sampled patients with co-occurring mental health problems (substance use disorder and depression, respectively). Three studies included only female participants, and one study included only male participants. Participants received a range of CPG recommended psychotherapies for PTSD, most frequently PE (*k* = 10 studies), CPT (*k* = 6 studies), EMDR (*k* = 4 studies), or Tf-CBT (*k* = 3). Total sample size ranged from *N* = 57-612. All but one of the studies were published between 2018 and 2022. Nine studies were conducted in the USA, three in Germany, three in the Netherlands, one in Australia, and one was an analysis of data from England by a team of researchers in Germany and the UK.

3.3. Risk of bias assessments with PROBAST

Detailed risk of bias assessments are presented in Supplementary Table 3. The first and second rater initially agreed on seven out of nine studies, corresponding to a Cohen's kappa = 0.4, indicating *fair* agreement. Following consultation with a third researcher consensus was reached on all nine studies. All seventeen studies were rated at high risk of bias overall, primarily as all studies were high risk of bias in the *Analysis* domain. None of the studies had a reasonable number of participants with the outcome, and for some studies the number of predictor parameters estimated was unclear (studies often reported the number of candidate variables but did not report dummy coding of categorical variables or whether psychometric measures were entered as total scores, factors, or items). Although nine studies reported metrics of prediction accuracy, error, and/or discrimination, none of the studies reported calibration and therefore relevant model performance metrics were not evaluated appropriately. Thirteen studies did not include all enrolled participants in the analysis. Three studies inappropriately handled missing data and six studies did not provide information on the handling of missing data. Seven studies were rated at risk of bias due to selection of participants for using routinely collected clinical data or retrospective cohort study data.

3.4. Study methods and results

Study methods are presented in Table 3 and results are presented in Supplementary Table 4.

3.4.1 Outcome variable

Fourteen studies sought to predict treatment response but operationalised response in a variety of different ways. Eight studies sought to predict treatment response as a continuous outcome, five of which predicted change in PTSD score, two predicted post-treatment PTSD score, and one predicted post-treatment depression score as a proxy outcome (Deisenhofer et al., 2018). Six studies sought to predict treatment response as a categorical outcome, two of which predicted percentage change in PTSD score (50% and 30% respectively) as a binary outcome, one predicted reliable change in PTSD score as a binary outcome, two predicted latent trajectory class membership as a polytomous outcome (Hendriks et al., 2018; Nixon et al., 2021), and one predicted latent trajectory class membership as two binary outcomes (Held et al., 2022). The remaining three studies sought to predict treatment retention, two of which predicted a count of the number of sessions attended, and one predicted dropout as a binary outcome (Keefe et al., 2018).

3.4.2 Candidate predictor variables

Thirteen studies employed psychometric data (e.g., self-report or clinician-report measures of PTSD, depression, anxiety) as candidate predictor variables, eleven of these also used demographic data (e.g., gender, age, employment status), and eleven also used clinical data (e.g., diagnoses, medication use). Eleven studies tested baseline PTSD symptoms and PTSD-related cognitions as candidate predictors, and seven of these also tested trauma characteristics such as type of trauma and time since trauma. Four studies explored the relationship between neuroimaging data (MRI and EEG) and PTSD treatment outcomes. Number of candidate predictor variables ranged from approximately 5 to 104. Studies that used neuroimaging data did not specify the number of candidate predictors. See Supplementary Table 5 for details of candidate predictor variables.

3.4.3 Predictors included in the final model

Among the fourteen studies sought to predict treatment response, all but one (Nixon et al., 2021) reported at least one significant pre-treatment predictor. Five studies included PTSD severity as a predictor in the final model, three of these also included trauma related variables; six studies included co-occurring mental health problems such as depression (k = 5) and emotion regulation difficulties (k = 2); and five included demographic variables such as age (k = 3) and gender (k = 3). Three studies using MRI data identified regions of the brain associated with treatment response, but there was no consensus (Etkin et al., 2019; Zhutovsky et al., 2019; Zilcha-Mano et al., 2020). Studies found that PTSD, trauma, and mental health related variables were stronger predictors of treatment response than demographic variables (Held et al., 2022; Herzog et al., 2021; Hoeboer et al., 2021; Stirman et al., 2021; Stuke et al., 2021). Three studies sought to predict treatment retention or dropout but there was no consensus among the predictors selected in the final model (Fleming et al., 2018; Keefe et al., 2018; López-Castro et al., 2021).

3.4.4. Machine learning methods

Studies used a range of different ML methods for various purposes. Fourteen studies used supervised ML methods. Eight studies used decision tree-based methods, and all but two of these used ensemble tree methods such as random forest and boosting algorithms (ADAboost, gradient boosted models). Three studies used a penalised regression method called *elastic net* (Held et al., 2022; Herzog et al., 2021; Stirman et al., 2021). Three studies used kernel methods (support vector machine, Gaussian process classifier) to analyse MRI data (Etkin et al., 2019; Zhutovsky et al., 2019; Zilcha-Mano et al., 2020). Five studies used unsupervised clustering (k-means) or dimension reduction methods (principal component analysis, independent component analysis). None of the studies used Bayesian ML methods or neural networks.

Five studies used the same ML method to perform feature selection, parameter estimation, and outcome prediction (Fleming et al., 2018; Held et al., 2022; Herzog et al., 2021; Kratzer et al., 2019; Nixon et al., 2021). Two studies used an unsupervised ML method for feature reduction and then used a supervised ML method for prediction (Stuke et al., 2021; Zhutovsky et al., 2019). Five studies used supervised ML methods to select predictors, and then used simpler statistical methods (e.g., linear regression, correlation) to test the relationship between the selected predictors and outcome (Hoeboer et al., 2021; Keefe et al., 2018; López-Castro et al., 2021; Stirman et al., 2021; Zilcha-Mano et al., 2020). One study used a genetic algorithm to select predictors for a linear regression model (Deisenhofer et al., 2018). One study used generalised linear modelling to select predictors and then used a supervised ML method to predict outcomes (Etkin et al., 2019).

Three studies used *k*-means cluster analysis: Zhang et al. (2021) used *k*-means to identify PTSD subtypes and then linear mixed models to test the relationship between subtypes and treatment outcome. Hendriks et al. (2018) used *k*-means to identify treatment response trajectory classes, and then used stepwise logistic regression to select predictors and predict trajectory class membership. Forbes et al. (2003) used *k*-means to test the reliability of PTSD subtypes identified by Ward's hierarchical cluster analysis, and then used generalised linear modelling to test differences in treatment response between subtypes.

Two studies compared the performance of more than one ML method (Etkin et al., 2019; Held et al., 2022), and two studies compared the performance of ML methods against that of traditional statistical methods (Held et al., 2022; Stuke et al., 2021).

3.4.5. Adherence to the ML pipeline

The number of studies that reported each section of the ML pipeline is presented in Figure 2.

3.4.5.1. Sample size calculation

None of the studies reported a sample size calculation. The number of participants with the outcome in a model development sample ranged from < 36 (Etkin et al., 2019) to 397 (Herzog et al., 2021).

3.4.5.2. Data pre-processing

Nine studies reported handling of missing data, six of which reported multiple imputation. Three studies performed multiple imputation via random forest, but only one reported out-of-bag error estimates (Stirman et al., 2021). One study reported listwise exclusion of participants with missing data (Held et al., 2022); one excluded participants missing follow-up data (Zhutovsky et al., 2019); one excluded participants missing a whole scale and imputed mean values where < 20% of a scale was missing (Stuke et al., 2021). Three studies reported reduction of categorical variables, one reported transformation of variables, one reported handling of class imbalance, and one reported case-control matching. Three of four studies that used neuroimaging data reported preprocessing of neuroimaging data. Four studies did not report any pre-processing of data.

3.4.5.3. Hyperparameter selection

Six studies reported using internal-cross validation to optimise hyperparameter settings, one of which also reported using default settings for some hyperparameters. Two studies only reported using default hyperparameter settings (López-Castro et al., 2021; Nixon et al., 2021). Two studies reported using statistical criteria (goodness of fit, gap statistic) to decide the number of *k*-means clusters (Hendriks et al., 2018; Zhang et al., 2021). Some studies reported setting for some but not all hyperparameters, and seven studies did not report hyperparameter setting. Most studies did not report the hyperparameters tested during optimisation, and none reported the optimal hyperparameter settings selected for the final model.

3.4.5.4. Cross validation and level of evidence

Eleven studies performed internal cross-validation: four performed *k*-fold, four performed leave-one-out, and two performed bootstrapping. One study also performed external validation in a randomly partitioned hold-out dataset (Herzog et al., 2021). As such, ten studies provided level 2 evidence and one study provided level 3 evidence.

Six studies did not perform internal cross-validation or external validation and therefore provided level 1 evidence. One of these studies (López-Castro et al., 2021) used the predictors selected in one dataset to make predictions in a second dataset, but repeated parameter estimation (model fitting) in the second dataset, and therefore performed replication rather than external validation. Another study (Zhang et al., 2021) divided the dataset into two cohorts and repeated *k*-means clustering and linear mixed modelling in the second cohort, again performing replication rather than external validation.

3.4.6. Evaluation metrics

Nine studies reported metrics of model prediction accuracy or error. These studies all applied internal cross-validation procedures, but it is important to note that only Herzog et al. (2021) performed external validation, and none had a reasonable number of participants with the outcome. Therefore, model performance metrics were estimated within a training sample of insufficient size, limiting the likelihood that they will generalise to independent samples. None of the studies that sought to predict treatment retention reported evaluation metrics. None of the studies reported calibration.

Among the eight studies that sought to predict a continuous outcome, three reported model prediction accuracy in the form of R^2 or R, and four reported prediction error in the form of root mean squared error (RMSE), mean absolute error (MAE), and true error. Two of these studies reported both accuracy and error, and four studies did not report either. Herzog et al. (2021) used elastic net and reported R^2 = 0.17 (MAE = 0.69, RMSE = 0.91) in the training set (with bootstrap internal-cross validation) and R^2 = 0.16 (MAE = 0.77, RMSE = 0.95) in the hold-out external validation.

Stirman et al. (2021) used elastic net to select predictors and reported $R^2 = 0.39$ (RMSE = 20.28) for prediction with linear regression mean averaged over 1000 repetitions of 10-fold internal cross-validation. Stuke et al. (2021) used principal component analysis to select predictors and reported $R = 0.162$ for prediction with ADAboost regressor and $R = 0.214$ for linear regression (when squared, ADAboost $R^2 = 0.03$ and linear regression $R^2 = 0.05$). Hoeboer et al. (2021) reported RMSE ranging from 4.06 to 7.24 when predicting change on two PTSD measures in two treatment groups (RMSE is referred to as *average error* in the publication and was clarified through correspondence with the author). Deisenhofer et al. (2018) reported *true error* (MAE of factual predictions) of 4.92 in one treatment group and 5.37 in the other.

Among the six studies that sought to predict a categorical outcome, two reported accuracy as raw accuracy or balanced accuracy, and three reported discrimination as area under the receiver operating characteristic curve (AUC-ROC), area under the precision-recall curve (AUC-PR), and/or sensitivity and specificity. Nixon et al. (2021) visually examined AUC-ROC but did not report statistics, and a further two studies did not report evaluation metrics for prediction of categorical outcomes. Held et al. (2022) tested six methods of developing a classification model and found that gradient boosted models produced the best predictions of fast response (AUC-PR = 0.466, AUC-ROC = 0.765) and elastic net produced the best predictions of minimal response (AUC-PR = 0.628, AUC-ROC = 0.826). Zhutovsky et al. (2019) used Gaussian process classifier to predict ≥ 30% reduction in PTSD score from MRI data and reported AUC-ROC = 0.929, balanced accuracy = 81.4%, sensitivity = 84.8%, specificity = 78%. Etkin et al. (2019) predicted ≥50% reduction in PTSD score from verbal memory delayed recall impairment and low within Ventral Attention Network connectivity (MRI) and reported accuracy = 85%, sensitivity = 80%, and specificity = 87% for linear SVM, and accuracy = 90%, sensitivity = 80%, and specificity = 93% for radial basis function SVM, but the sample size was particularly small ($n = 36$), the number of participants with the outcome was not reported, and class imbalance was not addressed.

3.4.7. Predicting differential treatment outcome

Five studies explored interactions between pre-treatment variables and treatment type. Three studies sought to retrospectively predict the optimal treatment for each participant by developing a personalised advantage index (Deisenhofer et al., 2018; Hoeboer et al., 2021; Keefe et al., 2018). Following a method suggested by Kessler et al. (2017), Deisenhofer et al. (2018) and Hoeboer et al. (2021) used ML methods to select predictors for a linear regression model for each treatment under investigation and identified each patient's optimal treatment by comparing the outcomes predicted by the two regression models. Both studies found a significantly greater improvement in symptoms among patients who had received their model indicated optimal treatment. Keefe et al. (2018) used ML methods to select predictors and moderators (i.e., variables that interact with treatment type) for a logistic regression model and found a significantly lower rate of dropout among patients who received their model-indicated optimal treatment.

Stirman et al. (2021) sought to identify patients most likely to benefit from the most efficacious of two treatments, and those for whom treatment type was unlikely to make a difference, by developing a prognostic index (composite predictor) and testing its interaction with treatment type. The interaction explained 39% of the variance in post-treatment PTSD severity. All four of the above studies reported that using ML methods in this way could potentially guide personalised treatment selection for PTSD.

Zhang et al. (2021) investigated whether patients with latent subtypes of PTSD identified via *k*-means of EEG data, and not identifiable through clinical measures or demographic data, responded differentially to two treatments. There was a significant difference in post-treatment severity between the two subtypes, but no interaction with treatment type. Patients in this study were not randomly allocated to treatment and this was not addressed, therefore there is potential confounding by indication (Kyriacou & Lewis, 2016).

4. Discussion

This review aimed to identify and synthesise studies that used ML methods to predict the outcome of psychological therapy for PTSD, and the degree to which these studies adhered to best practice via auditing the methods of the studies against the ML pipeline domains. Through searching four databases and eleven similar systematic reviews, conducting forward and backward citation searches, and contacting the authors of eligible papers, seventeen studies were identified that met the inclusion criteria. Sixteen were published within the previous four years, reflecting a recent surge of interest in ML methods in clinical psychology and psychiatry (Aafjes-van Doorn et al., 2021), driven partly by recent advances in technology and data collection (Jordan & Mitchell, 2015), and the potential applications of ML methods to psychological therapy data (Chekroud et al., 2021). The one exception was published almost 20 years earlier, but this study made no reference to ML and simply used *k*-means to test the reliability of clusters identified via a different clustering method (Forbes et al., 2003).

4.1. Considerations regarding risk of bias

Risk of bias assessments using PROBAST found all studies to be at high risk of bias. Specifically, all studies were rated high risk of bias in the *analysis* domain, primarily due to inadequate sample sizes for model training. Six studies were rated high risk of bias in the *participants* domain for using routinely collected practice data. Moons et al. (2019) suggest that routinely collected data is at higher risk of bias than RCT or prospectively collected data, as equivalent quality controls may not have been implemented. However, archival clinical practice data such as that of NHS Talking Therapies services is an available source of outcome data on a scale seldom seen in psychological therapy research, with treatments implemented with a high degree of standard training and supervision, and this may allow researchers to conveniently address the issue of sample size. More recently, mental health researchers have advocated the use of large electronic health records to optimise clinical prediction models, in view of the sample size limitations of typical clinical trials and the challenges related to data harmonization across clinical trial datasets, which often leads to sparse predictors (Delgadillo & Lutz, 2020; Kessler & Luedtke, 2021). Further, if the aim is to develop a prediction model for use in a particular mental health service, then using data from that same context may boost ecological validity and generalisability. Vieira et al. (2022) comment that using larger, more heterogeneous, naturalistic datasets may produce models with lower prediction accuracy but greater generalisability. Conversely, the finding that trauma related variables may be better predictors of outcome than demographic data presents a problem as many mental health services do not routinely collect this sort of data.

It is important to highlight that PROBAST was not developed to assess ML studies. Some argue that PROBAST may assess ML studies too harshly (Meehan et al., 2022), and others caution that ML methods may be at greater risk of bias under some conditions (Moons et al., 2019; van der Ploeg et al., 2014). Some important features of ML are not assessed by PROBAST, such as hyperparameter selection, which was not reported by seven out of the seventeen studies in this review and can lead to overfitting if performed inappropriately (Delgadillo & Atzil-Slonim, 2022). The inconsistent reporting and application of ML methods identified by this review reiterates the call for specific guidelines and risk of bias assessment tools (Meehan et al., 2022; Vieira et al., 2022), which were under development at the time when this review was conducted (Collins et al., 2021).

4.2. Sample size

The finding that none of the studies reported a sample size calculation is congruent with similar reviews of clinical prediction modelling with ML methods (Aafjes-van Doorn et al., 2021; Balki et al., 2019). Determining an appropriate sample size for a developing a clinical prediction model using ML methods is a complex task that depends on several factors. Riley et al. (2020) recently published guidelines for estimating the required sample size that go beyond the number of outcome events per variable (EPV) and other rules of thumb. However, the appropriate sample size also varies according to the particular ML method, with some methods requiring larger samples to develop stable models (Dalmaijer et al., 2022; Giesemann et al., 2023; Riley et al., 2021; van der Ploeg et al., 2014), and according to the internal cross-validation methods applied. In a simulation study, Vabalas et al. (2019) found that *k*-fold internal cross-validation yielded over-optimistic estimates of accuracy compared to nested *k*-fold and randomly partitioned hold-out validation, and the magnitude of the bias had an inverse relationship with sample size, increasing sharply with sample size $N < 100$. Additionally, Vabalas et al. (2019) found that the bias increased with the number of candidate predictor variables. A commonly applied rule of thumb is that a minimum of ten EPV is required to train a prediction model. However, this is contentious as it is not empirically-based, and Moons et al. (2019) suggest that an EPV of 20 may be more robust. More precisely it is the number of variable parameters in the model that is of interest (i.e., dummy coded categories and interactions between variables each require the estimation of additional parameters), and when the outcome is categorical the number of outcome events refers to the number of participants in the smallest category. Studies in this review did not consistently explicitly report the number of candidate predictor variables tested, and where they did it was unclear whether they were reporting the number of variables or number of parameters.

Notably, the four studies that used neuroimaging data did not report the number of candidate predictor parameters. Analysing neuroimaging data typically requires estimation of many parameters, and therefore a large number of participants with the outcome. However, Zhutovsky et al. (2019) and Etkin et al. (2019) had the two smallest samples, and Etkin et al., (2019) did not report the number of participants with the outcome. All four of these studies identified regions of the brain significantly associated with PTSD treatment response, but with little consensus, and none were externally validated. Etkin et al. (2019) and Zhutovsky et al. (2019) reported accuracies > 80% but this was likely due to overfitting. Similarly, Vieira et al. (2022) found that studies that used neuroimaging data to predict CBT outcomes reported higher accuracy but again had smaller sample sizes, suggesting that the higher estimates of accuracy were due to overfitting. Collecting neuroimaging data is more expensive, time-consuming, and imposes a higher degree of patient burden than collecting questionnaire or patient health record data. This makes the acquisition of an appropriate sample size to analyse high dimensional neuroimaging data even more challenging. Further, this raises doubts about the feasibility of implementing clinical prediction models that require this type of data at scale, particularly in large publicly funded health services.

