HHS Public Access

Author manuscript

J Anxiety Disord. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

J Anxiety Disord. 2018 December ; 60: 35–42. doi:10.1016/j.janxdis.2018.10.004.

Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization

Santiago Papini, MA[a], **Derek Pisner, BA**[a], **Jason Shumake, PhD**[a], **Mark B. Powers, PhD**[a,b], **Christopher G. Beevers, PhD**[a], **Evan E. Rainey, MS**[b], **Jasper A.J. Smits, PhD**[a], **Ann Marie Warren, PhD**[b]

[a]Department of Psychology and Institute for Mental Health Research, The University of Texas at Austin

[b]Baylor University Medical Center

Abstract

Posttraumatic stress disorder (PTSD) develops in a substantial minority of emergency room admits. Inexpensive and accurate person-level assessment of PTSD risk after trauma exposure is a critical precursor to large-scale deployment of early interventions that may reduce individual suffering and societal costs. Toward this aim, we applied ensemble machine learning to predict PTSD screening status three months after severe injury using cost-effective and minimally invasive data. Participants (N = 271) were recruited at a Level 1 Trauma Center where they provided variables routinely collected at the hospital, including pulse, injury severity, and demographics, as well as psychological variables, including self-reported current depression, psychiatric history, and social support. Participant zip codes were used to extract contextual variables including population total and density, average annual income, and health insurance coverage rates from publicly available U.S. Census data. Machine learning yielded good prediction of PTSD screening status 3 months post-hospitalization, AUC = 0.85 95% CI [0.83, 0.86], and significantly outperformed all benchmark comparison models in a cross-validation procedure designed to yield an unbiased estimate of performance. These results demonstrate that good prediction can be attained from variables that individually have relatively weak predictive value, pointing to the promise of ensemble machine learning approaches that do not rely on strong isolated risk factors.

1. Introduction

Posttraumatic stress disorder (PTSD) is characterized by heterogeneous combinations of re-experiencing, avoidance, cognitive, mood, and hyperarousal symptoms (Galatzer-Levy & Bryant, 2013). Although evidence from large epidemiological studies (Goldstein et al., 2016) and meta-analyses of trauma-exposed samples (Santiago et al., 2013) indicates recovery from traumatic stressors is the modal outcome, a substantial minority of trauma-exposed individuals develop a chronic form of PTSD that is associated with psychiatric

Corresponding Author: Santiago Papini, Institute for Mental Health Research, The University of Texas at Austin, 305 E. 23rd St., Stop E9000, Austin, TX, 78712, 512-471-7694 | spapini@utexas.edu.

Conflicts of Interest: The authors have declared that no competing interests exist.

The publisher version can be found at: https://doi.org/10.1016/j.janxdis.2018.10.004

comorbidity (Pietrzak, Goldstein, Southwick, & Grant, 2011) and generally follows a debilitating course (Kessler, 2000). Although evidence-based PTSD treatments are available (Watts et al., 2013), interventions delivered shortly after trauma but *prior to disorder onset* may reduce individual suffering and societal costs (Kearns, Ressler, Zatzick, & Rothbaum, 2012; Rothbaum et al., 2012). Efficient implementation of such preventive strategies requires a reliable and cost-effective approach to prospective prediction of PTSD.

Efforts to identify key PTSD predictors are facilitated by the diagnostic requirement that an external stressor precede symptom emergence (American Psychiatric Association, 2013), which provides a concrete point of reference for prediction. Prospective PTSD prediction from data collected pre-trauma is possible among subpopulations where trauma exposure is likely, such as soldiers (Beevers, Lee, Wells, Ellis, & Telch, 2011; Polusny et al., 2011; Telch et al., 2015; Telch, Rosenfield, Lee, & Pai, 2012) and police (Galatzer-Levy et al., 2013; Galatzer-Levy, Steenkamp, et al., 2014; Pole et al., 2009), or on data collected post-trauma but prior to PTSD onset from emergency room (ER) admits (Ehring, Ehlers, Cleare, & Glucksman, 2008; Morris, Hellman, Abelson, & Rao, 2016; van der Velden & Wittmann, 2008; Walsh et al., 2013) as well as survivors of terrorist attacks (Neria, DiGrande, & Adams, 2011) and other disasters (Neria, Nandi, & Galea, 2008). In contrast to prospective analyses, retrospective reports and controlled comparisons with trauma-exposed individuals without PTSD have identified biopsychosocial correlates of the disorder (D. G. Baker, Nievergelt, & O'Connor, 2012; Broekman, Olff, & Boer, 2007; Etkin & Wager, 2007; Liberzon & Sripada, 2007; Smoller, 2016). Together, these parallel approaches have elucidated several dimensions of PTSD, including etiological mechanisms (Admon, Milad, & Hendler, 2013) and potential treatment targets (Lokshina & Liberzon, 2017).

Despite these important advances, there has been limited impact on the application of preventive care because practical and methodological limitations impede development and implementation of predictive algorithms for this purpose. For example, amygdala and cingulate cortex abnormalities (Admon, Milad, & Hendler, 2013), the transporter linked polymorphic region (5-HTTLPR) genotype (Telch, et al. 2015), and eye gaze avoidance of fearful faces (Beevers et al. 2011) have all been found to predict PTSD, and though neuroimaging, genetic, and eye-tracking data may be critical to uncovering biological mechanisms, these data modalities may be too time consuming, invasive, or expensive to justify routine measurement for predictive purposes. Importantly, previous studies suggest that many routinely collected or easily collected hospital variables following traumatic injury may also be used to predict later PTSD including: age, dissociation, employment status, ethnicity, gender, heart rate, injury severity, mood/anxiety (current & history), race, social support, TBI, and trauma type (e.g. Galatzer-Levy et al. 2014; Morris et al. 2016). General linear model (GLM) approaches are limited when a large number of predictors are used to model a relatively small sample (Iniesta, Stahl, & McGuffin, 2016). The resulting models are prone to predicting the noise that is unique to a sample as opposed to the signal that is common across samples, such that the model gets worse at predicting future data as it gets better at predicting current data – a problem referred to as *overfitting* (Babyak, 2004; Yarkoni & Westfall, 2017). Moreover, the emphasis on isolating key causal features may be misguided, given the symptom heterogeneity, high comorbidity, and the variety of causal pathways involved in PTSD (Ruglass, Lopez-Castro, Cheref, Papini, & Hien, 2014).

Prediction for the purposes of guiding prevention is particularly sensitive to these limitations on feasibility and generalizability.

Ensemble machine learning (ML) approaches overcome these limitations by relying on the collective strength of simple models built from individually weak predictors. Recently, ML prediction of PTSD after ER hospitalization (Galatzer-Levy, Karstoft, Statnikov, & Shalev, 2014; Galatzer-Levy, Ma, Statnikov, Yehuda, & Shalev, 2017), yielded fair-to-good performance based on area under the receiver operator characteristic curve (AUC), a metric that incorporates both prediction sensitivity and specificity where AUC = 0.50 is prediction at chance and AUC = 1.00 is perfect prediction. One of these studies ($N = 957$) used hospital, demographic, and psychological features collected at the ER and in interviews 10 days post-hospitalization to predict non-remitting PTSD through 15 months of follow-up with AUC = 0.78, 95% CI, [0.74, 0.82] (Galatzer-Levy, Karstoft, et al., 2014). In another study ($N = 152$), models that included only ER features yielded good prediction, AUC = 0.82, 95% CI, [0.80, 0.85], which improved with the inclusion of follow-up data (Galatzer-Levy et al., 2017).

We built upon this prior work in several meaningful ways relevant to the implementation of preventive interventions for PTSD. First, instead of predicting non-remitting PTSD trajectories, we applied ML to predict PTSD diagnostic status 3-months post-hospitalization, which corresponds to chronic PTSD in DSM-IV (American Psychiatric Association, 2000) nosology, because preventive care may be desirable for these individuals—even if they remit several months to over one year later. Second, keeping in mind clinical application and implementation, our approach emphasizes feasibility: predictive features were extracted from routine hospital assessments and self-report instruments that can be administered with minimal staff training and participant burden. Consistent with our goal to use cost-effective and minimally invasive data, we selected measures that were either already part of the standard hospital assessment (i.e., Hospital measures), and validated psychological self-report measures of features that have been previously found to predict PTSD including past history of psychiatric diagnoses, current depression, posttraumatic stress symptoms, resilience, and social support (i.e., Psychological Measures). Additionally, we used participant zip codes to extract contextual variables from publicly available U.S. Census data. Third, we evaluated whether a machine learning model with all information outperformed several benchmarks including a machine learning model that utilized only information routinely collected at the ER, and a simple logistic regression model with only the strongest individual predictor (PTSD severity at the hospital). As opposed to the common comparisons to no-information or prediction-at-chance models, we selected benchmarks that allowed us to evaluate whether the addition of features and the complexity of the analytic approach provided a meaningful contribution to prediction (DeMasi, Kording, & Recht, 2017). Finally, we discuss considerations relevant to preventive care, which may guide future development and application of predictive algorithms.