4.3. External validation

The finding that only one study (Herzog et al., 2021) employed external validation procedures is congruent with recent reviews of prediction modelling in clinical psychology (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021; Meehan et al., 2022; Vieira et al., 2022). Moreover, this study only externally validated the model in a randomly split hold-out sample. Some argue that this is not an optimal form of external validation, as the training set and validation set are subsamples of the same dataset, are likely to be highly correlated, and provide overestimates of model performance (Aafjes-van Doorn et al., 2021; Steyerberg, 2019). Splitting the data by time (temporal validation) or geographic location (geographic validation) is a more stringent test of external validity (Steyerberg, 2019). Further, some studies had the opportunity to externally validate a model in an independent sample but replicated model fitting and reported the statistical significance of predictors instead of evaluating model performance metrics (López-Castro et al., 2021; Zhang et al., 2021). This suggests a reluctance amongst researchers to shift from testing hypotheses and seeking to explain relationships between variables, to developing pragmatic prediction models (Yarkoni & Westfall, 2017). As such, the extent to which any of the prediction models reviewed here would generalise beyond the respective training sample is unclear. Further, the generalisability of a prediction model is limited by the make-up of the training sample, and a number of studies in this review included only male or only female participants, potentially limiting the generalisability of the model.

4.4. Evaluating model performance

Eight studies did not report model performance evaluation metrics, including two that applied internal cross-validation (Keefe et al., 2018; Nixon et al., 2021), and none of the studies examined calibration. Therefore, it is unclear how efficacious and reliable these models are at predicting therapy outcomes for patients with PTSD. Only two studies compared the performance of ML methods to traditional statistical methods: Held et al. (2022) found that five different ML models outperformed logistic regression, but Stuke et al. (2021) found that ordinary linear regression performed slightly better than ADABoost (an ensemble decision tree method). Therefore, it is unclear whether ML methods offer an advantage over traditional statistical methods, particularly as neither of these studies tested prediction models in an independent validation sample. In a systematic comparison of ML methods and logistic regression, Christodoulou et al. (2019) found that ML methods were more accurate than logistic regression, but only when high risk of bias studies were included in the comparison, suggesting that any apparent advantage of ML methods over logistic regression are a product of study bias. However, this review included penalised logistic regressions in the non-ML logistic regression category, and some would consider these ML methods (Bi et al., 2019; Delgadillo & Atzil-Slonim, 2022; Webb et al., 2020). Further research is required to investigate whether the complexity added by ML methods improves clinical prediction accuracy to a meaningful degree and index the true extent of this advantage. Additionally, some ML methods may be better than others at predicting treatment outcomes, but only two studies compared the performance of more than one ML method (Etkin et al., 2019; Held et al., 2022).

Four studies applied supervised ML methods to develop a prediction model, but then entered the predictors into a simpler statistical model to estimate parameters and predict outcomes, thereby forgoing any potential advantages of the ML model. López-Castro et al., (2021) commented that variables selected by random forest were not all significant predictors in Poisson regression and suggest that this may be due to correlation with other variables (multicollinearity). However, Poisson regression also makes assumptions about the distribution of the data and the shape of the relationship between the predictor and outcome variables that random forest does not (Mushagalusa et al., 2022).

4.5. Predictors in final models

The finding that pre-treatment levels of PTSD, depression, and other mental health problems were among the most consistently selected predictors in a final model is congruent with previous systematic reviews and meta-analyses of predictors of PTSD treatment outcome, which found that these variables were associated with worse outcomes (Barawi et al., 2020; Dewar et al., 2020; Kline et al., 2021; Olatunji et al., 2010). Similarly, the finding that clinical variables were more important predictors than demographic variables is congruent with systematic reviews that found inconsistent support for demographic variables as predictors of PTSD treatment outcome (Barawi et al., 2020; Haagen et al., 2015).

4.6. Recommendations for future studies

To achieve the potential for ML methods to improve individual prediction of psychological therapy outcomes for PTSD, it is recommended that future studies demonstrate full adherence to the ML pipeline domains described above. Specifically, this can achieved in the following ways: [1] perform a sample size calculation and acquire a suitably powered dataset; [2] perform multiple imputation of missing data (stratified by treatment group; Zhang et al., 2021) and report data pre-processing in detail; [3] report all hyperparameter setting (using automated grid search or values selected *a priori*); [4] apply internal cross-validation during model development and testing; [5] externally validate (don't repeat model fitting) in a temporally and/or geographically independent samples (Steyerberg, 2019), and evaluate and report accuracy, error, discrimination, and calibration. Additionally, it is recommended that studies compare the performance of multiple ML methods against one another and against the simplest comparable method (e.g., linear regression or logistic regression).

If ML methods are applied in samples that are too small, with no internal cross-validation, and manual hyperparameter tuning, then it is likely that the model will be overfitted to the training data and estimates of model performance will be over-optimistic. Without external validation and calibration, the extent of the optimism and whether the model will generalise is unknown. A recent meta-analysis of ML models found a negative association between study quality and estimates of prediction accuracy, suggesting that poorer quality studies overestimate accuracy (Sajjadian et al., 2021). It is worth noting that for a prediction model to be clinically useful, the model's prediction accuracy does not necessarily need to be high, only better than expert clinical judgement (Ægisdóttir et al., 2006; Cearns et al., 2019). This can be tested in a prospective randomised trial once the external validity of a prediction model has been established (e.g., Delgadillo et al., 2022). The recently published TRIPOD+AI statement (Collins et al., 2024) guides transparent reporting of clinical prediction models, including those that used ML. This may also serve as an additional guide when designing studies.

4.7. Limitations

ML is an umbrella term that encompasses a broad range of methods, and studies do not always use the term "machine learning". Efforts were made to perform as wide a search as possible; nonetheless it is possible that some relevant studies were not found. Further, the distinction between ML and traditional methods is not clearly defined, and it is possible that some methods included in this review would not be considered ML by some, and vice versa (Bi et al., 2019). In line with the pre-registration, only studies published in peer reviewed journals were included. This is common practice in psychological therapy reviews, aids replicability of the search procedures, and reduces the likelihood of inclusion of poor-quality studies (Aafjes-van Doorn et al., 2021). However, some relevant studies may have been excluded for this reason (e.g., Cohen, 2018). This review focussed specifically on the prediction of outcome from pre-treatment or baseline data, in the interest of applying ML methods to predict the optimal treatment for individual patients. However, there are other ways that the application of ML methods could potentially improve PTSD treatment outcomes, for example by providing personalised outcome feedback and recommendations during treatment (Bone et al., 2021; Lutz et al., 2019). EndNote 20 reference management software was used to organise and screen search results, and citationchaser (Haddaway, 2021) used to conduct forward and backward citation searches. However, use of AI assisted systematic review tools, such as Rayyan (https://www.rayyan.ai/) and Covidence (https://www.covidence.org/), may have increased efficiency and accuracy of searching and screening.

4.8. Clinical implications

This systematic review highlights the need for services to critically evaluate clinical prediction models developed using ML before adopting and applying the recommendations into routine practice. In particular, services should consider the sample size, the level of evidence (indicated by the presence of internal and/or external cross-validation procedures), and assessments of calibration and discrimination (Delgadillo & Atzil-Slonim, 2022; Steyerberg, 2019).

4.9. Conclusion

Due to the methodological limitations and omissions of the studies identified by this systematic review, it is unclear whether ML methods offer any advantages over traditional statistical methods at predicting psychotherapy outcomes for PTSD. Studies neglected to recruit a sample of an appropriate size informed by a sample size calculation, report hyperparameter setting, perform internal and external cross-validation, and assess model calibration. Whilst ML methods may have the potential to improve the prediction of treatment outcomes for PTSD, in order to achieve this potential, ML methods need to be applied rigorously and be shown to offer an added benefit over traditional prediction methods.

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

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

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

Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S. C., Kong, D., Moody, A. R., & Tyrrell, P. N. (2019). Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. *Canadian Association of Radiologists Journal*, *70*(4), 344–353. https://doi.org/10.1016/j.carj.2019.06.002

Barawi, K. S., Lewis, C., Simon, N., & Bisson, J. I. (2020). A systematic review of factors associated with outcome of psychological treatments for post-traumatic stress disorder. *European Journal of Psychotraumatology*, *11*(1), 1774240. https://doi.org/10.1080/20008198.2020.1774240

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222–2239. https://doi.org/10.1093/aje/kwz189

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A.-K., Lutz, W., & Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*, *3*(4), e231–e240. https://doi.org/10.1016/S2589-7500(21)00018-2

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Campbell, Sarah. B., Trachik, B., Goldberg, S., & Simpson, Tracy. L. (2020). Identifying PTSD symptom typologies: A latent class analysis. *Psychiatry Research*, *285*, 112779. https://doi.org/10.1016/j.psychres.2020.112779

Campbell-Sills, L., Sun, X., Choi, K. W., He, F., Ursano, R. J., Kessler, R. C., Levey, D. F., Smoller, J. W., Gelernter, J., Jain, S., & Stein, M. B. (2022). Dissecting the heterogeneity of posttraumatic stress disorder: Differences in polygenic risk, stress exposures, and course of PTSD subtypes. *Psychological Medicine*, *52*(15), 3646–3654. https://doi.org/10.1017/S0033291721000428

Catarino, A., Fawcett, J. M., Ewbank, M. P., Bateup, S., Cummins, R., Tablan, V., & Blackwell, A. D. (2022). Refining our understanding of depressive states and state transitions in response to cognitive behavioural therapy using latent Markov modelling. *Psychological Medicine*, *52*(2), 332–341. https://doi.org/10.1017/S0033291720002032

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), 1–12. https://doi.org/10.1038/s41398-019-0607-2

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & 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

Chen, Z. S., Kulkarni, P. (Param), Galatzer-Levy, I. R., Bigio, B., Nasca, C., & Zhang, Y. (2022). Modern views of machine learning for precision psychiatry. *Patterns*, *3*(11), 100602. https://doi.org/10.1016/j.patter.2022.100602

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

Cohen, Z. D. (2018). *Treatment Selection: Understanding What Works for Whom in Mental Health* [Ph.D., University of Pennsylvania]. https://www.proquest.com/docview/2117235776/abstract/3CA5132CE7244197PQ/1

Cohen, Z. D., Delgadillo, J., & DeRubeis, R. J. (2021). Personalized treatment approaches. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 673–704). Wiley.

Collins, G. S., Dhiman, P., Navarro, C. L. A., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Calster, B. V., Smeden, M. van, Riley, R. D., & Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, *11*(7), e048008. https://doi.org/10.1136/bmjopen-2020-048008

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., Ghassemi, M., Liu, X., Reitsma, J. B., Smeden, M. van, Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*, *23*(1), 205. https://doi.org/10.1186/s12859-022-04675-1

Deisenhofer, A.-K., Delgadillo, J., Rubel, J. A., Bohnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, *35*(6), 541–550. https://doi.org/10.1002/da.22755

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

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z. D., DeRubeis, R. J., & Barkham, M. (2022). Stratified Care vs Stepped Care for Depression: A Cluster Randomized Clinical Trial. *JAMA Psychiatry*, *79*(2), 101–108. https://doi.org/10.1001/jamapsychiatry.2021.3539

Delgadillo, J., & Atzil-Slonim, D. (2022). Artificial intelligence, machine learning and mental health. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier. https://eprints.whiterose.ac.uk/197827/

Delgadillo, J., & Lutz, W. (2020). A Development Pathway Towards Precision Mental Health Care. *JAMA Psychiatry*, *77*(9), 889–890. https://doi.org/10.1001/jamapsychiatry.2020.1048

Department of Veterans Affairs and Department of Defense (VA/DoD). (2017). *VA/DoD clinical practice guideline for the management of posttraumatic stress disorder and acute stress disorder*. Author. https://www.healthquality.va.gov/guidelines/MH/ptsd/VADoDPTSDCPGFinal012418.pdf

Department of Veterans Affairs and Department of Defense (VA/DoD). (2023). *VA/DoD clinical practice guideline for the management of posttraumatic stress disorder and acute stress disorder (Version 4.0)*. Author. https://www.healthquality.va.gov/guidelines/MH/ptsd/VADoDPTSDCPGFinal012418.pdf

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. *The Canadian Journal of Psychiatry*, *65*(2), 71–86. https://doi.org/10.1177/0706743719875602

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Etkin, A., Maron-Katz, A., Wu, W., Fonzo, G. A., Huemer, J., Vertes, P. E., Patenaude, B., Richiardi, J., Goodkind, M. S., Keller, C. J., Ramos-Cejudo, J., Zaiko, Y. V., Peng, K. K., Shpigel, E., Longwell, P., Toll, R. T., Thompson, A., Zack, S., Gonzalez, B., … O'Hara, R. (2019). Using fMRI connectivity to define a treatment-resistant form of post-traumatic stress disorder. *Science Translational Medicine*, *11*(486), 1. https://doi.org/10.1126/scitranslmed.aal3236

Fear, N. T., Bridges, S., Hatch, S., Hawkins, V., & Wessely, S. (2016). Chapter 4: Posttraumatic stress disorder. In McManus S, Bebbington P, Jenkins R, & Brugha T (Eds.), *Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. NHS Digital.

Fleming, C. E., Kholodkov, T., Dillon, K. H., Belvet, B., & Crawford, E. F. (2018). Actuarial prediction of psychotherapy retention among Iraq-Afghanistan veterans with posttraumatic stress disorder. *Psychol Serv*, *15*(2), 172–180. https://doi.org/10.1037/ser0000139

Foa, E. B., Dancu, C. V., Hembree, E. A., Jaycox, L. H., Meadows, E. A., & Street, G. P. (1999). A comparison of exposure therapy, stress inoculation training, and their combination for reducing posttraumatic stress disorder in female assault victims. *Journal of Consulting and Clinical Psychology, 67*(2), 194–200. https://doi.org/10.1037/0022-006X.67.2.194

Forbes, D., Creamer, M., Allen, N., Elliott, P., McHugh, T., Debenham, P., & Hopwood, M. (2003). MMPI-2 Based Subgroups of Veterans with Combat-related PTSD: Differential Patterns of Symptom Change After Treatment. *The Journal of Nervous and Mental Disease*, *191*(8), 531–537. https://doi.org/10.1097/01.nmd.0000082181.79051.83

Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 Ways to Have Posttraumatic Stress Disorder. *Perspectives on Psychological Science*, *8*(6), 651–662. https://doi.org/10.1177/1745691613504115

Galovski, T. E., Blain, L. M., Mott, J. M., Elwood, L., & Houle, T. (2012). Manualized therapy for PTSD: Flexing the structure of cognitive processing therapy. *Journal of Consulting and Clinical Psychology*, *80*(6), 968–981. https://doi.org/10.1037/a0030600

Galovski, T. E., Harik, J. M., Blain, L. M., Elwood, L., Gloth, C., & Fletcher, T. D. (2016). Augmenting cognitive processing therapy to improve sleep impairment in PTSD: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *84*(2), 167–177. https://doi.org/10.1037/ccp0000059

Giesemann, J., Delgadillo, J., Schwartz, B., Bennemann, B., & Lutz, W. (2023). Predicting dropout from psychological treatment using different machine learning algorithms, resampling methods, and sample sizes. *Psychotherapy Research*, *33*(6), 683–695. https://doi.org/10.1080/10503307.2022.2161432

Glaz, A. L., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

Haagen, J. F., Smid, G. E., Knipscheer, J. W., & Kleber, R. J. (2015). The efficacy of recommended treatments for veterans with PTSD: A metaregression analysis. *Clinical Psychology Review*, *40*, 184–194. https://dx.doi.org/10.1016/j.cpr.2015.06.008

Haddaway, N. R. (2021). *citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching*. https://estech.shinyapps.io/citationchaser/

Hahn, T., Nierenberg, A. A., & Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: Applications, guidelines, challenges and perspectives. *Molecular Psychiatry*, *22*(1), 37–43. https://doi.org/10.1038/mp.2016.201

Hamblen, J. L., Norman, S. B., Sonis, J. H., Phelps, A. J., Bisson, J. I., Nunes, V. D., Megnin-Viggars, O., Forbes, D., Riggs, D. S., & Schnurr, P. P. (2019). A guide to guidelines for the treatment of posttraumatic stress disorder in adults: An update. *Psychotherapy*, *56*(3), 359–373. https://doi.org/10.1037/pst0000231

Held, P., Schubert, R. A., Pridgen, S., Kovacevic, M., Montes, M., Christ, N. M., Banerjee, U., & Smith, D. L. (2022). 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.

Hendriks, L., De Kleine, R. A., Broekman, T. G., Hendriks, G.-J., & Van Minnen, A. (2018). Intensive prolonged exposure therapy for chronic PTSD patients following multiple trauma and multiple treatment attempts. *European Journal of Psychotraumatology*, *9*(1). https://doi.org/10.1080/20008198.2018.1425574

Herzog, P., & Kaiser, T. (2022). Is it worth it to personalize the treatment of PTSD? – A variance-ratio meta-analysis and estimation of treatment effect heterogeneity in RCTs of PTSD. *Journal of Anxiety Disorders*, *91*, 102611. https://doi.org/10.1016/j.janxdis.2022.102611

Herzog, P., Voderholzer, U., Gärtner, T., Osen, B., Svitak, M., Doerr, R., Rolvering-Dijkstra, M., Feldmann, M., Rief, W., & Brakemeier, E. L. (2021). Predictors of outcome during inpatient psychotherapy for posttraumatic stress disorder: A single-treatment, multi-site, practice-based study. *Psychother Res*, *31*(4), 468–482. https://doi.org/10.1080/10503307.2020.1802081

Hien, D. A., Levin, F. R., Ruglass, L. M., López-Castro, T., Papini, S., Hu, M.-C., Cohen, L. R., & Herron, A. (2015). Combining seeking safety with sertraline for PTSD and alcohol use disorders: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 83*(2), 359–369. https://doi.org/10.1037/a0038719

Hoeboer, C. M., Oprel, D. A. C., De Kleine, R. A., Schwartz, B., Deisenhofer, A. K., Schoorl, M., Van Der Does, W. A. J., van Minnen, A., & Lutz, W. (2021). Personalization of treatment for patients with childhoodabuse-related posttraumatic stress disorder. *Journal of Clinical Medicine*, *10*(19). https://doi.org/10.3390/jcm10194522

Horn, S. R., Pietrzak, R. H., Schechter, C., Bromet, E. J., Katz, C. L., Reissman, D. B., Kotov, R., Crane, M., Harrison, D. J., Herbert, R., Luft, B. J., Moline, J. M., Stellman, J. M., Udasin, I. G., Landrigan, P. J., Zvolensky, M. J., Southwick, S. M., & Feder, A. (2016). Latent typologies of posttraumatic stress disorder in World Trade Center responders. *Journal of Psychiatric Research*, *83*, 151–159. https://doi.org/10.1016/j.jpsychires.2016.08.018

International Society for Traumatic Stress Studies (ISTSS). (2018). *ISTSS PTSD prevention and treatment guidelines: Methodology and recommendations*. http://www.istss.org/getattachment/TreatingTrauma/New-ISTSS-Prevention-and-Treatment-Guidelines/ISTSS_ PreventionTreatmentGuidelines_FNL-March-19-2019.pdf.aspx

Jericho, B., Luo, A., & Berle, D. (2021). Trauma-focused psychotherapies for post-traumatic stress disorder: A systematic review and network meta-analysis. *Acta Psychiatrica Scandinavica*. https://doi.org/10.1111/acps.13366

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Karatzias, T., Hyland, P., Bradley, A., Cloitre, M., Roberts, N. P., Bisson, J. I., & Shevlin, M. (2019). Risk factors and comorbidity of ICD-11 PTSD and complex PTSD: Findings from a trauma-exposed population based sample of adults in the United Kingdom. *Depression and Anxiety*, *36*(9), 887–894. https://doi.org/10.1002/da.22934