2. Method

2.1 Participants and Study Design

Participants (N = 271) admitted to the Level I Trauma Center at Baylor University Medical Center between March 2012 and May 2014 enrolled in the study if they were at least 18 years old, spoke English or Spanish, provided informed consent and a phone number for follow-up. Eligible participants were interviewed at bedside by trained staff. Procedures were approved by the medical center's Institutional Review Board. A preliminary analysis of a subset of participants collected through 2013 (N = 227) used logistic regression to identify PTSD predictors (Powers et al., 2014). This prior model reduced the number of features by first categorizing them into broad domains such as demographic, psychological, hospital, injury, and substance use, and eliminating variables that did not have significant predictive value in a step-wise fashion. The final model included predictors, without interactions, from all domains that survived the elimination process. The key distinction of our decision tree modeling approach is that it incorporated all predictive features and allowed them to interact in complex nonlinear ways. Importantly, we protected against overfitting – which can lead to biased, inflated estimates of performance – by evaluating our model on a subsample of the dataset that was not included in the model-building process.

2.2. Measures

2.2.1 Predictive Features—Hospital measures including pulse, length of stay in the intensive care unit, overall length of stay, insurance (uninsured, public, or private), injury etiology (e.g., fall, motor vehicle collision, assault) and type (e.g., penetrating, orthopedic), the Glasgow Coma Scale (Rowley & Fielding, 1991), which assesses level of consciousness after injury, and the Injury Severity Score (S. P. Baker, O'Neill, Haddon Jr, & Long, 1974), which measures overall severity from multiple injuries across anatomical regions, were extracted from electronic medical records. Patients provided demographic features including age, race, ethnicity, marital status, employment, education, and annual income. An in-house search engine developed in Python used participant zip codes to extract contextual variables from publicly available U.S. Census data, including land area, population total and density, average annual income, and health insurance coverage rates.

Psychological Measures included history of mood and anxiety diagnoses, current depression (Patient Health Questionnaire-8; Kroenke et al., 2009), physical and mental health functioning (Veterans RAND 12-item Health survey; Selim et al., 2009), social support (Social Provisions Scale; Cutrona & Russell, 1987), resilience (Connor Davidson Resilience Scale; Connor & Davidson, 2003), alcohol use (Alcohol Use Disorder Identification Test-Consumption; Bush, Kivlahan, McDonell, Fihn, & Bradley, 1998), pain (Fraenkel et al., 2012), and current posttraumatic stress symptoms with the Primary Care Posttraumatic Stress Disorder Screen (Cameron & Gusman, 2003) (described below).

2.2.2. PTSD Follow-up Assessments—The Primary Care Posttraumatic Stress Disorder Screen (PC-PTSD; Cameron & Gusman, 2003) determined PTSD screening status at 3-, 6-, and 12-month follow-up applying DSM-IV criteria (data collection was initiated prior to the publication of DSM-5 and subsequent development of instruments for measuring

DSM-5 symptoms). Nightmares and other intrusive recollections, avoidance, hyperarousal, and numbness or detachment are measured with four yes or no questions that have been psychometrically validated for determining PTSD screening status in Level I trauma centers (Hanley, Brasel, & others, 2013). PTSD+ status was assigned for scores ≥ 3. This cut-off has shown 85% diagnostic efficiency, 78% sensitivity, and 87% specificity compared to the Clinician Administered PTSD Scale (Cameron & Gusman, 2003). Of the 505 enrolled participants, we included 271 participants who were assessed within 1 week of admission and provided 3-month follow-up.

2.3. Data analysis

2.3.1 Data preprocessing—Categorical features were dummy coded (0 = “no” or absent, and 1 = “yes” or present) and features with near zero variance were removed (, resulting in 41 features: 22 hospital and demographic features, 5 zip code features, and 14 psychological features (see Table 1 for descriptive statistics).

2.3.2 Machine Learning—Probability of PTSD at 3-months was predicted using XGBoost (Chen & He, 2015) (version 0.6–4) for the open-source statistical software R (version 3.3.0; R. Core Team, 2000). This ensemble ML algorithm uses gradient-boosted decision trees that rely on the collective performance of individually weak classifiers. Each model is comprised of decision trees with branches representing logical structures terminating at a leaf representing a probability weight. The number of trees in a model and branches in a tree vary as a function of model parameters. A participant’s data determines their “path” through decision trees; for dichotomous features paths diverge for yes (present) or no (absent) responses, and for continuous features, cut-offs provide decision boundaries that steer paths. The final probability estimate is the sum of the individual weights across all trees in the model.

We selected gradient-boosted decision trees for several reasons. Missing data is handled without data imputation or exclusion; instead, specific paths are defined for missing data. This allows prediction even for patients who do not provide all information. Model features are not transformed, which can lead to overfitting (for example, models with features scaled to the maximum and minimum values of the dataset under analysis may not perform well on new datasets with different maxima or minima). Finally, gradient-boosted decision trees capture nonlinear interactions among categorical and continuous features with varied distributions. Although complex interactions may be difficult to interpret as theoretical moderators and mediators, the goal of ML approaches is accurate and reproducible prediction(Yarkoni & Westfall, 2017).

2.3.3. Nested Cross-validation—Nested cross-validation was implemented with the Machine Learning in R package (MLR, version 2.11; Bischl et al., 2016) using 10-times repeated 5-fold cross-validation for both inner and outer cross-validation nests, with the inner cross-validation used to tune model parameters and the outer cross-validation used to estimate predictive performance on new data. This an important distinction from ML studies that use a single (non-nested) cross-validation to both tune the model and to estimate predictive performance. When cross-validation is used to guide any aspect of model fitting,

including the choice of tuning parameters, it no longer provides an unbiased estimate of the true prediction error. This is because tuning parameters are selected based on which model (out of hundreds or thousands of models, depending on how many combinations of tuning parameters are searched) yields the very best cross-validation performance, which circularly selects for the most optimistic estimate of cross-validation error that any model was capable of producing. Nesting the tuning cross-validation within the test cross-validation removes this optimistic bias because it creates an additional set of test cases that are withheld from *both* training and tuning (Cawley & Talbot, 2010; Varma & Simon, 2006). This entire procedure was further repeated 10 times under different randomizations to ensure that model prediction metrics are not dependent on a single chance data partition and to estimate a confidence interval around the mean of these prediction metrics based on the observed variability among the 50 outer cross-validation resamples (10 X 5 folds = 50 models). Tuning parameter statistics are presented in Supplementary Table 1.

2.3.4. Performance Metrics—Overall accuracy is the proportion of predictions that were correct; however, when an outcome is unbalanced, high accuracy can be achieved by predicting the most common outcome for all cases. To account for this, models were trained to maximize AUC, which incorporates sensitivity (the proportion of PTSD+ cases that were correctly predicted to be PTSD+) and specificity (the proportion of PTSD− cases that were correctly predicted to be PTSD−). Several other metrics were calculated. Positive predictive value (PPV) is the proportion of predicted PTSD+ cases that developed PTSD, and negative predictive value (NPV) is the proportion of predicted PTSD− cases that did not develop PTSD. Bootstrapping with 10,000 iterations was used to estimate mean performance and 95% confidence intervals for each metric across models applied to both training and testing subset.

2.3.5. Feature Importance—The importance of individual predictors in the model was ranked according to the mean information gain attributed to each predictor across the 50 resampled models. Information gain is an information theoretic quantity defined as the reduction of entropy that results from splitting a parent node into two child nodes, e.g., making a different prediction of PTSD for those who have nightmares vs. those who do not. This concept is analogous to the reduction of variance from including a predictor in a linear regression; a feature that leads to a larger reduction in variance would be considered to "explain more" than a feature that leads to a smaller reduction in variance. Similarly, features with larger information gains contributed to a greater reduction in the uncertainty of PTSD diagnosis than did those with smaller information gains.