Keefe, J. R., Stirman, S. W., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and Anxiety*, *35*(4), 330–338. https://doi.org/10.1002/da.22731

Kessler, R. C., Loo, H. M. van, Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., Jonge, P. de, Nierenberg, A. A., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36. https://doi.org/10.1017/S2045796016000020

Kessler, R. C., & Luedtke, A. (2021). Pragmatic Precision Psychiatry—A New Direction for Optimizing Treatment Selection. *JAMA Psychiatry*, *78*(12), 1384–1390. https://doi.org/10.1001/jamapsychiatry.2021.2500

Kline, A. C., Cooper, A. A., Rytwinski, N. K., & Feeny, N. C. (2021). The Effect of Concurrent Depression on PTSD Outcomes in Trauma-Focused Psychotherapy: A Meta-Analysis of Randomized Controlled Trials. *Behavior Therapy*, *52*(1), 250–266. https://doi.org/10.1016/j.beth.2020.04.015

Koenen, K. C., Ratanatharathorn, A., Ng, L., McLaughlin, K. A., Bromet, E. J., Stein, D. J., Karam, E. G., Ruscio, A. M., Benjet, C., Scott, K., Atwoli, L., Petukhova, M., Lim, C. C. W., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Bunting, B., Ciutan, M., Girolamo, G. de, … Kessler, R. C. (2017). Posttraumatic stress disorder in the World Mental Health Surveys. *Psychological Medicine*, *47*(13), 2260–2274. https://doi.org/10.1017/S0033291717000708

Kratzer, L., Heinz, P., Schennach, R., Schiepek, G. K., Padberg, F., & Jobst, A. (2019). Inpatient Treatment of Complex PTSD Following Childhood Abuse: Effectiveness and Predictors of Treatment Outcome. *PPmP Psychotherapie Psychosomatik Medizinische Psychologie*, *69*(3–4), 114–122. https://doi.org/10.1055/a-0591-3962

Kyriacou, D. N., & Lewis, R. J. (2016). Confounding by Indication in Clinical Research. *JAMA*, *316*(17), 1818–1819. https://doi.org/10.1001/jama.2016.16435

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Lewis, C., Roberts, N. P., Andrew, M., Starling, E., & Bisson, J. I. (2020). Psychological therapies for post-traumatic stress disorder in adults: Systematic review and meta-analysis. *European Journal of Psychotraumatology*, *11*(1), 1729633. https://doi.org/10.1080/20008198.2020.1729633

Lewis, C., Roberts, N. P., Gibson, S., & Bisson, J. I. (2020). Dropout from psychological therapies for post-traumatic stress disorder (PTSD) in adults: Systematic review and meta-analysis. *European Journal of Psychotraumatology*, *11*(1), 1709709. https://doi.org/10.1080/20008198.2019.1709709

López-Castro, T., Zhao, Y., Fitzpatrick, S., Ruglass, L. M., & Hien, D. A. (2021). Seeing the forest for the trees: Predicting attendance in trials for co-occurring PTSD and substance use disorders with a machine learning approach. *Journal of Consulting and Clinical Psychology*, *89*(10), 869–884. https://doi.org/10.1037/ccp0000688

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A.-K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

Malejko, K., Abler, B., Plener, P. L., & Straub, J. (2017). Neural Correlates of Psychotherapeutic Treatment of Post-traumatic Stress Disorder: A Systematic Literature Review. *Frontiers in Psychiatry*, *8*, 85. https://doi.org/10.3389/fpsyt.2017.00085

Malgaroli, M., & Schultebraucks, K. (2021). Artificial Intelligence and Posttraumatic Stress Disorder (PTSD). *European Psychologist*, *25*(4), 272–282. https://doi.org/10.1027/1016-9040/a000423

Manchia, M., Pisanu, C., Squassina, A., & Carpiniello, B. (2020). Challenges and Future Prospects of Precision Medicine in Psychiatry. *Pharmacogenomics and Personalized Medicine*, *13*, 127–140.

Martin, A., Naunton, M., Kosari, S., Peterson, G., Thomas, J., & Christenson, J. K. (2021). Treatment Guidelines for PTSD: A Systematic Review. *Journal of Clinical Medicine*, *10*(18), 4175. https://doi.org/10.3390/jcm10184175

Mavranezouli, I., Megnin-Viggars, O., Daly, C., Dias, S., Welton, N. J., Stockton, S., Bhutani, G., Grey, N., Leach, J., Greenberg, N., Katona, C., El-Leithy, S., & Pilling, S. (2020). Psychological treatments for post-traumatic stress disorder in adults: A network meta-analysis. *Psychological Medicine*, *50*(4), 542–555. https://doi.org/10.1017/S0033291720000070

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry*, *27*(6), Article 6. https://doi.org/10.1038/s41380-022-01528-4

Merz, J., Schwarzer, G., & Gerger, H. (2019). Comparative Efficacy and Acceptability of Pharmacological, Psychotherapeutic, and Combination Treatments in Adults with Posttraumatic Stress Disorder: A Network Meta-analysis. *JAMA Psychiatry*, *76*(9), 904–913. https://doi.org/10.1001/jamapsychiatry.2019.0951

Moons, K. G. M., de Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, *170*(1), 1–33. https://doi.org/10.7326/M18-1377

Mushagalusa, C. A., Fandohan, A. B., & Glèlè Kakaï, R. (2022). Random Forests in Count Data Modelling: An Analysis of the Influence of Data Features and Overdispersion on Regression Performance. *Journal of Probability and Statistics*, *2022*, 1–21. https://doi.org/10.1155/2022/2833537

Najavits, L. M. (2015). The problem of dropout from "gold standard" PTSD therapies. *F1000Prime Reports*, *7*, 43. https://doi.org/10.12703/P7-43

National Institute for Health and Care Excellence [NICE]. (2018). *Post-traumatic stress disorder (NICE guideline [NG116])*. NICE. https://www.nice.org.uk/guidance/ng116

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

Olatunji, B. O., Cisler, J. M., & Tolin, D. F. (2010). A meta-analysis of the influence of comorbidity on treatment outcome in the anxiety disorders. *Clinical Psychology Review*, *30*(6), 642–654. https://dx.doi.org/10.1016/j.cpr.2010.04.008

Oprel, D. A. C., Hoeboer, C. M., Schoorl, M., Kleine, R. A. de, Cloitre, M., Wigard, I. G., … van der Does, W. (2021). Effect of Prolonged Exposure, intensified Prolonged Exposure and STAIR+Prolonged Exposure in patients with PTSD related to childhood abuse: a randomized controlled trial. *European Journal of Psychotraumatology*, *12*(1). https://doi.org/10.1080/20008198.2020.1851511

Pacella, M. L., Hruska, B., & Delahanty, D. L. (2013). The physical health consequences of PTSD and PTSD symptoms: A meta-analytic review. *Journal of Anxiety Disorders*, *27*(1), 33–46. https://doi.org/10.1016/j.janxdis.2012.08.004

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Palazón-Bru, A., Martín-Pérez, F., Mares-García, E., Beneyto-Ripoll, C., Gil-Guillén, V. F., Pérez-Sempere, Á., & Carbonell-Torregrosa, M. Á. (2020). A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models. *Statistics in Medicine*, *39*(23), 3207–3225. https://doi.org/10.1002/sim.8660

Phoenix Australia Centre for Posttraumatic Mental Health. (2021). *Australian Guidelines for the Prevention and Treatment of Acute Stress Disorder, Posttraumatic Stress Disorder and Complex PTSD*. Author. https://www.phoenixaustralia.org/australian-guidelines-for-ptsd/

Pietrzak, R. H., el-Gabalawy, R., Tsai, J., Sareen, J., Neumeister, A., & Southwick, S. M. (2014). Typologies of posttraumatic stress disorder in the U.S. adult population. *Journal of Affective Disorders*, *162*, 102–106. https://doi.org/10.1016/j.jad.2014.03.024

Ramos-Lima, L. F., Waikamp, V., Antonelli-Salgado, T., Passos, I. C., & Freitas, L. H. M. (2020). The use of machine learning techniques in trauma-related disorders: A systematic review. *Journal of Psychiatric Research*, *121*, 159–172. https://doi.org/10.1016/j.jpsychires.2019.12.001

Ravi, M., Powers, A., Rothbaum, B. O., Stevens, J. S., & Michopoulos, V. (2023). Neighborhood poverty prospectively predicts PTSD symptoms six-months following trauma exposure. *Mental Health Science*, *1*(4), 213–221. https://doi.org/10.1002/mhs2.35

Resick, P. A. (2001). Cognitive therapy for posttraumatic stress disorder. *Journal of Cognitive Psychotherapy*, *15*(4), 321. https://doi.org/10.1891/0889-8391.15.4.321

Resick, P. A., Galovski, T. E., Uhlmansiek, M. O., Scher, C. D., Clum, G. A., & 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

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & Smeden, M. van. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, *132*, 88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005

Robinson, L., Delgadillo, J., & Kellett, S. (2020). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research*, *30*(1), 79–96. https://doi.org/10.1080/10503307.2019.1566676

Ruglass, L. M., Lopez-Castro, T., Papini, S., Killeen, T., Back, S. E., & Hien, D. A. (2017). Concurrent treatment with prolonged exposure for co-occurring full or subthreshold posttraumatic stress disorder and substance use disorders: A randomized clinical trial. *Psychotherapy and psychosomatics*, *86*(3), 150-161. https://doi.org/10.1159/000462977

Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., Strother, S. C., Farzan, F., Kennedy, S. H., & Uher, R. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine*, *51*(16), 2742–2751. https://doi.org/10.1017/S0033291721003871

Schnurr, P. P., Friedman, M. J., Engel, C. C., Foa, E. B., Shea, M. T., Chow, B. K., ... & Bernardy, N. (2007). Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *Jama*, *297*(8), 820-830. https://doi.org/10.1001/jama.297.8.820

Schottenbauer, M. A., Glass, C. R., Arnkoff, D. B., Tendick, V., & Gray, S. H. (2008). Nonresponse and Dropout Rates in Outcome Studies on PTSD: Review and Methodological Considerations. *Psychiatry: Interpersonal and Biological Processes*, *71*(2), 134–168. https://doi.org/10.1521/psyc.2008.71.2.134

Shalev, A., Liberzon, I., & Marmar, C. (2017). Post-Traumatic Stress Disorder. *New England Journal of Medicine*, *376*(25), 2459–2469. https://doi.org/10.1056/NEJMra1612499

Shekelle, P. G., Ortiz, E., Rhodes, S., Morton, S. C., Eccles, M. P., Grimshaw, J. M., & Woolf, S. H. (2001). Validity of the Agency for Healthcare Research and Quality Clinical Practice GuidelinesHow Quickly Do Guidelines Become Outdated? *JAMA*, *286*(12), 1461–1467. https://doi.org/10.1001/jama.286.12.1461

Simiola, V., Neilson, E. C., Thompson, R., & Cook, J. M. (2015). Preferences for trauma treatment: A systematic review of the empirical literature. *Psychological Trauma: Theory, Research, Practice, and Policy*, *7*(6), 516–524. https://doi.org/10.1037/tra0000038

Simmonds-Buckley, M., Catarino, A., & Delgadillo, J. (2021). Depression subtypes and their response to cognitive behavioral therapy: A latent transition analysis. *Depression and Anxiety*, *38*(9), 907–916. https://doi.org/10.1002/da.23161

Steenkamp, M. M., Litz, B. T., Hoge, C. W., & Marmar, C. R. (2015). Psychotherapy for Military-Related PTSD: A Review of Randomized Clinical Trials. *JAMA*, *314*(5), 489–500. https://doi.org/10.1001/jama.2015.8370

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer Nature. https://link.springer.com/book/10.1007/978-3-030-16399-0

Stirman, S., Cohen, Z., Lunney, C., DeRubeis, R., Wiley, J., & Schnurr, P. (2021). A personalized index to inform selection of a trauma-focused or non-trauma-focused treatment for PTSD. *Behaviour Research and Therapy*, *142*. https://doi.org/10.1016/j.brat.2021.103872

Stuke, H., Schoofs, N., Johanssen, H., Bermpohl, F., Ülsmann, D., Schulte-Herbrüggen, O., & Priebe, K. (2021). Predicting outcome of daycare cognitive behavioural therapy in a naturalistic sample of patients with PTSD: a machine learning approach. *European Journal of Psychotraumatology*, *12*(1). https://doi.org/10.1080/20008198.2021.1958471

Swift, J. K., Greenberg, R. P., Tompkins, K. A., & Parkin, S. R. (2017). Treatment refusal and premature termination in psychotherapy, pharmacotherapy, and their combination: A meta-analysis of head-to-head comparisons. *Psychotherapy*, *54*(1), 47–57. https://doi.org/10.1037/pst0000104

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 137. https://doi.org/10.1186/1471-2288-14-137

Vieira, S., Liang, X., Guiomar, R., & Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review*, 102193.

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. https://doi.org/10.1037/ccp0000451

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yehuda, R., Hoge, C. W., McFarlane, A. C., Vermetten, E., Lanius, R. A., Nievergelt, C. M., Hobfoll, S. E., Koenen, K. C., Neylan, T. C., & Hyman, S. E. (2015). Post-traumatic stress disorder. *Nature Reviews Disease Primers*, *1*(1), 1–22.

Zhang, Y., Wu, W., Toll, R. T., Naparstek, S., Maron-Katz, A., Watts, M., Gordon, J., Jeong, J., Astolfi, L., Shpigel, E., Longwell, P., Sarhadi, K., El-Said, D., Li, Y., Cooper, C., Chin-Fatt, C., Arns, M., Goodkind, M. S., Trivedi, M. H., … Etkin, A. (2021). Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography. *Nature Biomedical Engineering*, *5*(4), 309–323. https://doi.org/10.1038/s41551-020-00614-8

Zhutovsky, P., Thomas, R., Olff, M., van Rooij, S., Kennis, M., van Wingen, G., & Geuze, E. (2019). Individual prediction of psychotherapy outcome in posttraumatic stress disorder using neuroimaging data. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0663-7

Zilcha-Mano, S., Zhu, X., Suarez-Jimenez, B., Pickover, A., Tal, S., Such, S., Marohasy, C., Chrisanthopoulos, M., Salzman, C., Lazarov, A., Neria, Y., & Rutherford, B. R. (2020). Diagnostic and Predictive Neuroimaging Biomarkers for Posttraumatic Stress Disorder.

Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, *5*(7), 688–696. https://doi.org/10.1016/j.bpsc.2020.03.010

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Figure 1

PRISMA flow diagram

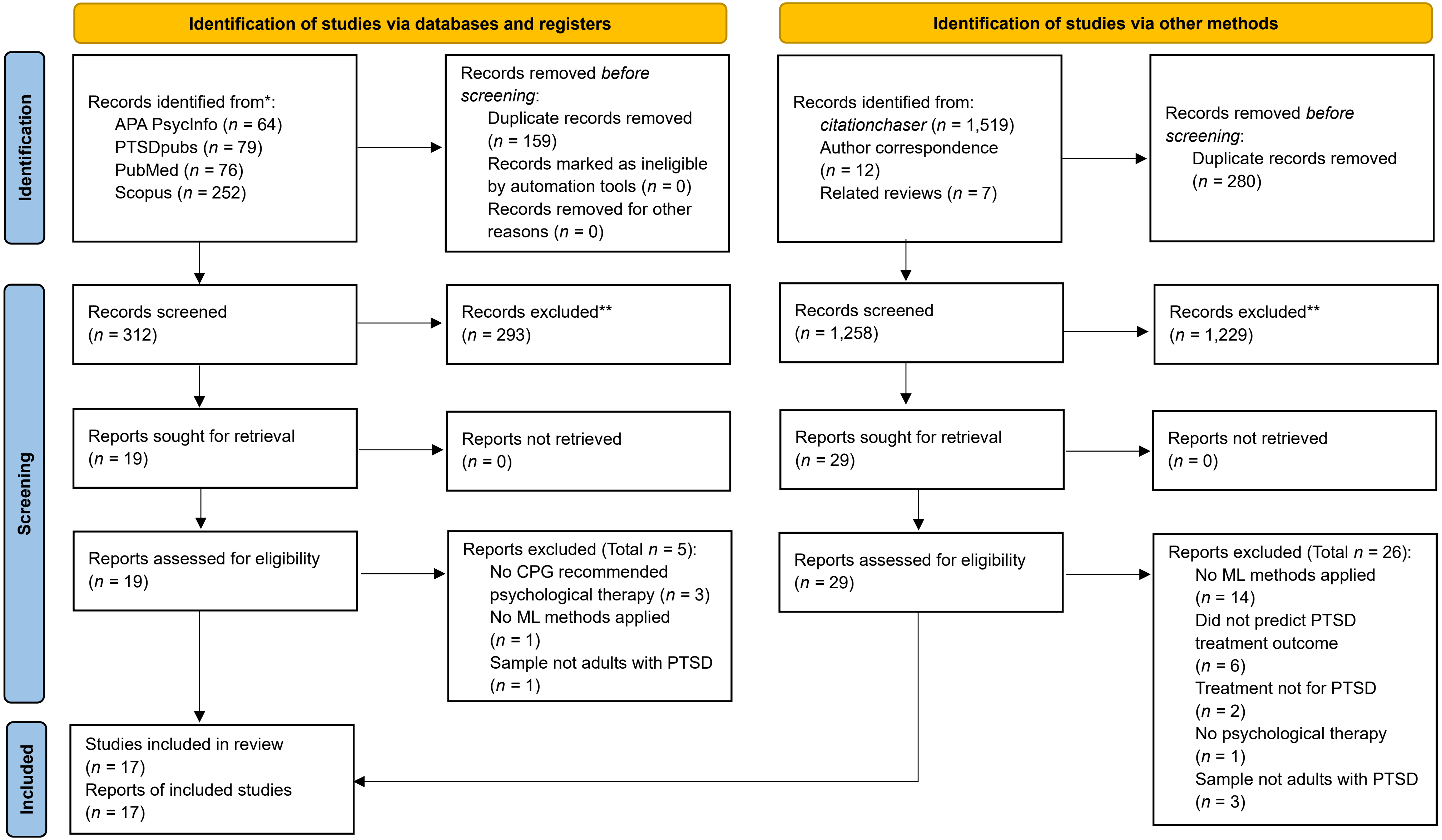
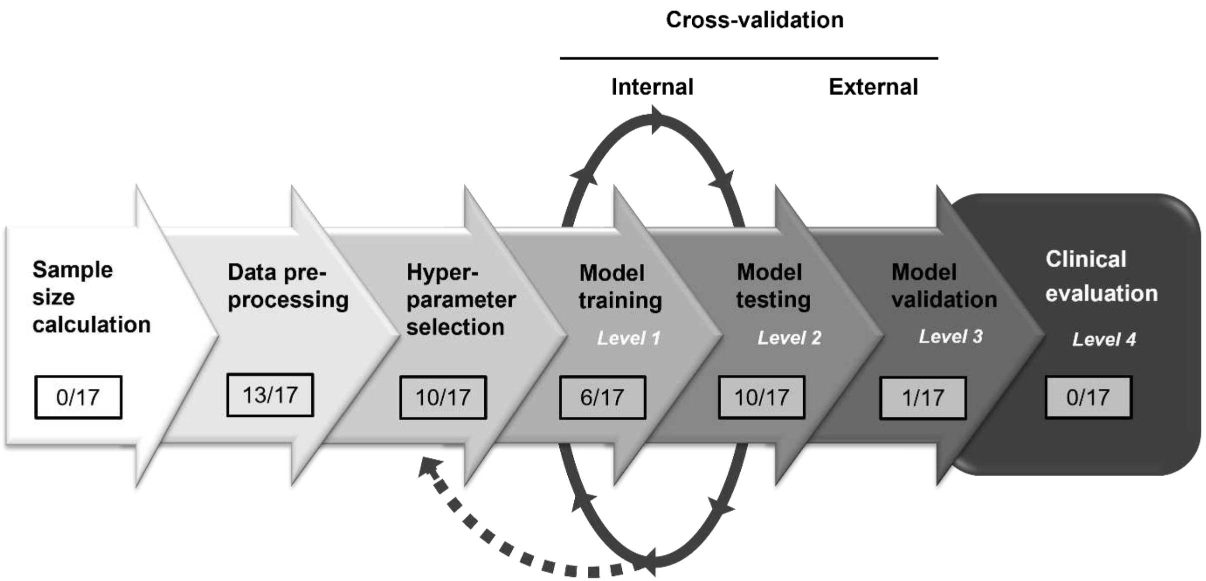


Figure 2.2

Proportion of studies that reported each step of the machine learning pipeline



Note. Figure adapted from Delgadillo and Atzil-Slonim (2022)

Table 1

Inclusion and exclusion criteria

	Inclusion Criteria	***Exclusion Criteria***
Population	Adults (aged 18 and over) who received clinical practice guideline recommended psychological therapy for current PTSD.	Children and adolescents under the age of 18. Adults receiving treatment for a condition other than PTSD.
Intervention	Evidence-based psychological therapies recommended for the treatment of current symptoms of PTSD in adults by current clinical practice guidelines.	Psychological therapy intended to treat a different condition. Psychological therapy intended to prevent the onset or relapse of PTSD. Pharmacological therapy. Non-psychological therapy (e.g., acupuncture or yoga). Psychological therapy not recommended by clinical practice guidelines. (If any of the above were delivered alongside or in comparison to an intervention that met the inclusion criteria then that study would be included.)
Outcome to be predicted	Continuous or categorical outcomes of psychotherapy for PTSD, including remission, change in symptoms, dropout, and retention.	Future onset or relapse of PTSD. Current presence (diagnosis) of PTSD.