2.3.6. Partial Dependence—Partial dependence analyses calculated model predicted probabilities across values of a feature when all other model features are set to their mean values. These analyses show whether increasing (for continuous) or "yes" (for dichotomous) values of a feature are positively or negatively associated with PTSD+ status.

2.3.7. Cost-benefit Analyses—To illustrate how cost-benefit considerations that reflect tradeoffs between sensitivity and specificity can guide the selection of prediction

threshold, we plotted Receiver Operating Characteristic (ROC) curves and the distribution of predicted probabilities from the final model.

3. Results

3.1 Sample Characteristics

Table 1 provides descriptive statistics of all model features. The three most common causes of injury were falls (29%), automobile collisions (22%) and motorcycle collisions (12%). Most patients had an orthopedic injury (56%), and 26% experienced a traumatic brain injury, which includes concussion. Most of the sample identified as White (68%) and were employed (55%). Among psychological features, 36% reported a history of mood disorders and 25% a history of anxiety disorders. Avoidance was the most frequently endorsed PTSD symptom during hospitalization (44%), followed by nightmares (38%), hypervigilance (34%), and emotional numbing (30%). Most features had less than 5% missing data except the social support measure (12% missing), which was incorporated into the psychological battery shortly after recruitment began, and the income measure (26% missing), which many participants preferred to not respond. At the 3-month follow-up, 110 (32%) screened PTSD+ and 231 (68%) screened PTSD−.

3.2 PTSD Prediction

Table 2 presents performance metrics for the full ML model, and two benchmark comparisons: 1) an ML model that used only features routinely collected at the hospital to assess the value of collecting additional information, and 2) a simple logistic regression that only used the strongest predictor, PTSD severity at the hospital, to assess the value of applying ML to a large feature set. The full model demonstrated significantly better prediction, AUC = 0.85, 95% CI [0.83, 0.86], compared to ML with hospital-based measures only, AUC = 0.75, 95% CI [0.73, 0.76], and logistic regression with PTSD severity at the hospital only, AUC = 0.78, 95% CI [0.76, 0.80].

3.3 Feature Importance

Figure 1 shows the relative importance, assessed by information gain, of features in the final model. Of the nine features with an information gain > 0.05, five were from the psychological battery and four were from routine collection. Most features provided little to no information gain, which reflects their small contributions to the final prediction (Supplementary Table 2).

3.4 Partial Dependence

Figure 2 shows partial dependence graphs for the features with highest relative importance. These graphs show whether features are positively or negatively associated with PTSD after controlling for all other features. However, they should be interpreted with attention to the small changes in predicted PTSD probability across values of even the most important features (visualized when the y-axis range is 30–55%). Of note, none of the individual features were strong enough to capture a shift in the probability of PTSD above the 50% threshold—this underscores the value of our analytic approach, which yielded good prediction by learning complex interactions among individually weak predictors.

3.5 Cost-benefit Analyses

Figure 3A shows the ROC curves for the full model and benchmark comparisons. The selection of threshold can be informed by tradeoffs between sensitivity and specificity. We can consider several hypothetical examples for illustrative purposes. If preventive care were easily accessible the emphasis may be to only exclude individuals who are very unlikely to get PTSD by lowering the threshold to 35%, which yields sensitivity = 0.82 and NPV = 0.90. Conversely, if the cost of preventive care were high, the emphasis may be to only include individuals who are very likely to get PTSD; raising the threshold to 65% yields PPV = 0.75. Another option is to designate an "inconclusive" range of predictions reflecting the decreased accuracy near the 50% threshold (Figure 3B). For example, if probabilities in the 40–70% range were deemed inconclusive, 20% of the sample would not receive a prognosis. However, this would increase the sensitivity (0.76) and specificity (0.90) of predictions for the remaining cases.