Time span of prediction	From pre-treatment to post-treatment. The outcome timepoint of interest is the end of treatment, or the follow-up nearest to the end of treatment.	
Intended moment of model use	Initial patient assessment, prior to the start of treatment.	During or after treatment.
Modelling approach	Prognostic models that applied supervised or unsupervised machine learning methods in the prediction of treatment outcomes from patients' pre-treatment or baseline features.	Diagnostic models that predict the presence of PTSD. Prognostic models that predict onset of PTSD. Modelling approaches that did not use any machine-learning methods.
Scope/intended purpose of models	To guide clinical decision-making and treatment planning.	

Note. PTSD = post-traumatic stress disorder.

Table 2

Study characteristics

Study	Data source	Population (total sample N)	Setting (country)	Treatment (group n)	Treatment duration
Deisenhofer et al. (2018)	Routine clinical practice (Retrospective)	Adults with PTSD (317)	NHS primary care outpatient mental health service (England)	Tf-CBT (242) EMDR (75)	≤ 20 weekly sessions (Session duration not reported)
Etkin et al. (2019)	RCT (Prospective)	Adults with PTSD (76)	University (U.S.A.)	PE (36) Wait-list control (30)	9 or 12 weekly or twice-weekly 90-minute sessions
Fleming et al. (2018)	Routine clinical practice (Retrospective)	Military veterans with PTSD (124)	Veterans Affairs speciality outpatient clinic (U.S.A.)	PE (49) CPT (53) Opted out of psychological therapy following introductory psychoeducation session (22)	Mean (SD) *n* sessions attended = 6.78 (7.03) (Session duration not reported)
Forbes et al. (2003)	Routine clinical practice (Retrospective)	Military veterans with PTSD (166)	Veterans PTSD treatment programme (Australia)	Group and individual therapy, primarily cognitive-behavioural in orientation, with trauma-focussed sessions (166)	16 sessions of individual therapy over 12 weeks (4 weeks inpatient, 8 weeks outpatient) (Session duration not reported)
Held et al. (2022)	Cohort study (Prospective)	Military veterans with PTSD (502)	University Medical Centre Intensive Outpatient Treatment Program (U.S.A.)	CPT based intensive PTSD treatment program (502)	14 once-daily 50-minute sessions of individual CPT over 3 weeks
Hendriks et al. (2018)	Cohort study (Prospective)	Adults with PTSD and history of multiple	Outpatient mental health clinic (Netherlands)	Intensive PE (73)	12 sessions over 4 days within 1 week (4.5 hours per-day), followed

Study	Data source	Population (total sample N)	Setting (country)	Treatment (group n)	Treatment duration
		interpersonal traumas (73)			by 4 weekly 90-minute booster sessions with homework
Herzog et al. (2021)	Routine clinical practice (Retrospective)	Adults with PTSD (612)	Five specialised inpatient clinics (Germany)	Individual exposure therapy (PE, IRRT, or EMDR), plus group Tf-CBT and a range of supplementary psychological and non-psychological therapies (612)	8 to 10 weeks, 1 hour per week individual exposure therapy, 8 hours per week of group Tf-CBT, plus an average of 11 hours per week of multimodal and transdiagnostic interventions (total 152-200 therapy hours) Sample mean (SD, range) length of stay (days) = 54.3 (15.5, 6 - 98)
Hoeboer et al. (2021)	RCT (Retrospective; Oprel et al., 2021)	Adults with childhood-abuse-related PTSD (149)	Two specialist outpatient mental health services (Netherlands)	PE (48) Intensified PE (51) STAIR+PE (50)	PE: 16 weekly 90-minute sessions Intensified PE: Three PE sessions per-week for 4 weeks, followed by booster PE sessions after 1 month and 2 months (total 14 sessions) STAIR+PE: Eight sessions of STAIR followed by eight sessions of PE

Study	Data source	Population (total sample N)	Setting (country)	Treatment (group n)	Treatment duration
Keefe et al. (2018)	RCT (Retrospective; Resick et al., 2002)	Women with rape-trauma PTSD (160)	(U.S.A.)	CPT (79) PE (81)	Total 13 hours for each treatment over 6 weeks CPT: 12 sessions of 50-60 minutes, with 30 minutes added to each of the two writing exposure sessions (sessions 4 and 5) PE: Nine sessions; one 60-minute initial session followed by eight 90-minute sessions
Kratzer et al. (2019)	Routine clinical practice (Retrospective)	Inpatients with complex PTSD following childhood physical and childhood sexual abuse (150)	Specialist inpatient clinic (Germany)	Tf-CBT, often with integrated exposure and EMDR. Patients also offered group psychotherapies. (150)	≤ 20 individual psychotherapy sessions of 75-minutes each
López-Castro et al. (2021)	RCT (Retrospective; Ruglass et al., 2017; Hien et al., 2015)	Adults with PTSD and SUD (130)	Community based outpatient mental-health treatment programme (U.S.A.)	Sample 1 (Ruglass et al., 2017): 1. COPE (33) 2. RPT (37) Sample 2 (Hien et al., 2015):	Sample 1 (Ruglass et al., 2017): All participants were offered 12 weekly 90-min individual sessions Sample 2 (Hien et al., 2015): All participants were offered 12

Study	Data source	Population (total sample N)	Setting (country)	Treatment (group n)	Treatment duration
				1. Seeking Safety plus placebo (29) 2. Seeking Safety plus ADM (31)	weekly 60-min individual psychotherapy sessions, and ADM (sertraline) dosage started on 50 mg/day and increased up to 200 mg/day over 2 weeks throughout the active study period
Nixon et al. (2021)	RCT (Retrospective; Galovski et al., 2012, Galovski, Harik, Blain, Elwood, et al., 2016; Resick et al., 2002, 2008)	Female interpersonal trauma survivors (216)	Community (U.S.A.)	CPT (216)	12 weekly or bi-weekly 60-min sessions
Stirman et al. (2021)	RCT (Retrospective; Schnurr et al., 2007)	Female military veterans and active-duty service members with PTSD (267)	Nine VA medical centres, two VA readjustment counselling centres, and a military hospital (U.S.A.)	PE (135) Present-Centred Therapy (132)	10 weekly 90-minute sessions
Stuke et al. (2021)	Routine clinical practice (Retrospective)	Adults with PTSD (209)	Specialist day clinic (Germany)	CBT based day-care programme including individual CPT (209)	Four sessions per-week of individual CPT, plus group trauma-focussed therapy 5 days per-week, for a mean of 8.59 weeks (SD = 1.4)

Study	Data source	Population (total sample N)	Setting (country)	Treatment (group n)	Treatment duration
					(Session duration not reported)
Zhang et al. (2021)	Cohort study (Prospective); non-randomised clinical trial (Prospective)	Military veterans with PTSD (241); trauma-exposed controls (95)	University; Veterans Affairs PSTD clinic (U.S.A.)	PE or CPT (135)	Based on manualised protocols (Foa et al., 1999; Resick, 2001) (Number of sessions and session duration not reported)
Zhutovsky et al. (2019)	Cohort study (Prospective)	Male military veterans with PTSD (57); combat-exposed controls (29)	Four military mental-healthcare outpatient clinics (Netherlands)	Tf-CBT (8) EMDR (28) Tf-CBT+EMDR (8)	Mean (SD) number of treatment sessions: Responders = 9.86 (6.29) Non-responders = 10.05 (4.22) (Session duration not reported)
Zilcha-Mano et al. (2020)	Cohort study (Retrospective)	Adults with PTSD (51); adults with PTSD and depression (52); trauma-exposed controls (76)	State Psychiatric Institute (U.S.A.)	PE (55)	10-week standard PE protocol (Session duration not reported)

Note. ADM = anti-depressant medication; CBT = cognitive behavioural therapy; COPE = concurrent treatment for substance use disorder and post-traumatic stress disorder combining prolonged exposure and relapse prevention therapy; CPT = cognitive processing therapy; EMDR = eye movement desensitization and reprocessing; IRRT = imagery rescripting and reprocessing therapy; NHS = National Health Service; PE = prolonged exposure; PTSD = post-traumatic stress disorder; RCT = randomised control trial; RPT = relapse prevention therapy (treatment for substance use disorder); Seeking Safety = skills-based intervention for concurrent post-traumatic stress disorder and substance use disorder; STAIR = Skills Training in Affective and Interpersonal Regulation; SUD = substance use disorder; Tf-CBT = Trauma-focussed cognitive behavioural therapy; VA = Veterans Affairs.

Table 3

Study methods

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
Deisenhofer et al. (2018)	Post-treatment symptom severity (continuous, PHQ-9 as a proxy measure of PTSD) Optimal treatment for each patient	Final treatment session	Clinical, demographic, psychometric (11)	Genetic algorithm (feature selection, *n* = 150; 75)	Linear regression (parameter estimation, calculate PAI, *n* = 150; 75) Chi-squared test (compare rate of reliable improvement between patients who received model indicated optimal vs. suboptimal treatment, *n* = 225)	NR	Multiple imputation via random forest (on whole sample) Categorical predictors reduced to dichotomous variables (employment, medication) Propensity score matching	Importance threshold set at 80% Other hyperparameter settings not reported	Leave-one-out cross-validation	2
Etkin et al. (2019)	≥50% reduction in PTSD score (binary, CAPS)	4 weeks after final treatment session	MRI, EEG, neurocognitive tests (unclear)	Linear support vector machine; Non-linear radial basis function support vector machine (predict treatment outcome, *n* = 36, outcome frequency not reported)	Generalised linear modelling (neurocognitive feature selection, *n* = 92 including *n* = 36 controls; neuroimaging feature selection, *n* = 87 including *n* = 36 healthy controls) Generalised linear mixed modelling (test interactions with	NR	Threshold in delayed recall score indicative of impaired recall identified by discriminant analysis (*n* = 92) Preprocessing of neuroimaging data described in supplementary materials	NR	Leave-one-out cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
					treatment, *n* = 36, vs. control, *n* = 30)					
Fleming et al. (2018)	Retention (count, *n* sessions completed)	Final treatment session	Clinical, demographic, psychometric, military service characteristics, trauma characteristics (51)	Exhaustive CHAID classification tree (feature selection, parameter estimation, prediction, *n* = 122)		NR	NR	NR	NR	1
Forbes et al. (2003)	Change in symptom score (continuous, PCL)	3 months post-treatment; 9 months post-treatment (*n* = 136)	Psychometric (16)	*k*-means cluster analysis (test reliability of subgroups identified by Ward's cluster analysis, *n* = 158)	Ward's hierarchical cluster analysis (identify subgroups of PTSD patients, *n* =158) Second order principal components analysis (reduce MMPI-2 scale and aid interpretation of results, *n* = 158) Multivariate generalised linear modelling (explore differences in outcome and independent variables between clusters, *n* = 158) Repeated measures multivariate generalised linear	NR	NR	NR	NR	1

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
					modelling (examine differences in treatment response between subgroups, *n* = 158)					
Held et al. (2022)	Minimal response (binary, PCL-5); Fast response (binary, PCL-5)	Intake, treatment days 2, 3, 5, 6, 8, 11, and 13, and post-treatment	Demographic, psychometric, military service characteristics, trauma characteristics (104)	Elastic Net classification; Gradient Boosted Models; Random Forest; Ridge classification; Logistic Regression with Max-Min Parent-Child variable selection (feature selection, parameter estimation, prediction, *n* = 432 including *n* = 73 with minimal response outcome and *n* = 61 with fast response outcome)	Group-based trajectory modelling (identify response trajectory class) Logistic Regression (comparison with ML methods)	NR	Listwise exclusion of participants with missing data One-hot-encoding of categorical variables Performance assessed by area under the precision-recall curve to account for class imbalance	Hyperparameter optimisation via five-fold cross-validated grid search within inner loop of nested five-fold cross validation Hyperparameter tuning not required for logistic regression or logistic regression with max-min parent-child variable selection Hyperparameter settings not reported	Five-fold cross-validation	2
Hendriks et al. (2018)	Response trajectory class (polytomous, CAPS)	Baseline, 3 month follow up, 6	Clinical, demographic, psychometric (14)	*k*-means cluster analysis (identify response trajectory class, *n* = 69)	Stepwise multinominal logistic regression (feature	NR	Multiple imputation of missing data following a	Varied number of clusters from 3 to	NR	1

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
		month follow up			selection and prediction, *n* = 69)		framework for multiple imputation in cluster analysis Participants missing baseline CAPS score were excluded (*n* = 4)	6 and evaluated goodness of fit Other hyperparameter settings not reported		
Herzog et al. (2021)	Change in symptom score (continuous, IES-R)	First and last day of treatment	Clinical, demographic, psychometric (≥ 46)	Elastic net (feature selection, parameter estimation, prediction, *n* = 397)		NR	Participants missing > 60% and variables missing > 40% were excluded Univariate outlier values removed Time-event data log-transformed Categorical variables were reduced to binary or continuous variables (details not reported), ICD-10 medical diagnoses were dummy coded Binary variables	L1 and L2 penalty weighting alpha set to 0.5 Optimal lambda value estimated by *k*-fold cross-validation averaged across 10 runs (within training set) Optimal lambda value not reported	Tested on randomly partitioned (35%) hold-out validation set (*n* = 215) Bootstrap internal cross-validation in training set (*n* = 397)	3

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
							with class imbalance were excluded Multiple imputation via random forest (separately on training and test set)			
Hoeboer et al. (2021)	Change in symptom score (continuous, CAPS-5; PCL-5) Optimal treatment for each patient	4 weeks, 8 weeks, and 16 weeks after start of treatment	Clinical, demographic, psychometric (24)	Boruta algorithm random forest classifier (feature selection, *n* = 99; 50)	Linear mixed-effect modelling (estimate change in symptoms over the course of treatment for each participant, *n* = 149) Linear regression (parameter estimation, prediction, *n* = 99; 50)	NR	NR	NR	Bootstrapping (feature selection) Leave-one-out cross-validation internal cross-validation (prediction, PAI)	2
Keefe et al. (2018)	Dropout (binary, treatment completion) Optimal treatment for each patient	Final treatment session	Clinical, demographic, psychometric, trauma characteristics (20)	Bootstrapped, random forest variant of model-based recursive partitioning (MoB), and bootstrapped variant of an AIC-based backward selection model (feature selection, *n* = 160 including *n* = 49 with dropout outcome)	Logistic regression (parameter estimation, prediction, *n* = 160)	NR	Participants who dropped out prior to randomisation excluded from analyses (*n* = 11) Single-dataset random forest imputation strategy using all available pre-treatment and outcome data	NR	Five-fold cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
Kratzer et al. (2019)	Reliable change (binary, IES-R)	Before discharge	Clinical, psychometric (5)	Conditional inference tree (feature selection and prediction, *n* = 150 including *n* = 78 with reliable change outcome)		NR	Bayesian multiple imputation	NR	NR	1
López-Castro et al. (2021)	Treatment attendance (count, *n* sessions attended)	Final treatment session	Clinical, demographic, psychometric, trauma characteristics (28)	Iterative Random Forest (feature selection, *n* = 70)	Poisson regression (parameter estimation, prediction, *n* = 70; 60)	NR	NR	Default hyperparameter settings used, values not reported	Parameter estimation repeated in independent dataset (training set *n* = 70; replication set *n* = 60)	1
Nixon et al. (2021)	Response trajectory class (polytomous, PDS/PSS)	Post-treatment, follow up 3 to 9 months after final session	Clinical, demographic, psychometric, trauma characteristics (38)	Random forests of conditional inference trees (feature selection and prediction, *n* = 179)		NR	Classified response trajectories identified based on symptom scores at session 1, session 6, posttreatment and follow-up	Default hyperparameter settings used, values not reported	Internal validation as part of random forest	2
Stirman et al. (2021)	Post-treatment symptom severity (continuous, CAPS)	Post-treatment	Clinical, demographic, psychometric, trauma characteristics (29)	Elastic net, five iterations, features retained if selected on all five iterations. Then stepwise AIC-penalised bootstrapped variable selection with 10,000	Linear regression with 10-fold cross-validation, coefficients mean averaged across 1000 runs (parameter estimation, generate PI, *n* = 267)	NR	Binary variables effect-coded Continuous predictors standardised	Elastic net alpha parameter set to .75, lambda optimised via 10-fold cross-validation	10-fold cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
				bootstrapped samples, features retained if selected in > 60% samples (feature selection, *n* = 267)	Linear regression (test association between PI and outcome, and interaction between PI and treatment type, *n* = 267)		Multiple imputation via random forest (OOB error estimates reported)	Optimal lambda not reported		
Stuke et al. (2021)	Change in symptom score (continuous, DTS)	Discharge	Clinical, demographic, psychometric, trauma characteristics (12)	Principal component analysis (feature reduction, *n* = 115) ADAboost regressor (parameter estimation, prediction, *n* = 115)	Linear regression (comparison with ADAboost regressor, *n* = 115)	NR	Participants missing responses to a whole scale excluded; scale mean imputed where participants were missing < 20% responses to scale (*n* = 10)	Optimal number of components for each participant estimated via hyperparameter optimisation with 10-fold cross-validation in (*N* - 1) training set, varying number of components from 1-10 and comparing squared error ADAboost: *n* estimators optimised with 10-fold cross-validation in training set (candidates: 2, 5, 10, 20, 40);	Leave-one-out cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
								default settings used for other hyperparameters Hyperparameter settings not reported		
Zhang et al. (2021)	Post-treatment symptom severity (continuous, CAPS; CAPS-5)	NR for PTSD data	EEG/PEC (unclear)	Sparse *k*-means clustering (identify PTSD subtypes, *n* = 106)	Linear mixed models (predict outcome from subtype, *n* = 72; *n* = 63)	NR	Multiple imputation reported for depression dataset but not for PTSD dataset EEG and MRI preprocessing reported in methods section	Number of clusters (2) determined and assessed by the gap statistic Sparsity parameter (*s*) optimised by inner-loop cross-validation, value not reported	*k*-means repeated on 100 randomly selected subsamples (random 90% of the sample in each subsample) PTSD treatment outcomes dataset divided into two cohorts, cluster analysis applied, and linear mixed modelling repeated in the second cohort	1