4. Discussion

We used ensemble machine learning (ML) to predict PTSD screening status at 3-months using psychological and contextual features in addition to information that is routinely collected during hospitalization. The final models with hospital, demographic, contextual, and psychological features achieved superior prediction to two meaningful benchmarks: an ML model limited to routinely collected data, and a simple logistic regression with PTSD severity at the hospital. Consistent with the expectations of ensemble ML, good prediction was achieved despite the relatively weak predictive value of individual features. Importantly, our predictions were evaluated in a nested cross-validation procedure that separated the testing subset from all aspects of model building, providing an accurate measure of how these models may perform on new data from a similar population.

Our model's performance, measured by AUC, was comparable to two recent ML applications that used ER data to predict non-remitting PTSD (Galatzer-Levy, Karstoft, et al., 2014; Galatzer-Levy et al., 2017). Several of the most important features in our models, including nightmares, age, social support, and avoidance were also key features in prior analyses, despite differences in the ML approach, patient sample, instruments, and the type of PTSD outcome predicted, pointing to some consistency in their predictive value (Galatzer-Levy, Karstoft, et al., 2014). Moreover, several key predictors identified in a preliminary GLM analysis of the current dataset such as age, current depression, and psychiatric history (Powers et al., 2014), remained important in our final model, despite the inclusion of additional participants, and the difference in analytic approach and model evaluation. Given that our approach applied repeated cross-validation to reduce the risk of overfitting, the convergence of key features across analyses provides further evidence of their reliability as predictors.

Several limitations of the current work that can inform future development of predictive algorithms. Multisite data collection would ensure that training models are exposed to a wider diversity of patients and increase the likelihood that predictive accuracy generalizes. Future applications of this approach to multisite projects with the same measures, or combining data from different projects with similar measures that may be harmonized.

Multisite datasets with greater geographic diversity will also provide a more rigorous test of the value of contextual data, which can be extracted from participant zip codes. Some features had little to no impact on the final prediction, even in the context of the ensemble approach, suggesting they either had very limited predictive value or overlapped with other features that yielded better prediction. These can be trimmed from future data collections to make room for other measures without increasing burden on patients and clinicians. We measured PTSD screening status using a self-report measure based on DSM-IV criteria. Subsequent studies can apply the same methodological approaches to predict PTSD status using a clinician-rated instrument based on DSM-5 criteria. This would also allow the clinician to determine whether PTSD status at 3-months post-hospitalization represents a chronic diagnosis, or delayed onset expression. Future studies can also increase the number of assessments for a closer examination of the temporal course of PTSD symptoms. Although our study applied a categorical versus dimensional outcome, similar methods can be applied to predict symptom severity to identify individuals with subthreshold PTSD symptoms that are at risk for experiencing significant functional impairment. Moreover, future studies can apply a similar approach to predicting a variety of psychopathological and functional outcomes that may identify individuals who would benefit from a corresponding early intervention.

In contrast to research aimed at identifying potential treatment *targets*, our approach aims to accurately, cost-effectively, and efficiently predict *risk* for PTSD development. These contrasting aims have different applications. Whereas the former may lead to development of new treatments or improvement of existing ones, our approach may facilitate the implementation of already existing early interventions. Early interventions for PTSD in trauma centers are becoming more efficient and gaining empirical support (Kearns et al. 2012), but these cannot be prescribed across the board. As such, there are several cost-benefit considerations that should guide the application of predictive algorithms for preventive care. These include whether to emphasize overall prevention even if it leads to unnecessary treatment and increased costs, or to reduce false positives even if it means preventive care is withheld from some individuals who go on to develop PTSD. These considerations require multidisciplinary collaboration across key stakeholders.

Importantly, our participants were hospital patients who volunteered to complete additional measures without financial incentive and the collection of measures did not pose a significant training or implementation burden on hospital staff. Our approach complements the aims of evidence-based practice across health care settings, including hospitals where the use of data to drive treatment plans is increasingly emphasized. Altogether, this adds to the evidence provided by prior emergency room studies, in terms of the feasibility and desirability of expanding this line of research. We selected a machine learning algorithm that can accommodate missing data, which is critical for assessing risk in patients who cannot or do not want to complete all measures, or applying the algorithm at sites that do not collect all of the same hospital measures. Moreover, this method can incorporate site-specific features to enhance prediction in locations where the algorithm underperforms.