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* participants analysed, *n* with categorical outcome)	Additional methods (purpose, *n* participants analysed)	Sample size calculation	Data pre-processing	Hyperparameter setting	Validation methods	Evidence level
Zhutovsky et al. (2019)	≥30% reduction in PTSD score (binary, CAPS)	6 to 8 months from baseline assessment	MRI (unclear)	Independent component analysis using the meta-ICA approach (dimension reduction, *n* = 28 controls) Gaussian process classifier (feature selection and prediction, *n* = 44 including *n* = 20 with treatment response outcome)	Univariate analysis with threshold-free cluster enhancement and permutation analysis (dimension reduction, *n* = 44)	NR	Participants missing follow-up data were excluded from analysis, and 3 participants were excluded due to excessive movement during MRI MRI pre-processing reported in supplementary material	NR	10-fold cross-validation	2
Zilcha-Mano et al. (2020)	Change in symptom score (continuous, CAPS)	Pre to post-treatment	MRI (unclear)	Linear kernel support vector machine with *t*-test filtering and wrapper based sequential feature selection (feature reduction and selection, *n* = 179)	Pearson correlations (test correlation between features and treatment outcome, *n* = 55)	NR	Excluded 3 participants due to excessive movement during MRI Features regressed for age, sex, and MRI scanner, and normalised (-1, 1) MRI preprocessing reported in supplementary materials	Hyperparameter optimisation during 10-fold cross-validation, settings not reported	10-fold cross-validation during support vector machine training Correlations not cross-validated	1

Note. AIC = Akaike Information Criterion; CAPS = Clinician-Administered PTSD Scale; CAPS-5 = Clinician-Administered PTSD Scale for DSM-5; CHAID = Chi-square Automatic Interaction Detection; DTS = Davidson Trauma Scale; EEG = Electroencephalography; ICD-10 = International Classification of Diseases 10th Revision; IES-R = Impact-of-Event-Scale-Revised; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; MRI = Magnetic Resonance Imaging; NR = Not Reported; OOB = Out-Of-Bag; PAI = Personalised Advantage Index; PCL = PTSD Checklist; PCL-5 = PTSD Checklist for DSM-5; PDS = Posttraumatic Stress Diagnostic Scale; PEC = Power Envelope Connectivity; PHQ-9 = Patient Health Questionnaire-9; PI = Prognostic Index; PSS = Post-traumatic Symptoms Scale; PTSD = Post-traumatic Stress Syndrome; RCT = Randomised Controlled Trial.

Database Search Terms

Scopus

((TITLE-ABS-KEY("psychotherapy" OR "psychological therapy" OR "Cognitive behavio* therapy" OR "cognitive therapy" OR "CBT" OR "cognitive processing therapy" OR "Exposure therapy" OR "Prolonged exposure" OR "narrative exposure" OR "Eye Movement Desensiti*ation and Reprocessing" OR "EMDR" OR "trauma focussed therapy" OR "Brief Eclectic Psychotherapy"))

AND

(TITLE-ABS-KEY("posttraumatic" OR "post traumatic" OR "post-traumatic" OR "traumatic stress" OR "traumatic memor*" OR "ptsd" OR "cptsd"))

AND

(TITLE-ABS-KEY("machine learning" OR "machine-learning" OR "supervised learning" OR "unsupervised learning" OR algorithm OR "statistical learning" OR "artificial intelligence" OR "AI" OR "data mining" OR "deep learning" OR "kernel" OR "personali*ed advantage ind*" OR "regularized" OR "ridge" OR "least absolute shrinkage and selection operator" OR "LASSO" OR "elastic net" OR "decision tree*" OR "random forest*" OR "regression tree*" OR "classification tree*" OR "nearest neighb*" OR "k-nn" OR "k-means" OR "support vector" OR "vector machine" OR "SVM" OR "SVR" OR "naïve Bayes" OR "Bayesian network" OR "neural network" OR "perceptron" OR "radial basis function" OR "cluster analysis" OR "principal components analysis" OR "latent transition" OR "autoencoder" OR "dimensionality reduction" OR "latent Dirichlet allocation" OR "LDA" OR "chi-square automatic interaction detection" OR "CHAID" OR "XGBoost" OR "bootstrap*" OR "bagging" OR "boosting")))

PubMed

(("psychotherapy"[MeSH Terms] OR "psychotherapy"[Title/Abstract] OR "Cognitive behavio* therapy"[Title/Abstract] OR "cognitive therapy"[Title/Abstract] OR "CBT"[Title/Abstract] OR "cognitive processing therapy"[Title/Abstract] OR "Exposure therapy"[Title/Abstract] OR "Prolonged exposure"[Title/Abstract] OR "narrative exposure"[Title/Abstract] OR "Eye Movement Desensiti*ation and Reprocessing"[Title/Abstract] OR "EMDR"[Title/Abstract] OR "trauma focused therapy"[Title/Abstract] OR "Brief Eclectic Psychotherapy"[Title/Abstract])

AND

("Stress Disorders, Post-Traumatic"[MeSH Terms] OR "post-traumatic"[Title/Abstract] "posttraumatic"[Title/Abstract] OR "post traumatic"[Title/Abstract] OR "traumatic stress"[Title/Abstract] OR "traumatic memor*"[Title/Abstract] OR "ptsd"[Title/Abstract] OR "cptsd"[Title/Abstract])

AND

(("machine learning"[MeSH Terms] OR "machine-learning"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "supervised learning"[Title/Abstract] OR "unsupervised learning"[Title/Abstract] OR "algorithm"[Title/Abstract] OR "statistical learning"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "AI"[Title/Abstract] OR "data mining"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "kernel"[Title/Abstract] OR "personali*ed advantage ind*"[Title/Abstract] OR "regularized"[Title/Abstract] OR "ridge"[Title/Abstract] OR "least absolute shrinkage and selection operator"[Title/Abstract] OR "LASSO"[Title/Abstract] OR "elastic net"[Title/Abstract] OR "decision tree*"[Title/Abstract] OR "random forest*"[Title/Abstract] OR "regression tree*"[Title/Abstract] OR "classification tree*"[Title/Abstract] OR "nearest neighb*"[Title/Abstract] OR "k-nn"[Title/Abstract] OR "k-means"[Title/Abstract] OR "support vector"[Title/Abstract] OR "vector machine"[Title/Abstract] OR "SVM"[Title/Abstract] OR "SVR"[Title/Abstract] OR "naïve Bayes"[Title/Abstract] OR "Bayesian network"[Title/Abstract] OR "neural network"[Title/Abstract] OR "perceptron"[Title/Abstract] OR "radial basis function"[Title/Abstract] OR "cluster analysis"[Title/Abstract] OR "principal components analysis"[Title/Abstract] OR "latent transition"[Title/Abstract] OR "autoencoder"[Title/Abstract] OR "dimensionality reduction"[Title/Abstract] OR "latent Dirichlet allocation"[Title/Abstract] OR "LDA"[Title/Abstract] OR "chi-square automatic interaction detection"[Title/Abstract] OR "CHAID"[Title/Abstract] OR "XGBoost"[Title/Abstract] OR "bootstrap*"[Title/Abstract] OR "bagging"[Title/Abstract] OR "boosting"[Title/Abstract])))

APA PsycInfo via Ovid

(("psychotherapy" OR "psychological therapy" OR "Cognitive behavio$ therapy" OR "cognitive therapy" OR "CBT" OR "Exposure therapy" OR "Prolonged exposure" OR "narrative exposure" OR "Eye Movement Desensiti$ation and Reprocessing" OR "EMDR" OR "trauma focussed therapy" OR "Brief Eclectic Psychotherapy").ab,ti,id.

AND

("posttraumatic" OR "post traumatic" OR "traumatic stress" OR "traumatic memor$" OR "ptsd" OR "cptsd").ab,ti,id.

AND

("machine learning" OR "machine-learning" OR "supervised learning" OR "unsupervised learning" OR algorithm OR "statistical learning" OR "artificial intelligence" OR "AI" OR "data mining" OR "deep learning" OR "kernel" OR "personali*ed advantage ind*" OR "regularized" OR "ridge" OR "least absolute shrinkage and selection operator" OR "LASSO" OR "elastic net" OR "decision tree*" OR "random forest*" OR "regression tree*" OR "classification tree*" OR "nearest neighb*" OR "k-nn" OR "k-means" OR "support vector" OR "vector machine" OR "SVM" OR "SVR" OR "naïve Bayes" OR "Bayesian network" OR "neural network" OR "perceptron" OR "radial basis function" OR "cluster analysis" OR "principal components analysis" OR "latent transition" OR "autoencoder" OR "dimensionality reduction" OR "latent Dirichlet allocation" OR "LDA" OR "chi-square automatic interaction detection" OR "CHAID" OR "XGBoost" OR "bootstrap*" OR "bagging" OR "boosting").ab,ti,id.)

PTSDpubs via ProQuest

(NOFT(("psychotherapy" OR "psychological therapy" OR "Cognitive behavio* therapy" OR "CBT" OR "cognitive therapy" OR "cognitive processing therapy" OR "Exposure therapy" OR "Prolonged exposure" OR "narrative exposure" OR "Eye Movement Desensiti*ation and Reprocessing" OR "EMDR" OR "trauma focussed therapy" OR "Brief Eclectic Psychotherapy") AND ("posttraumatic" OR "post traumatic" OR "post-traumatic" OR "traumatic stress" OR "traumatic memor*" OR "ptsd" OR "cptsd") AND ("machine learning" OR "machine-learning" OR "supervised learning" OR "unsupervised learning" OR algorithm OR "statistical learning" OR "artificial intelligence" OR "AI" OR "data mining" OR "deep learning" OR "kernel" OR "personali*ed advantage ind*" OR "regularized" OR "ridge" OR "least absolute shrinkage and selection operator" OR "LASSO" OR "elastic net" OR "decision tree*" OR "random forest*" OR "regression tree*" OR "classification tree*" OR "nearest neighb*" OR "k-nn" OR "k-means" OR "support vector" OR "vector machine" OR "SVM" OR "SVR" OR "naïve Bayes" OR "Bayesian network" OR "neural network" OR "perceptron" OR "radial basis function" OR "cluster analysis" OR "principal components analysis" OR "latent transition" OR "autoencoder" OR "dimensionality reduction" OR "latent Dirichlet allocation" OR "LDA" OR "chi-square automatic interaction detection" OR "CHAID" OR "XGBoost" OR "bootstrap*" OR "bagging" OR "boosting")))

Supplementary Table 1

Psychological therapies recommended as first line treatment for current posttraumatic-stress disorder in adults by clinical practice guidelines

Treatment	**Clinical Practice Guideline**				
	APA (2017)[1]	ISTSS (2020)[2]	NICE (2018)[3]	PAC (2020)[4]	Va/DoD (2017)[1]
Trauma focussed-cognitive behavioural therapy		✓✓	✓✓	✓✓	✓✓
Cognitive processing therapy	✓✓	✓✓	✓✓	✓✓	✓✓
Cognitive therapy for post-traumatic stress disorder			✓✓		
Cognitive behavioural therapy	✓✓	✓			
Cognitive therapy	✓✓	✓✓		✓✓	
Eye movement desensitisation and reprocessing	✓	✓✓	✓	✓✓	✓✓
Prolonged exposure	✓✓	✓✓	✓✓	✓✓	✓✓
Narrative exposure therapy	✓	✓	✓✓	✓	✓✓
Written Narrative exposure					✓✓
Brief Eclectic psychotherapy	✓				✓✓

Note. APA = American Psychological Association; ISTSS = International Society for Traumatic Stress Studies; NICE = National Institute for Health and Care Excellence; PAC = Phoenix Australia Centre for Posttraumatic Mental Health; Va/DoD = Veterans Affairs/Department of Defence.

✓✓ = Primary/strong recommendation.

✓ = Secondary/conditional recommendation.

[1] USA

[2] International

[3] UK

[4] Australia

Supplementary Table 2

Studies excluded during full text screening

Author	Title	DOI	Reason For Exclusion
Barnes et al. (2019)	Developing predictive models to enhance clinician prediction of suicide attempts among veterans with and without PTSD	https://doi.org/10.1111/sltb.12511	Did not predict PTSD treatment outcome (Predicted suicide)
Bryant et al. (2008)	Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder	https://doi.org/10.1017/s0033291707002231	No machine learning methods
de Kleine et al. (2014)	Prescriptive variables for d-cycloserine augmentation of exposure therapy for posttraumatic stress disorder	https://doi.org/10.1016/j.jpsychires.2013.10.008	No machine learning methods
Dorrepaal et al. (2013)	Treatment compliance and effectiveness in complex PTSD patients with co-morbid personality disorder undergoing stabilizing cognitive behavioral group treatment: A preliminary study	https://doi.org/10.3402/ejpt.v4i0.21171	Psychological therapy for PTSD not recommended by CPG (Group CBT only)
Galatzer-Levy et al. (2014)	Quantitative forecasting of PTSD from early trauma responses: A machine learning application	https://doi.org/10.1016/j.jpsychires.2014.08.017	Did not predict treatment outcome (Predicted onset of PTSD)
Galatzer-Levy et al. (2017)	Utilization of machine learning for prediction of post-traumatic stress: A re-examination of cortisol in the prediction and pathways to non-remitting PTSD	https://doi.org/10.1038/tp.2017.38	Did not predict treatment outcome (Predicted onset of PTSD)
Hilbert et al. (2021)	Identifying CBT non-response among OCD outpatients: A machine-learning approach	https://doi.org/10.1080/10503307.2020.1839140	Treatment not for PTSD (Treatment for OCD)
Karstoft et al. (2015)	Bridging a translational gap: Using machine learning to improve the prediction of PTSD	https://doi.org/10.1186/s12888-015-0399-8	Did not predict treatment outcome (Predicted onset of PTSD)
Korgaonkar et al. (2020)	Intrinsic connectomes underlying response to trauma-focused psychotherapy in post-traumatic stress disorder	https://doi.org/10.1038/s41398-020-00938-8	No machine learning methods
Lutz et al. (2019)	Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN)	https://doi.org/10.1016/j.brat.2019.103438	Sample not adults with PTSD (Only 5.7% of participants were seeking treatment for PTSD)
Rizvi et al. (2009)	Cognitive and affective predictors of treatment outcome in Cognitive Processing Therapy and Prolonged Exposure for posttraumatic stress disorder	https://doi.org/10.1016/j.brat.2009.06.003	No machine learning methods
Roberge et al. (2019)	Predicting response to cognitive processing therapy: Does trauma history matter?	https://doi.org/10.1037/tra0000530	No machine learning methods
Rubel et al. (2019)	Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration	https://doi.org/10.1080/10503307.2019.1597994	Sample not adults with PTSD (Only 3.9% of the sample sought treatment for PTSD)

Author	Title	DOI	Reason For Exclusion
Schultebraucks et al. (2021)	Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: A machine-learning approach for analyzing multivariate predictors	https://doi.org/10.1038/s41380-020-0789-2	Did not predict treatment outcome (Predicted onset of PTSD)
Schultebraucks et al. (2020)	A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor	https://doi.org/10.1038/s41591-020-0951-z	Did not predict treatment outcome (Predicted onset of PTSD)
Schumm et al. (2013)	Latent class differences explain variability in PTSD symptom changes during cognitive processing therapy for veterans	https://doi.org/10.1037/a0030359	No machine learning methods
Senger et al. (2021)	Predicting optimal treatment outcomes using the Personalized Advantage Index for patients with persistent somatic symptoms	https://doi.org/10.1080/10503307.2021.1916120	Treatment not for PTSD (Treatment for persistent somatic symptoms)
Smith et al. (2019)	Predictors of dropout from residential treatment for posttraumatic stress disorder among military veterans	https://doi.org/10.3389/fpsyg.2019.00362	No machine learning methods
Sonne et al. (2016)	Psychosocial predictors of treatment outcome for trauma-affected refugees	https://doi.org/10.3402/ejpt.v7.30907	No machine learning methods
Stein et al. (2012)	Trajectories of response to treatment for posttraumatic stress disorder	https://doi.org/10.1016/j.beth.2012.04.003	No machine learning methods
Taylor (2003)	Outcome predictors for three PTSD treatments: Exposure therapy, EMDR, and relaxation training	https://dx.doi.org/10.1891/jcop.17.2.149.57432	No machine learning methods
Taylor et al. (2001)	Posttraumatic stress disorder arising after road traffic collisions: Patterns of response to cognitive-behavior therapy	https://doi.org/10.1037/0022-006X.69.3.541	Psychological therapy for PTSD not recommended by CPG (Group CBT)
van Geusau et al. (2021)	Predicting outcome in an intensive outpatient PTSD treatment program using daily measures	https://doi.org/10.3390/jcm10184152	No machine learning methods
van Minnen et al. (2002)	Prolonged exposure in patients with chronic PTSD: Predictors of treatment outcome and dropout	https://doi.org/10.1016/s0005-7967(01)00024-9	No machine learning methods
van Rooij et al. (2014)	Neural correlates of inhibition and contextual cue processing related to treatment response in PTSD	https://doi.org/10.1038/npp.2014.220	No machine learning methods
van Rooij et al. (2015)	Predicting treatment outcome in PTSD: A longitudinal functional MRI study on trauma-unrelated emotional processing	https://doi.org/10.1038/npp.2015.257	No machine learning methods
Vöhringer et al. (2020)	Should I stay or must I go? Predictors of dropout in an internet-based psychotherapy programme for posttraumatic stress disorder in Arabic	https://doi.org/10.1080/20008198.2019.1706297	Psychological therapy for PTSD not recommended by CPG (Internet-based CBT)

Author	Title	DOI	Reason For Exclusion
Wester et al. (2022)	Covariate selection for estimating individual treatment effects in psychotherapy research: A simulation study and empirical example	https://doi.org/10.1177/21677026211071043	Sample not adults with PTSD (Simulation study)
Yuan et al. (2018)	Pre-treatment resting-state functional MR imaging predicts the long-term clinical outcome after short-term Paroxtine treatment in post-traumatic stress disorder	https://doi.org/10.3389/fpsyt.2018.00532	No psychological therapy (Pharmacological treatment only)
Zegerius et al. (2021)	Modelling metaplasticity and memory reconsolidation during an eye-movement desensitization and reprocessing treatment	https://doi.org/10.1007/978-3-030-65596-9_74	No machine learning methods
Zhutovsky et al. (2021)	Individual prediction of trauma-focused psychotherapy response in youth with posttraumatic stress disorder using resting-state functional connectivity	https://doi.org/10.1016/j.nicl.2021.102898	Sample not adults with PTSD (Sample < 18 years old)

Note. CBT = cognitive behavioural therapy; CPG = clinical practice guidelines; OCD = obsessive compulsive disorder; PTSD = post-traumatic stress disorder.

Supplementary Table 3

Risk of bias assessments with PROBAST

	PROBAST Signalling Question	Deisenhofer et al. (2018)	Etkin et al. (2019)	Fleming et al. (2018)	Forbes et al. (2003)	Held et al. (2022)	Hendriks et al. (2018)	Herzog et al. (2021)	Hoeboer et al. (2021)	Keefe et al. (2018)	Kratzer et al. (2019)	López-Castro et al. (2021)	Nixon et al. (2021)	Stirman et al. (2021)	Stuke et al. (2021)	Zhang et al. (2021)	Zhutovsky et al. (2019)	Zilcha-Mano et al. (2020)
1.1	Were appropriate data sources used, e.g., cohort, RCT or nested case-control study data?	N	Y	N	N	Y	Y	N	Y	Y	N	Y	Y	Y	N	Y	Y	Y
1.2	Were all inclusions and exclusions of participants appropriate?	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Risk of bias introduced by selection of participants	High	Low	High	High	Low	Low	High	Low	Low	High	Low	Low	Low	High	Low	Low	Low
2.1	Were predictors defined and assessed in a similar way for all participants?	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
2.2	Were predictor assessments made without knowledge of outcome data?	Y	PY	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
2.3	Are all predictors available at the time the model is intended to be used?	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
	Risk of bias introduced by predictors or their assessment	Low	Low	Low	Low	Low	High	Low	Low	Low	Low	High	Low	Low	Low	Low	Low	Low
3.1	Was the outcome determined appropriately?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.2	Was a pre-specified or standard outcome definition used?	Y	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.3	Were predictors excluded from the outcome definition?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.4	Was the outcome defined and determined in a similar way for all participants?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

3.5	Was the outcome determined without knowledge of predictor information?	Y	NI	Y	Y	Y	Y	Y	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.6	Was the time interval between predictor assessment and outcome determination appropriate?	PY	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Risk of bias introduced by the outcome or its determination	Low	Unclear	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
4.1	Were there a reasonable number of participants with the outcome?	N	N	N	N	N	N	PY	N	N	N	N	N	N	N	PN	N	N
4.2	Were continuous and categorical predictors handled appropriately?	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
4.3	Were all enrolled participants included in the analysis?	N	N	N	N	N	N	Y	Y	N	N	N	N	N	N	Y	N	Y
4.4	Were participants with missing data handled appropriately?	Y	NI	NI	PY	N	Y	Y	Y	Y	Y	NI	N	Y	N	NI	NI	NI
4.5	Was selection of predictors based on univariable analysis avoided?	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y	N	Y
4.6	Were complexities in the data (e.g., censoring, competing risks, sampling of controls) accounted for appropriately?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
4.6	Were relevant model performance measures evaluated appropriately?	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	NA
4.8	Were model overfitting and optimism in model performance accounted for?	Y	Y	N	N	Y	N	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y
4.9	Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	Y	NI	Y	NI	NI	Y	Y	Y	Y	Y	N	NI	Y	NI	NI	NI	NA
	Risk of bias introduced by the analysis	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
	Overall risk of bias assessment	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**

Note. N = no; NA = not applicable; NI = no information; PN = probably no; PY = probably yes; Y = yes.

Supplementary Table 4

Study results and study author's interpretation

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
Deisenhofer et al. (2018)	*Tf-CBT:* Functional impairment (WSAS), employment status, age, gender (4) *EMDR:* Depression (PHQ-9), medication (2)	True error: Tf-CBT = 4.92 EMDR = 5.37 Whole analysed sample = 5.07	For Tf-CBT, higher functional impairment (WSAS; $\beta = 0.32$, $p < .001$), being male ($\beta = -0.14$, $p \geq .05$), being unemployed ($\beta = -0.33$, $p < .001$), and younger age ($\beta = -0.17$, $p < .05$) predicted more severe symptoms post-treatment. For EMDR, higher baseline depression (PHQ-9; $\beta = 0.40$, $p < .01$) and being prescribed antidepressants ($\beta = 0.29$, $p < .001$) predicted more severe symptoms post-treatment. Patients who received their model indicated optimal treatment (n = 124) reported significantly better outcomes than those who received their model indicated suboptimal treatment (n = 101; Cohen's d = 0.40, 95% CI [0.13, 0.67], number needed to treat = 4.49). There was a significantly higher rate of reliable improvement among patients who received their model indicated optimal treatment (62.9%) versus those who received their suboptimal treatment (33.66%; $\chi^2(1, n = 225) = 19.54$, $p < .001$).	Results consistent with previous literature. Limitations: Exploratory modelling using retrospective data, small sample, and propensity score matching, no PTSD symptom measure. Prospective testing is required with a larger sample and PTSD symptom outcome measure. Integrating a PAI into routine clinical practice could potentially improve PTSD outcomes by guiding selection of optimal treatment.
Etkin et al. (2019)	Verbal memory delayed recall impairment, within ventral attention network functional connectivity (2)	Linear support vector machine: Accuracy = 85% Sensitivity = 80% Specificity = 87% Radial basis function support vector machine: Accuracy = 90% Sensitivity = 80% Specificity = 93%	Generalized linear model results indicated that impaired verbal memory, low within Ventral Attention Network connectivity, and their interaction, were not significantly associated with baseline PTSD severity, depression severity, PTSD symptom clusters, dissociative symptoms, comorbid diagnoses, alcohol use, traumatic brain injury, or quality of life. Generalized linear mixed model results indicated that the interaction between verbal memory recall and within Ventral Attention Network functional connectivity moderated treatment outcome (PE vs. wait list), in that impaired verbal memory and low within Ventral Attention Network connectivity predicted no change in PTSD symptom score in the PE group. Support vector machine results indicated that the combination of low within Ventral Attention Network connectivity and impaired verbal memory predicted non-response to PE.	Findings consistent with recent calls to move away from symptom focussed definitions of PTSD heterogeneity to cognitive-neurological definitions. Replication is needed.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			Either variable alone did not predict outcome (accuracies ≤ 65%, $p > 0.18$).	
Fleming et al. (2018)	Days between introductory session and therapy invitation, URICA item 9 (“*I have been successful in working on my trauma issues, but I’m not sure I can keep up the effort on my own*”), URICA item 5 (“*I’m not the one with trauma issues, it doesn’t make much sense for me to be here*”), presence of traumatic brain injury, prior PTSD treatment (5)	None	Wait time before treatment was the best predictor of attendance, followed by motivational readiness to address trauma, then traumatic brain injury and prior PTSD treatment. Exhaustive CHAID classification tree indicated that participants who waited > 68 days to begin treatment attended significantly fewer treatment sessions; and among those who waited > 68 days, participants who disagreed with the URICA scale item 9 attended significantly fewer sessions. Among those who waited ≤ 68 days, participants who strongly disagreed with URICA item 5 completed significantly more sessions; among those who strongly disagreed with URICA item 5, those with traumatic brain injury attended significantly more sessions; and among those who didn't strongly disagree with URICA item 5, participants with no prior PTSD treatment attended significantly more sessions.	The finding that shorter wait time and motivation is associated with greater retention is congruent with previous research. However, contrary to previous research and theory, the current findings suggest that these are most important predictors of engagement above demographic and clinical variables; this requires further study and replication. This study was exploratory, and the use of naturalistic clinical data limits internal validity. There may be extraneous variables associated with treatment retention. Clinical implications: Engage veterans quickly and focus on motivation and preparation for treatment.
Forbes et al. (2003)	Clinical subscales and validity subscales of MMPI-2, anxiety and depression (HADS), alcohol use (AUDIT) (16)	None	Ward's cluster analysis identified three groups, and this was supported by the *k*-means cluster analysis with three-cluster classification agreement in 125 of the 158 cases ($n = 158$, alpha = 0.68, $p < 0.001$). Group 1 ($n = 62$) scored highly on PTSD symptom severity, but lower in other psychiatric measures; Group 2 ($n = 38$) scored significantly lower on PTSD symptom severity, and low on other psychiatric measures; and Group 3 ($n = 36$) scored as highly as Group 1 on PTSD symptom severity, and significantly higher on personality disturbance, other forms of psychopathology, emotional vulnerability, and expressed distress. Repeated-measures multivariate general linear model found significant effect of time ($F(2,132) = 22.56$, $p < 0.001$) and group ($F(2,133) = 12.51$, $p < 0.001$), and a significant time by group interaction ($F(4,266) = 3.35$, $p < 0.02$). Univariate post-hoc analyses revealed that Group 1 and 3 significantly improved from baseline to 3 months, but Group 2 did not. Cross-sectional analyses revealed a significant difference between groups mean PCL scores at baseline	Based on MMPI-2 characteristics groups may be labelled as follows: Group 1 = High PTSD-introversion/ somatization. Group 2 = Low PTSD-subclinical personality pathology. Group 3 = High PTSD-disinhibition/externalization. Group 2 represents a "subclinical" group, with lower intake PTSD, anxiety, and depression scores than Groups 1 and 3. Groups 1 and 3 reported similarly severe intake symptom scores, but Group 3 scored higher on Factor 2 (indicating greater personality disturbance and psychopathology), and lower on Factor 3 (indicating greater emotional vulnerability and expressed distress). There are some similarities to/consistencies with subgroups identified in previous studies.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			($F(2,133) = 15.91$, $p < 0.001$), 3 months post-treatment ($F(2,133) = 5.28$, $p < 0.01$), and 9 months post-treatment ($F(2,133) = 8.48$, $p < 0.001$); and Scheffe post-hoc analyses revealed that Group 2 had significantly less severe PTSD symptoms at baseline than Groups 1 and 3 (with no difference between 1 and 3), but at 9 month follow up Group 3 had significantly more severe PTSD symptoms than Groups 1 and 2 (with no difference between 1 and 2).	Personality pathology and psychopathology of Group 3 may have negatively affected long term symptom change, consistent with previous research on other disorders with comorbid personality disorder. Poor treatment response of Group 2 may be due to a floor effect, mild symptoms at intake allows less capacity for change. Limitations: Imprecision of procedure of identifying optimal solution in cluster analysis; small sample size limits generalizability; no non-treatment control group, no comparison to natural pattern of symptom change over time.
Held et al. (2022)	*Gradient Boosted Models predicting Fast Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5), post-traumatic cognitions (item level responses to PTCI) (10) *Elastic Net predicting Minimal Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5), post-traumatic cognitions (item level responses to PTCI), marriage/domestic partnership status, age, gender, level of education, military service branch (60)	Predicting fast response AUC-PR: Gradient Boosted Models = 0.466 Random Forest = 0.457 MMPC LR = 0.450 Elastic Net = 0.405 Ridge Classification = 0.394 Logistic Regression = 0.224 AUC-ROC: Gradient Boosted Models = 0.765 Predicting minimal response AUC-PR: Gradient Boosted Models = 0.583 Random Forest = 0.595 MMPC LR = 0.579 Elastic Net = 0.628 Ridge Classification = 0.611	Group Based Trajectory Modelling classified $n = 61$ participants (14.1%) as fast responders, and $n = 73$ participants (16.9%) as minimal responders. When predicting fast response class membership Gradient Boosted Models performed best and selected 10 predictor variables, all of which were trauma related and were item level scores from PCL-5, PTCI, and CAPS-5 measurements. When predicting minimal response class membership Elastic Net performed best and selected 60 predictor variables. This included 7 demographic variables, but the trauma related clinical variables were the most important predictors.	These findings are congruent with those of similar, recent studies. Strengths include large sample size, large number of variables, generalizability of naturalistic clinical sample. Limitations include reliance on self-report measures, outcome class imbalance, trajectories based on probabilistic trajectory modelling (which can be sample dependent and could reduce generalizability), and the likelihood of misclassification limits clinical usefulness. It is not recommended that clinicians rely on this model alone when deciding whether to include/exclude a patient from treatment.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
		Logistic Regression = 0.288 AUC-ROC: Elastic Net = 0.826		
Hendriks et al. (2018)	Living condition, between-session fear habituation (2)	None	*k*-means clustering analysis identified four response trajectories: fast responders (n = 9), slow responders (n = 18), partial responders (n = 22), and non-responders (n = 20). Living Condition was the only significant pre-treatment predictor of response trajectory class, in that living alone significantly predicted partial response cluster membership (B = -1.89, SE = .70, $p < .05$, OR = .15, 95% CI [.04, .59]). Additionally, greater between-session fear habituation significantly predicted Fast Response cluster membership (B = .76, SE = .39, p = .05, OR = 2.13, 95% CI [1.00, 4.54]).	The response trajectories identified corresponded to those identified in previous research. The finding that between-session, but not within-session, fear habituation was associated with outcome is coherent with previous research. The finding that patients in the partial-responders cluster were more likely to be living alone than those in the non-responders cluster conflicts with previous findings; this may be due to complex PTSD and partners facilitating PTSD-related avoidance behaviour. Early treatment process variables more robust predictors of intensive PTSD treatment outcomes than baseline demographic and clinical variables. Replication in a randomised controlled design is required. Other limitations include the low number of participants per cluster; standardised, validated measures were not used to measure treatment resistance, suicidal ideation, self-harm, aggressive behaviour, and sense of losing control; therapist adherence was measured via self-report; DSM-IV-TR (CAPS) diagnostic criteria was used instead of DSM-5 (CAPS-5).
Herzog et al. (2021)	PTSD severity (IES-R total), psychoticism (BSI), avoidance (IES-R subscale), wish to retire, depression (BDI-II), number of comorbid diagnoses, age, bronchial asthma, physical symptoms (PHQ-15), outpatient psychiatric care, children, being retired, work disability in past year, outpatient psychotherapy,	Training set: R^2 = 0.17 MAE = 0.69 RMSE = 0.91 Test set: R^2 = 0.16 MAE = 0.77 RMSE = 0.95	Elastic Net Regularization selected 11 clinical/psychological factors, and 4 sociodemographic factors. The strongest and most stable predictors were clinical/psychological, and the top three were baseline PTSD score (IES-R, β = .207), psychoticism (BSI subscale, β = -.110), and avoidance (IES-R subscale, β = .088). The strongest and most stable sociodemographic predictors were 'wish to retire' (β = -.078) and older age (β = -.038). More severe baseline PTSD (IES-R; β = .207) and PTSD-related avoidance (IES-R subscale; β = .088) were associated with better	A prediction model combining predictors from multiple domains (sociodemographic, clinical, psychometric) could help to predict variance in treatment outcome between different patients. Findings regarding predictors associated with outcome were in line with previous theory and research. The proportion of variance (R^2) explained by the model is satisfactory, but some variables that were not measured in routine care may be important predictors, such as

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
	somatization (BSI) (15)		outcomes; whereas higher psychoticism (BSI subscale; β = -.110), wish to retire (β = -.078), higher baseline depression (BDI-II; β = -.068), greater number of comorbid diagnoses (β = -.057), older age (β = -.038), bronchial asthma (β = -.024), being retired (β = -.007), and work disability in last year (β = -.005) were associated with poorer treatment outcomes.	expectations, childhood maltreatment, treatment resistance and chronicity. The proportion of variance explained in the test set is comparable to that in the training set, suggesting that overfitting of the model to the data did not occur. Limitations: Risk of lower internal validity, exploratory, naturalistic study using retrospective data, lack of strict manual for treatment limits generalizability, no assessment of therapist effects, no follow up measures, no control group, no randomization, no blinding of outcome assessors, large percentage of patients were excluded due to missing outcome measure data, generalizability limited to inpatient multi-modal treatment setting (common in Germany but not elsewhere), did not control for concurrent psychopharmacological interventions.
Hoeboer et al. (2021)	*CAPS-5 change following PE/IPE:* Depression (BDI), social support (MOS), concurrent mental health problems (MINI axis 1), childhood sexual abuse (CTQ) (4) *CAPS-5 change following STAIR+PE:* General health status (EQ-5D-5L), emotion regulation difficulties (DERS), PTSD severity (CAPS-5) (3) *PCL-5 change following PE or IPE:* Depression (BDI), social support (MOS) (2)	RMSE (referred to as *average error* in the publication) *CAPS-5:* PE/IPE = 5.09 (SD = 7.57) STAIR+PE = 4.06 (SD = 7.25) *PCL-5:* PE/IPE = 7.09 (SD = 6.16) STAIR+PE = 7.24 (SD = 4.74)	When predicting change in observer rated PTSD symptom (CAPS-5) score following PE or IPE, higher depression score (BDI), higher childhood sexual abuse score (CTQ), lower social support score (MOS), and more DSM-IV axis-1 diagnoses (MINI), predicted poorer response to treatment. When predicting change in observer rated PTSD symptom (CAPS-5) score following STAIR+PE, higher difficulties in emotion regulation (DERS), higher baseline PTSD (CAPS-5), and lower general health status (EQ-5D-5L), predicted poorer response to treatment. With CAPS-5 as outcome, 50% of patients were randomised to their optimal treatment, patients randomised to their optimal treatment reported a significantly larger reduction in symptoms (Mean (SD) reduction = 22.96 (6.99)) than patients randomised to their suboptimal treatment (Mean (SD) reduction = 18.94 (7.57); $F(1,147) = 11.36$, $p < 0.001$; Cohen's $d = 0.55$, 95% CI [0.23, 0.88]). When predicting change in self-rated PTSD symptom (PCL-5) score following PE or IPE, higher depression (BDI) score, and lower social	Approximately half of the patients were randomised to their suboptimal treatment and may have benefitted from model-based treatment selection. Predictors selected were consistent with previous personalised PTSD treatment studies, except that no demographic variables were selected as predictors. This may be due to the greater number of candidate clinical predictors, which appear to be better predictors of outcome than demographics. The finding that more emotion regulation difficulties predicted poorer outcomes in the STAIR+PE condition is in contrast to previous findings, and is important as STAIR+PE was developed for patients with severe emotion regulation difficulties who may find it difficult to tolerate PE. Many variables often identified as predictors of PTSD treatment outcomes were not selected as predictors in

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
	PCL-5 change following STAIR and PE: General health status (EQ-5D-5L), emotion regulation difficulties (DERS) (2)		support (MOS) score predicted poorer response to treatment. When predicting change in self-rated PTSD symptom (PCL-5) score following STAIR+PE, lower general health status (EQ-5D-5L) score and higher difficulties in emotion regulation (DERS) scores predicted poorer response to treatment. With PCL-5 as outcome, 63% of patients were randomised to their optimal treatment, patients randomised to their optimal treatment reported a significantly larger reduction in symptoms (Mean (SD) reduction = 25.65 (10.4)) than patients randomised to their suboptimal treatment (Mean (SD) reduction = 20.96 (9.84); $F(1,147) = 7.67$, $p = 0.006$; Cohen's $d = 0.47$, 95% CI [0.13, 0.81]).	either model, such as dissociation and personality disorders. Limitations: Sample size did not allow for *k*-fold cross validation or partitioning of a holdout sample; PAI based on linear combination of predictors, and some predictors identified by the Boruta algorithm but dropped during bootstrapping may be non-linearly related to outcome; relationship between baseline CAPS-5 and change in CAPS-5 (outcome) may be due to regression to the mean and mathematical coupling.
Keefe et al. (2018)	*Prescriptive variables:* Childhood physical abuse, current relationship abuse, trait anger, race *Prognostic variables:* Years of education, estimated IQ score (6)	None	A total $n = 49$ participants (30.6%) dropped out after starting treatment, including $n = 25$ from the PE group (30.9%) and $n = 24$ from the CPT group (30.4%). The final PAI model consisted of four moderator (prescriptive) variables, and two predictor (prognostic) variables. (Non-significant trends were included when they were consistently selected by the bootstrapped variable selection process.) Prescriptive variables: Patients were more likely to dropout from PE, relative to CPT, when they reported higher current relationship abuse (log odds = −1.08, 95% CI [−2.20, −0.13], SE = 0.52, $p = .037$); belonging to a racial minority (log odds = 1.96, 95% CI [0.17, 3.88], SE = 0.94, $p = .037$); reported more severe childhood physical abuse (log odds = −0.83, 95% CI [−1.80, 0.07], SE = 0.47, $p = .078$); and higher levels of anger (log odds = −0.90, 95% CI [−1.94, 0.07], SE = 0.51, $p = .075$). Prognostic variables: Patients were more likely to complete either treatment when they scored higher on the quick IQ test (log odds = 0.60, 95% CI [0.10, 1.13], SE = 0.26, $p = .021$); or completed more years of education (log odds = 0.45, 95% CI [−0.05, 0.98], SE = 0.26, $p = .091$). 19.7% of patients who received their model-indicated optimal treatment dropped out, compared to 40.5% of patients received	Findings of this exploratory analysis suggest that machine-learning and bootstrapping methodologies may be used to effectively predict optimal treatment for each patient, and reduce the likelihood that patients will dropout before benefitting from treatment. Models such as this could potentially inform decision-support tools to be used in clinical practice. Limitations: It is not clear whether this will generalize to other primary trauma populations; several potentially important variables were not included such as biomarkers and patient preference; all selected variables predicted greater probability of dropout in PE, which could suggest that variables important for predicting dropout from CPT were not included in the analysis; patients classified as dropouts may have benefitted from treatment before they dropped out; the model was only tested on the same data used to train it (internal cross-validation); only investigated two commonly used treatments for PTSD, CPT protocol has been updated since the RCT took place and there is evidence that the new protocol may reduce dropout.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			their model-indicated suboptimal treatment (log odds = 1.02, 95% CI [0.32, 1.75], $z = 2.80$, $p = .005$; relative risk of dropout = 0.49, 95% CI [0.29, 0.82], number needed to treat = 4.8, 95% CI [2.9, 14.7]).	
Kratzer et al. (2019)	Somatoform symptoms (HEALTH-49), complex dissociative disorder, mindfulness (FMI) (3)	None	52% of the sample ($n = 78$) had a reliable change in symptoms. The conditional inference tree identified three predictors of reliable change: Lower somatoform symptom (HEALTH-49) score, absence of diagnosis of a complex dissociative disorder, and higher mindfulness (FMI) score. In the subgroup of patients who scored above 3.29 for somatoform symptoms (node 7; $n = 29$), the rate of reliable improvement fell to 27.6%. In the subgroup of patients who scored below 3.29 for somatoform symptoms and were diagnosed with a complex dissociative disorder (Node 6; $n = 33$), the rate of reliable improvement was 39.4% ($p < 0.05$). The subgroup of patients who scored below 3.29 for somatoform disorders but were not diagnosed with a complex dissociative disorder (node 3; $n = 88$) could be divided further into subgroups based on their mindfulness scores. In the subgroup of patients who scored below 3.29 for somatoform disorders, were not diagnosed with a complex dissociative disorder, and scored above 52 on the FMI (Node 5; $n = 70$), the rate of reliable improvement was 71.4%. In the subgroup of patients who scored below 3.29 for somatoform disorders, were not diagnosed with a complex dissociative disorder, and scored 52 or below on the FMI (Node 4; $n = 18$), the rate of reliable improvement was 38.9%.	Results consistent with evidence that dissociation negatively affects therapy outcome. The significance of somatoform symptoms has been overlooked thus far. The finding that mindfulness deficits impair PTSD treatment outcome is novel. The role of mindfulness in development, maintenance, and treatment of PTSD is unclear. Limitations: Severity of PTSD and comorbid disorders was not assessed through structured interview; the duration of stay was not included as a covariate, and exact duration of trauma-focussed treatment was not recorded for each patient (this could also be included as a covariate); lack of follow-up measures means that long-term benefit of treatment was not assessed.
López-Castro et al. (2021)	Age, slope of improvement in PTSD (MPSS-SR), years since last traumatic event, baseline PTSD severity (CAPS, MPSS-SR), age at earliest traumatic event, slope of improvement in problem substance use, employment, interaction between age and slope of	None	Age × Weekly Improvement in PTSD Symptoms (MPSS-R) interaction effect was significant in the test set (β = 0.0104, SE = 0.0037, $p = .01$), and suggests that older patients with more weekly improvement in symptoms were likely to attend more sessions, whereas younger patients with more weekly improvement in symptoms were likely to attend fewer sessions. This effect was also significant in the validation set (β = 0.0050, SE = 0.0026, $p = .05$). Employment was a significant predictor in the test set (β = 0.213, SE	Random forest can identify reproducible predictors of PTSD/SUD treatment attendance, including an interaction between predictors. Some variables selected by random forest were not found to be statistically significant predictors in the Poisson regression analysis, potentially because they were highly correlated with age, PTSD symptom improvement, and/or

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
	improvement in PTSD (MPSS-SR) (9) Emotion regulation (DERS) and baseline primary substance use were selected by the random forest, but were omitted from the regression model as they were not present in the replication dataset		= 0.0969, *p* = .030), indicating that patients without employment in the past three years were likely to attend more sessions than patients with full-time or part-time jobs. However, this effect was not significant in the validation set (β = 0.014, SE = 0.1211, *p* = .905). In the validation set, Weekly Improvement in PTSD symptoms (MPSS-SR) was a significant predictor of number of sessions attended (β = −0.2219, SE = 0.1117, *p* = .05), but this was not a significant predictor in the training set. No other predictors were significant.	their interaction. Within treatment improvement in symptoms seems to be an important predictor of attendance regardless of treatment type, this is congruent with the only other published study of this variable in PTSD+SUD treatment, however the finding that baseline PTSD did not interact with change in PTSD is not. The finding that baseline employment status was not significant in the validation set reflects broader pattern in the literature of baseline predictors of PTSD+SUD trial attendance failing to replicate. The finding that age interacted with weekly improvement adds depth to previous findings that older age is associated with greater session attendance in clinical trials of treatment for PTSD+SUD. Limitations: Analysis only included those who attended at least one session, therefore this study does not generalise to patients who drop out before starting treatment; limited by sample size and measures of the RCT datasets; the validation dataset was missing two of the candidate predictors of the training dataset; many variables identified by literature review as potentially related to attendance were missing from both datasets, (e.g., income, cognitive functioning, anxiety sensitivity, comorbid personality disorder, therapeutic alliance).
Nixon et al. (2021)	No pre-treatment variables were associated with outcome	Examined AUC-ROC but did not report statistics, only presented and visually interpreted plots	Participants were classified as follows: Clear responders (*n* = 94) reported a reliable (≥ sample RCI) decrease in PDS score by session 6, and did not meet CAPS PTSD diagnostic criteria at post-treatment or follow-up. Delayed responders (*n* = 52) did not report reliable decrease in PDS score by session 6, but did report a reliable decrease and did not meet CAPS PTSD diagnostic criteria at post-treatment or follow-up. Partial responders (*n* = 17) reported a reliable decrease in PDS score	Results of this exploratory study suggest that it is not possible to predict response pattern pre-treatment from the predictors included in this study. Session-by-session progress data is more informative. Limitations: RCT data, findings require replication in other contexts and samples; completer only analysis, provides no information about predictors of dropout, or the relationship between dropout and non-response;

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			at post-treatment or follow-up but still met CAPS PTSD diagnostic criteria. Non-responders (*n* = 16) did not report a reliable decrease in PDS score and still met the CAPS PTSD diagnostic criteria at post-treatment or follow-up. The classes were not distinguishable using pre-treatment predictor data.	potentially important predictors missing from the dataset such as treatment credibility ratings, homework compliance, and therapeutic alliance; used slightly different versions of self-report PTSD and depression measures (PDS used in 3/4 studies and PSS used in the other, BDI-II used in 3/4 studies and BDI used in the other); the best statistical method for analyses such as these is as yet unknown.
Stirman et al. (2021)	Clinician-rated PTSD symptom severity (CAPS), military sexual trauma (MSIW), physical functioning (PCS), mental functioning (MCS), perceived treatment credibility (CEQ) (5)	R^2 and RMSE with 10 fold cross validation, mean averaged over 1000 runs: R^2 = 0.39, 95% CI [0.38587, 0.38789] RMSE = 20.28, 95% CI [20.27, 20.28]	In the final regression model higher baseline PTSD symptoms (β = 0.46, SE = 1.37) and experience of military sexual trauma (β = 0.12, SE = 2.86) were associated with higher posttreatment PTSD symptoms (CAPS), whereas better physical (β = − 0.23, SE = 1.29) and mental functioning (β = − 0.17, SE = 1.41) and higher perception of treatment credibility (β = − 0.17, SE = 1.24) were associated with lower post-treatment PTSD symptoms (CAPS). These variables were combined to form the PI. Regression model predicting post-treatment CAPS from Treatment Type, the PI, and the Treatment Type by PI interaction explained 39% of the variance in post-treatment PTSD severity. The interaction term was significant (β = 0.2999, SE = 0.1526, 95% CI [0.0008, 0.59906], *t* = 1.97, *p* = 0.0494), as well as the effect of Treatment Type (β = −6.7088, SE = 2.4854, 95% CI [−11.580, −1.8375], *t* = −2.70, *p* = 0.0069), and the PI (β = 0.9154, SE = 0.0765, 95% CI [0.7654, 1.06547], *t* = 11.96, *p* < .0001), indicating that the interaction between the PI and treatment type moderated treatment outcome. For the 64% of patients who were predicted the best treatment outcomes, PE was associated with statistically significantly better outcomes than PCT. Whereas for the 36% of patients with the worst predicted treatment outcomes, treatment type was not associated with outcome. PCT was not associated with better treatment outcomes at any point on the PI continuum.	Although on average there was a slight advantage of PE over PCT, for patients with the best overall prognosis PE was associated with substantially better outcomes, but for the patients with poorer prognosis there was no difference in outcome between PE and PCT. Replication is necessary before drawing conclusions. Findings suggest that PI may inform treatment selection by identifying patients for whom trauma focussed interventions confer a significant advantage. Limitations: Variable selection and imputation was performed using the whole sample, risking invalid statistical inference, model overfitting, inflated relationships, and overconfidence; Luedtke et al. (2019) suggest *n* = 300 per treatment arm to detect reliable improvements in outcomes related to treatment selection models, RCTs with samples this size currently unavailable for PTSD, sample size too small to facilitate holdout sample, external validation required; source RCT not designed to inform treatment selection; sample all female, mostly veterans, with high chronicity and diagnostic complexity; PE therapists mostly inexperienced, further research with more diverse patient and therapist characteristics recommended, model may not generalize beyond female military population; other predictor variables (e.g., trauma history

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
				details) and outcome variables (e.g., quality of life, functioning) could be included.
Stuke et al. (2021)	Posttraumatic cognitions (PTCI), centrality of trauma event to person's identity and life story (CES-7), depression (BDI), gender, general psychopathology (BSI), PTSD symptoms (DTS, PDS), comorbid affective disorder, psychosocial functioning (IMET), rumination (PTQ), age, comorbid substance use disorder (12)	Linear regression: $R = 0.214$, $p = .021$ ADAboost regressor: $R = 0.162$, $p = .081$	Univariate correlations indicated that severe posttraumatic cognitions (PTCI; $r = .277$), greater centrality of traumatic event to patients' identity and life story (CES-7; $r = .202$), and more severe depression (BDI; $r = .201$) were associated with poorer outcome.	Treatment outcome could be significantly predicted from pre-treatment total scores on psychometric measures using linear regression. The overall predictive power of the model was low compared to similar studies using machine learning methods, and arguably too low to be clinically useful. No single predictor was particularly strong, but a higher level of posttraumatic cognitions was the strongest predictor of poor outcome. This finding is in line with one previous study but contradicts numerous others that found no effect of posttraumatic cognitions. The finding that higher baseline PTSD predicted greater improvement in symptoms is in-line with some previous studies but contradicts others that found the opposite effect (could be due to ceiling effect or over-reporting of baseline symptoms). There is a need to identify more powerful predictor variables and include interactions. Study strengthened by cross-validation methods and naturalistic design. Limitations: Routinely collected data, self-report outcome, lack of control group, lack of formal control for therapy adherence, may have omitted important predictor variables (e.g., employment and social problems), small sample size, large proportion of participants excluded for missing values or dropout, cannot rule out bias due to completer only analysis.
Zhang et al. (2021)	Resting state EEG/PEC features primarily selected from the beta frequency band and eyes-open	None	In the first sample, cluster analysis identified two subtypes in the subsample that met diagnostic criteria for PTSD (*n* = 106), and the stability analysis (100 repetitions on random 90% subsamples)	Subtype 1 represents a subgroup of patients who meet CAPS/CAPS-5 diagnostic criteria for PTSD but do not differ biologically from healthy controls. This is consistent with

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
	condition (NR)		confirmed this as the most stable solution. The clustering stability was significantly lower and more variable in the subsample of healthy controls (n = 95). The two subtypes primarily differed in PEC patterns in regions located in the frontoparietal control network and the default mode network; compared with subtype 2, subtype 1 PEC was stronger between the frontal cortex and other regions but weaker between the parietal cortex and other regions. There were no significant differences in clinical or demographic variables between the two subtypes. The cluster analysis was successfully replicated in the second sample. Again, there was no significant difference in clinical or demographic variables between subtype 1 and 2. In the first cohort (n = 72), subtype 1 had significantly better treatment outcomes than subtype 2 (group x time interaction: $F(1,123) = 9.04$, $p = 0.0032$, Cohen's $d = 0.80$ for CAPS-IV; $F(1,123) = 4.38$, $p = 0.039$, Cohen's d = 0.59 for CAPS-5). Splitting the cohort by treatment, subtype 1 had significantly better outcomes for both PE ($F(1,38) = 7.23$, $p = 0.011$ for CAPS-IV; $F(1,38) = 2.90$, $p = 0.097$ for CAPS-5) and CPT ($F(1,81)$ = 4.75, $p = 0.032$ for CAPS-IV; $F(1,81)=2.41$, $p = 0.12$ for CAPS-5). This was replicated in the second cohort (n = 63): Subtype 1 had significantly better treatment outcomes than subtype 2 (group × time interaction: $F(1,109) = 4.76$, $p = 0.031$, Cohen's $d = 0.56$ for CAPS-IV; $F(1,109) = 4.46$, $p = 0.037$, Cohen's $d = 0.55$ for CAPS-5). Splitting the cohort by treatment, Subtype 1 had significantly better outcomes for both PE ($F(1,34) = 3.09$, $p = 0.088$ for CAPS-IV; $F(1,34)$ = 9.31, $p = 0.0044$ for CAPS-5) and CPT ($F(1,71)=2.13$, $p = 0.15$ for CAPS-IV and $F(1,71) = 0.74$, $p = 0.39$ for CAPS-5). Comparing the percentage of treatment responders, there were significantly more treatment responders in subtype 1 than subtype 2 ($X^2 = 4.07$, $p = 0.044$, number needed to treat = 5.1 for CAPS-IV). The clustering analysis was repeated using only clinical and	previous psychiatric neuroimaging findings. However, it may be that patients in subtype 1 differ from healthy controls on other neurological variables that were not measured in this study. Unlike previous studies, no *a priori* assumptions were made about brain regions or frequency bands of interest, leading to unbiased, data-driven, identification of significant biomarkers, which generalized across independent datasets. The findings of this study support the findings of previous fMRI studies. Limitations: Analysis of clusters as predictors of treatment response require replication; EEG vulnerable to confounding neural signals and artifacts of volume conduction; possible change in neural connectivity over time was not accounted for in this study.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			demographic variables, and the subtypes identified using PEC data could not be identified using clinical and demographic data alone.	
Zhutovsky et al. (2019)	A network centred around the pre-supplementary motor area (NR)	Balanced accuracy = 81.4% Sensitivity = 84.8% Specificity = 78% AUC-ROC (SD) = 0.929 (SD = 0.149)	Of the *n* = 44 participants treated for PTSD, *n* = 24 met the criteria for treatment response and *n* = 20 did not. The univariate group analysis indicated heightened connectivity in the frontal polar area in non-responders, particularly in the right superior frontal gyrus (*p* FWE = 0.04). The multivariate Gaussian process classification analysis identified a network centred around the pre-supplementary motor area that could be used to classify responders and non-responders with a high degree of accuracy. After Bonferroni correction was applied, no other areas were identified as significant predictors, including the frontal polar area, however this area was significant before applying the Bonferroni correction.	The results demonstrate that it is feasible to predict individual response to PTSD treatment using resting state fMRI data, this provides a proof-of-concept that PTSD treatment can be personalized using biomarker predictors. Considered alongside previous research into depression treatment response, the results suggest that pre-supplementary motor area connectivity may influence response to treatment regardless of treatment type or disorder. The network identified by this study is not comparable to the ventral attention network investigated by Etkin et al. (2019). Although also using resting state fMRI, Etkin et al. (2019) discovered the ventral attention network by comparing patients with PTSD to healthy controls, rather than comparing responders to non-responders, and did not investigate any networks other than ventral attention network. Therefore, the predictive power of both the pre-supplementary motor area and ventral attention network still require replication. However, both studies demonstrate that resting state fMRI may be used to predict individual treatment response in PTSD. Findings contrast with previous research, this may be due to methodological differences such as previous studies investigating a pre-defined region of interest (as opposed to the data-driven, whole brain approach taken here), previous studies primarily focussed on task-induced changes rather than resting state fMRI, different PTSD populations and types of trauma, different treatment and outcome criteria, different study designs and methods of measurement/analysis.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
				Limitations: Sample size small for machine learning methods; all male veteran sample limits generalizability; mix of treatments means that results are not specific to one particular treatment, and patient by treatment interactions may be obscured.
Zilcha-Mano et al. (2020)	Within-network connectivity in the Executive Control Network (Lateral Prefrontal Cortex right – Posterior Parietal Cortex right; Frontal Pole right – Lateral Prefrontal Cortex right) (NR)	None	Support vector machine identified 18 functional connectivity features that distinguished participants with PTSD (with or without depression) from trauma-exposed healthy controls. Participants with PTSD (with or without depression) displayed lower connectivity in the Within-Executive Control, Within-Salience, Salience–Dorsal Attention, Salience–Default Mode, and Default Mode–Executive Control Networks, and higher connectivity in the Default Mode–Dorsal Attention and Salience–Default Mode Networks (compared to controls). Support vector machine identified 20 functional connectivity features that distinguished participants with PTSD from patients with PTSD+Depression. For those with PTSD alone, within-network connectivity was higher in the Basal Ganglia Network, but lower in the Executive Control, Salience, and Dorsal Attention Networks. The model could classify participants with or without PTSD with an accuracy of 70.6%, and classify those with PTSD with or without MDD with an accuracy of 76.7%. Of the functional connectivity features that distinguished participants with PTSD (with or without MDD) from trauma-exposed healthy controls, there was a significant, positive correlation between within-network connectivity in the Executive Control Network and reduction in PTSD symptoms over the course of PE (Lateral Prefrontal Cortex right–Posterior Parietal Cortex right: $r = .455$, $p < .001$; Frontal Pole right-Lateral Prefrontal Cortex right: $r = .415$, $p = .002$). Such that participants who showed higher levels of functional connectivity in the Executive Control Network pre-treatment reported a greater reduction in PTSD symptoms post-treatment. No significant correlations with change in depression	The model identified baseline differences in functional connectivity between patients with PTSD (with or without concurrent depression) and healthy controls, which were significantly associated with change in symptoms over the course PE for PTSD and are therefore clinically useful. The finding that differences in functional connectivity in the identified networks distinguish patients with PTSD from trauma-exposed healthy controls is in-line with some previous findings but not others. The finding that the biomarkers distinguishing PTSD alone from PTSD+Depression did not significantly correlate with treatment response may be due to fact that the treatment was PTSD focussed, and not a treatment for depression. Limitations: Combined data from three trials with different exclusion criteria and different MRI scanners; focus on difference between disorders ignored heterogeneity within disorders and transdiagnostic processes; assumed that PTSD and PTSD+Depression are distinct subpopulations, unsupervised ML may be better suited to investigate this; sample size prohibited exploration of symptom cluster associations with functional connectivity; lack of non-trauma-exposed control group prevented exploration of the effect of trauma exposure.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			were observed, and there were no significant correlations between the biomarkers identified by the PTSD vs. PTSD+Depression classification and change in PTSD or depression symptoms.	

Note. AUC-PR = Area Under the Precision-Recall Curve; AUC-ROC = Area Under the Receiver Operating Characteristic Curve; AUDIT = Alcohol Use Disorder Identification Test; BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory – II; BSI = Brief Symptom Inventory; CAPS = Clinician-Administered PTSD Scale; CAPS-5 = Clinician-Administered PTSD Scale for DSM-5; CEQ = Credibility/Expectancy Questionnaire; CES-7 = Centrality of Event Scale; CHAID = Chi-square Automatic Interaction Detection; CI = Confidence Interval; CPT = Cognitive Processing Therapy; CTQ = Childhood Trauma Questionnaire; DERS = Difficulties in Emotion Regulation Scale; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; DSM-IV-TR = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision; DTS = Davidson Trauma Scale; EEG = Electroencephalography; EMDR = Eye Movement Desensitisation and Reprocessing; EQ-5D-5L = EuroQoL 5 Dimensions 5 Levels; FMI = Freiburg Mindfulness Inventory; fMRI = Functional Magnetic Resonance Imaging; FWE = Family Wise Error; HADS = Hospital Anxiety and Depression Scale; HEALTH-49 = Hamburg Modules for the Assessment of Psychosocial Health; IES-R = Impact of Event Scale – Revised; IMET = Index zur Messung von Einschränkungen der Teilhabe; IPE = Intensified Prolonged Exposure; IQ = Intelligence Quotient; MAE = Mean Absolute Error; MINI = Mini International Neuropsychiatric Interview; MMPC LR = Logistic Regression with Max-Min Parent-Child variable selection; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; MOS = Medical Outcome Study; MPSS-SR = Modified Post-traumatic Stress Disorder Symptom Scale Self-Report; MSIW = Military Stress Inventory for Women; NR = Not Reported; OR = Odds Ratio; PAI = Personalised Advantage Index; PCL-5 = PTSD Checklist for DSM-5; PCS = Physical Component Summary Scale; PDS = Posttraumatic Stress Diagnostic Scale; PE = Prolonged Exposure; PEC = Power Envelope Connectivity; PHQ-15 = Patient Health Questionnaire – 15; PHQ-9 = Patient Health Questionnaire-9; PI = Prognostic Index; PSS = Post-traumatic Symptoms Scale; PTCI = Posttraumatic Cognitions Inventory; PTQ = Perseverative Thinking Questionnaire; PTSD = Post-Traumatic Stress Disorder; RCT = Randomised Controlled Trial; RMSE = Root Mean Square Error; SD = Standard Deviation; SE = Standard Error; STAIR+PE = Skills Training in Affective and Interpersonal Regulation plus Prolonged Exposure; SUD = Substance Use Disorder; Tf-CBT = Trauma-focussed Cognitive Behavioural Therapy; URICA = University of Rhode Island Change Assessment; WSAS = Work and Social Adjustment Scale.

Supplementary Table 5

Candidate predictor variables

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
Deisenhofer et al. (2018)	Depression (PHQ-9) Functional impairment (WSAS) Anxiety (GAD-7) Long-term medical condition (yes / no) Disability (yes / no) Antidepressant medication (prescribed / not prescribed) Gender (male / female) Age Employment status (employed, student / unemployed, long-term sick, other)	9	Genetic algorithm	*Tf-CBT:* Functional impairment (WSAS) Employment status Age Gender (4) *EMDR:* Depression (PHQ-9) Medication (2)
Etkin et al. (2019)	MRI EEG Sustained attention (task) Working memory (task) Verbal memory (task) Inhibitory control (task) Response inhibition (task) Flexibility (task) Processing speed (task)	Unclear	Generalized linear modelling	Verbal memory delayed recall impairment Within ventral attention network functional connectivity (2)
Fleming et al. (2018)	Marital Status (single / married / divorced / widow(er) / separated engaged) Military Branch (Army / Navy / Marines / Air Force / National Guard / multiple branches) Active Duty (yes / no) Conflict (OIF / OEF / both)	51	Exhaustive CHAID	Days between introductory session and therapy invitation URICA item 9 ("*I have been successful in working on my trauma issues, but I'm not sure I can keep up the effort on my own*") URICA item 5 ("*I'm not the one with trauma issues, it doesn't make much sense for me to be here*")

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Prior treatment for PTSD (yes / no)			Presence of traumatic brain injury
	Number of children			Prior PTSD treatment (5)
	Number of combat tours			
	Years of education			
	Traumatic brain injury severity (none / mild / moderate or severe)			
	Depression (yes / no)			
	Anxiety disorder (yes / no)			
	Alcohol use disorder (yes / no)			
	Substance use disorder (yes / no)			
	Psychotropic medication (taking / not taking)			
	Trauma type (combat only / sexual only / both)			
	Pre-military trauma (yes / no)			
	Pending disability claim (yes / no)			
	Total VA Disability Rating (10% or below / 20-40% / 50-70% / ≥ 80%)			
	PTSD Disability Rating (10% or below / 20-40% / 50-70% / ≥ 80%)			
	PTSD symptom severity (PCL total, subscales, and items)			
	Days between mental health consult and mental health appointment			
	Days from psychotherapy referral to introduction session			
	Days from completed introduction session to therapy invitation			
	Motivational readiness to address trauma (URICA total, subscales, and items)			
	PTSD Information Session Survey (total and items [*n* = 9])			
	Format of introduction session contact (group / individual)			
	Weighs pros or cons more (pros / cons / equal weight)			

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Readiness category (precontemplation / contemplation / preparation) Treatment choice (PE / CPT group / CPT individual / CPT individual or group / other / none)			
Forbes et al. (2003)	Psychopathology (MMPI-2 clinical subscales [*n* = 10] validity subscales [*n* = 3] and factors [*n* = 3]) PTSD (PCL) Anxiety (HADS) Depression (HADS) Alcohol use (AUDIT) Combat exposure (CES)	21	Ward's hierarchical cluster analysis	Clinical subscales and validity subscales of MMPI-2 anxiety and depression (HADS) Alcohol use (AUDIT) (16)
Held et al. (2022)	Age Gender (male / female) Race (white / all other races) Ethnicity (non-Hispanic / Hispanic or Latino) Education (8 categories dummy coded) Marital Status (married or domestic partnership / not married) Military Branch (5 categories dummy coded) Deployed (yes / no) Served in military after 11th September 2001 (yes / no) Military sexual trauma (three items measured yes / no) Referral source (6 categories dummy coded) Clinician assessed PTSD symptoms (CAPS-5) Self-reported PTSD symptoms (PCL-5) Depression (PHQ-9) Alcohol use disorder (AUDIT-C) Neurobehavioral symptom exaggeration (NSI-Valid)	104	Elastic net classification Gradient boosted models Random forest Ridge classification Logistic regression Logistic regression with Max-Min Parent-Child variable selection	*Gradient Boosted Models predicting Fast Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5) Post-traumatic cognitions (item level responses to PTCI) (10) *Elastic Net predicting Minimal Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5) Post-traumatic cognitions (item level responses to PTCI) Marriage/domestic partnership status Age Gender Level of education Military service branch (60)

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Trauma related cognitions (PTCI)			
Hendriks et al. (2018)	Age Educational level Living alone PTSD symptom severity (PSS-SR) Depression (BDI-II) Dissociative symptom severity (DES) Borderline personality disorder (BPD-47) Psychoactive medication use Fear activation during first exposure session Within-session fear habituation during the first session Between session fear habituation	14	Stepwise multinominal logistic regression	Living condition Between-session fear habituation (2)
Herzog et al. (2021)	Age Gender Number of children Inability to work Wish to retire School-leaving qualification Professional qualification Occupational status (6 categories dummy coded) Marital status (4 categories dummy coded) In a relationship (yes / no) Living situation (7 categories dummy coded) Previous outpatient psychiatric care Previous outpatient psychotherapy	≥ 46 (unclear)	Elastic net	PTSD severity (IES-R total) Psychoticism (BSI) Avoidance (IES-R subscale) Wish to retire Depression (BDI-II) Number of comorbid diagnoses Age Bronchial asthma Somatic symptoms (PHQ-15) Outpatient psychiatric care Children Being retired Work disability in past year Outpatient psychotherapy

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	ICD-10 medical diagnoses (dummy coded, including but not limited to: Bronchial Asthma; Endocrine, nutritional and metabolic disease; Personality Disorder; Obesity) Number of comorbid diagnoses PTSD symptoms (IES-R) Psychiatric symptoms and distress (BSI, 9 subscales) Life satisfaction (SWLS) Depression symptoms (BDI-II) Depression, anxiety, and somatic symptoms (PHQ, 3 scales) Psychosocial functioning (GAF) Work disability last year Duration of work disability			Somatization (BSI) (15)
Hoeboer et al. (2021)	Patient expectancies (expectancy of burden and credibility questionnaire) Age Gender Cultural background Education Employment Social support (MOS) Childhood trauma background (CTQ) General health status (EQ-5D-5L) Depression (BDI) Post-traumatic cognitions (PTCI) Interpersonal problems (IIP) Self-esteem (RSES) Emotion regulation difficulties (DERS)	24	Boruta algorithm	*CAPS-5 change following PE/IPE:* Depression (BDI) Social support (MOS) Concurrent mental health problems (MINI axis 1) Childhood sexual abuse (CTQ) (4) *CAPS-5 change following STAIR+PE:* General health status (EQ-5D-5L) Emotion regulation difficulties (DERS) PTSD severity (CAPS-5) (3) *PCL-5 change following PE or IPE:* Depression (BDI) Social support (MOS) (2) *PCL-5 change following STAIR and PE:* General health status (EQ-5D-5L)

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Somatoform dissociation (SDQ-5) Presence of personality disorder (SCID-2) Number of DSM-IV-defined Axis-1 disorders excluding PTSD (MINI) Dissociation (DSP-I) PTSD symptom severity (CAPS-5) Psychotropic medication			Emotion regulation difficulties (DERS) (2)
Keefe et al. (2018)	Age Race Years of education Estimated IQ Years since index rape Severity of childhood physical abuse Severity of childhood sexual abuse Abuse by current partner Total sex crime exposures CAPS total score PTSD avoidance (PSS) PTSD arousal (PSS) PTSD re-experiencing (PSS) Depression (BDI-II) Dissociation (DES) Hopelessness (BHS) Trait anger (STAXI) Total trauma cognitions (TRGI)	20	Bootstrapped, random forest variant of model-based recursive partitioning, and bootstrapped variant of an AIC-based backward selection model	*Prescriptive variables:* Childhood physical abuse Current relationship abuse Trait anger Race *Prognostic variables:* Years of education Estimated IQ score (6)
Kratzer et al. (2019)	Depression (HEALTH-49)	5	Conditional inference tree	Somatoform symptoms (HEALTH-49)

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Somatoform symptoms (HEALTH-49)			Complex dissociative disorder
	Well-being (HEALTH-49)			Mindfulness (FMI) (3)
	Presence of a complex dissociative disorder			
	Mindfulness (FMI)			
	Candidate variables excluded through univariate analysis:			
	Gender			
	Age			
	Presence of a personality disorder			
	Childhood trauma (CTQ)			
	PTSD symptoms (IES-R)			
	Interactional difficulties (HEALTH-49)			
	Self-efficacy (HEALTH-49)			
	Phobic anxiety (HEALTH-49)			
	Social distress (HEALTH-49)			
	Social support (HEALTH-49)			
	Activity and participation (HEALTH-49)			
	Dissociation (DES-T)			
López-Castro et al. (2021)	Age	28	Random forest	Age
	Gender			Slope of improvement in PTSD (MPSS-SR)
	Race and ethnicity			Years since last traumatic event
	Employment			Baseline PTSD severity (CAPS, MPSS-SR)
	Education			Age at earliest traumatic event
	Marital status			Slope of improvement in problem substance use
	Severity of substance use			Employment
	Substance use disorder type (alcohol vs. substance)			

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Comorbid substance use disorder diagnosis PTSD symptom severity (CAPS) Trauma characteristics (more than one traumatic event, number of traumatic events, trauma before age 18, age of first trauma, sexual assault, physical assault, other trauma, accident) Type of intervention (COPE vs. RPT) Depression diagnosis Emotion regulation (DERS) Within-treatment substance use Within treatment PTSD symptom change (MPSS-SR)			Interaction between age and slope of improvement in PTSD (MPSS-SR) (9) Emotion regulation (DERS) and baseline primary substance use were selected by random forest, but were omitted from the regression model as they were not present in the replication dataset
Nixon et al. (2021)	Race Marital status Income level Age Years of education Childhood sexual abuse Childhood physical abuse Endorsed adult sexual violence Adult physical violence Intimate partner violence independent of other forms of violence Index trauma type Depression diagnosis Depression severity (BDI) Panic disorder diagnosis PTSD symptoms (PDS, CAPS)	38	Random forest	No pre-treatment variables were associated with outcome (0)
Stirman et al. (2021)	Clinician-rated PTSD symptom severity (CAPS)	29	Elastic net with stepwise AIC-penalized bootstrapped variable selection	Clinician-rated PTSD symptom severity (CAPS)

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Re-experiencing (PCL)			Military sexual trauma (MSIW)
	Avoidance (PCL)			Physical functioning (PCS)
	Numbing (PCL)			Mental functioning (MCS)
	Hyperarousal (PCL)			Perceived treatment credibility (CEQ) (5)
	Time since trauma (CAPS)			
	Sexual index trauma (CAPS)			
	Military sexual trauma (MSIW)			
	Military stress exposure (MSIW)			
	Number of trauma types			
	Age			
	White race			
	College Education			
	Married/living as married			
	Working			
	Current mood disorder (SCID-P)			
	Current anxiety disorder (SCID-P)			
	Borderline personality disorder (SCID-P)			
	Other personality disorder (SCID-P)			
	Depression symptoms (BDI)			
	Anxiety symptoms (SSAI)			
	Dissociative symptoms (TSI)			
	Anger symptoms (TSI)			
	Physical Functioning (PCS)			
	Mental Functioning (MCS)			
	Self-reported quality of life (QOLI)			
	Treatment credibility (CEQ)			

Study	Candidate predictor variables	N variables	Predictor selection method	Predictors in final model(s) (*n* predictors in model)
	Psychoactive medication use at screening Benzodiazepine use at screening			
Stuke et al. (2021)	Age Gender Comorbid ICD-10 psychiatric disorder Trauma details (LEC scale of PDS) PTSD symptoms (DTS, PDS) Posttraumatic cognitions (PTCI) Centrality of trauma event to person's identity and life story (CES-7) Rumination (PTQ) Depression (BDI) General psychopathology (BSI) Psychosocial functioning (IMET)	12	Principal components analysis	Posttraumatic cognitions (PTCI) Centrality of trauma event to person's identity and life story (CES-7) Depression (BDI) Gender General psychopathology (BSI) PTSD symptoms (DTS, PDS) Comorbid affective disorder Psychosocial functioning (IMET) Rumination (PTQ) Age Comorbid substance use disorder (12)
Zhang et al. (2021)	PEC features from eight EEG conditions (four frequency bands across eyes-open/eyes-closed conditions)	Unclear	Sparse k-means clustering	Resting state EEG/PEC features primarily selected from the beta frequency band and eyes-open condition (NR)
Zhutovsky et al. (2019)	Structural MRI and resting state functional MRI data with age and total intracranial volume as covariates	Unclear	Gaussian process classifier	A network centred around the pre-supplementary motor area (NR)
Zilcha-Mano et al. (2020)	43 MRI regions of interest	Unclear	Support vector machine	Within-network connectivity in the Executive Control Network (Lateral Prefrontal Cortex right – Posterior Parietal Cortex right; Frontal Pole right – Lateral Prefrontal Cortex right) (NR)

Note. AIC = Akaike Information Criterion; AUDIT = Alcohol Use Disorder Identification Test; AUDIT-C = Alcohol Use Disorder Identification Test - Consumption; BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory - II; BHS = Beck Hopelessness Scale; BPD-47 = Borderline Personality Disorder symptom checklist; BSI = Brief Symptom Inventory; CAPS = Clinician-Administered PTSD Scale; CAPS-5 = Clinician-Administered PTSD Scale for DSM-5; CEQ = Credibility/Expectancy Questionnaire; CES = Combat Exposure Scale; CES-7 = Centrality of Event Scale; CHAID = Chi Squared Automatic Interaction Detection; COPE = combined prolonged exposure and relapse prevention therapy; CTQ = Childhood Trauma Questionnaire; DERS = Difficulties in Emotion Regulation Scale; DES = Dissociative Experiences Scale; DES-T = Dissociative Experiences Scale - Taxon; DSP-I = Dissociative subtype of PTSD Interview; DTS = Davidson Trauma Scale; EEG = Electroencephalography; EQ-5D-5L = EuroQoL 5 Dimensions 5 Levels; FMI = Freiburg Mindfulness Inventory; GAD-7 = Generalized Anxiety Disorder 7; GAF = Global Assessment of Functioning; HADS = Hospital Anxiety and

Depression Scale; HEALTH-49 = Hamburg Modules for the Assessment of Psychosocial Health; IES-R = Impact of Event Scale - Revised; IIP = Inventory of Interpersonal Problems; IMET = Index zur Messung von Einschränkungen der Teilhabe; LEC = Life Events Checklist; MCS = Mental Component Summary Scale; MINI = Mini International Neuropsychiatric Interview; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; MOS = Medical Outcome Study; MPSS-SR = Modified PTSD Symptom Scale Self-Report; MRI = Magnetic Resonance Imaging; MSIW = Military Stress Inventory for Women; NSI-Valid = Neurobehavioral Symptoms Inventory Validity-10; PCL = PTSD Checklist; PCL-5 = PTSD Checklist for DSM-5; PCS = Physical Component Summary Scale; PDS = Posttraumatic Stress Diagnostic Scale; PEC = Power Envelope Connectivity; PHQ = Patient Health Questionnaire; PHQ-9 = Patient Health Questionnaire-9; PSS = Post-traumatic Symptoms Scale; PSS-SR = PTSD Symptom Scale, Self-Report; PTCI = Posttraumatic Cognitions Inventory; PTQ = Perseverative Thinking Questionnaire; PTSD = Post-traumatic Stress Disorder; QOLI = Quality of Life Inventory; RPT = Relapse Prevention Therapy; RSES = Rosenberg Self-esteem Scale; SCID-2 = Clinical Interview for DSM-IV Personality Disorders; SCID-P = Structured Clinical Interview for DSM-IV Patient Version; SDQ-5 = Somatoform Dissociation Questionnaire ; SSAI = Spielberger State Anxiety Inventory ; STAXI = State-Trait Anger Expression Inventory; SWLS = Satisfaction With Life Scale; TRGI = Trauma-Related Guilt Inventory; TSI = Trauma Symptom Inventory; URICA = University of Rhode Island Change Assessment; WSAS = Work and Social Adjustment Scale.