5. Conclusions

Ensemble ML is a promising approach for improving the prediction of PTSD development after emergency room hospitalization. Progress in identifying more useful predictors that can be routinely collected, and eliminating unnecessary ones, may increase efficiency while improving the accuracy of algorithms that can guide decision-making among patients and providers considering a targeted preventive care intervention after trauma exposure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. Support for this research was provided by The Donald D. Harrington Foundation (Papini) and NIH R21MH110758 (Beevers & Shumake). Funding for the original data collection was provided by the Stanley Seeger Surgical Fund of the Baylor Health Care System Foundation.

Drs. Shumake, Beevers, and Smits receive funding from NIH. Dr. Powers receives funding from NIDA (K01 DA035930). Dr. Smits has received monetary compensation for his work as consultant to Microtransponder, Inc. and Aptinyx, Inc., and royalties from various book publishers.

References

Admon R, Milad MR, & Hendler T (2013). A causal model of post-traumatic stress disorder: disentangling predisposed from acquired neural abnormalities. Trends in Cognitive Sciences, 17(7), 337–347. [PubMed: 23768722]

American Psychiatric Association. (2000). Diagnostic and statistical manual-text revision (DSM-IV-TRim, 2000).

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub. Retrieved from https://books.google.com/books?hl=en&lr=&id=-JivBAAAQBAJ&oi=fnd&pg=PT18&dq=DSM+5&ots=ceNN48IFy7&sig=0TrAsklYxPSG8cYRmQwOcACbyHc

Babyak MA (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic Medicine, 66(3), 411–421. [PubMed: 15184705]

Baker DG, Nievergelt CM, & O'Connor DT (2012). Biomarkers of PTSD: Neuropeptides and immune signaling. Neuropharmacology, 62(2), 663–673. 10.1016/j.neuropharm.2011.02.027 [PubMed: 21392516]

Baker SP, O'Neill B, Haddon W Jr, & Long WB (1974). The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. Journal of Trauma and Acute Care Surgery, 14(3), 187–196.

Beevers CG, Lee H-J, Wells TT, Ellis AJ, & Telch MJ (2011). Association of predeployment gaze bias for emotion stimuli with later symptoms of PTSD and depression in soldiers deployed in Iraq. American Journal of Psychiatry, 168(7), 735–741. [PubMed: 21454916]

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, … Jones ZM (2016). mlr: Machine learning in R. Journal of Machine Learning Research, 17(170), 1–5.

Broekman BFP, Olff M, & Boer F (2007). The genetic background to PTSD. Neuroscience & Biobehavioral Reviews, 31(3), 348–362. 10.1016/j.neubiorev.2006.10.001 [PubMed: 17126903]

Bush K, Kivlahan DR, McDonell MB, Fihn SD, & Bradley KA (1998). The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Archives of Internal Medicine, 158(16), 1789–1795. [PubMed: 9738608]

Cameron RP, & Gusman D (2003). The primary care PTSD screen (PC-PTSD): development and operating characteristics. Primary Care Psychiatry, 9(1), 9–14.

Cawley GC, & Talbot NL (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079–2107.

Chen T, & He T (2015). Xgboost: extreme gradient boosting. R Package Version 0.4–2. Retrieved from http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf

Connor KM, & Davidson JR (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). Depression and Anxiety, 18(2), 76–82. [PubMed: 12964174]

Cutrona CE, & Russell DW (1987). The provisions of social relationships and adaptation to stress. Advances in Personal Relationships, 1(1), 37–67.

DeMasi O, Kording K, & Recht B (2017). Meaningless comparisons lead to false optimism in medical machine learning. ArXiv Preprint ArXiv:1707.06289. Retrieved from https://arxiv.org/abs/1707.06289

Ehring T, Ehlers A, Cleare AJ, & Glucksman E (2008). Do acute psychological and psychobiological responses to trauma predict subsequent symptom severities of PTSD and depression? Psychiatry Research, 161(1), 67–75. [PubMed: 18789538]

Etkin A, & Wager TD (2007). Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. American Journal of Psychiatry, 164(10), 1476–1488. [PubMed: 17898336]

Fraenkel L, Falzer P, Fried T, Kohler M, Peters E, Kerns R, & Leventhal H (2012). Measuring pain impact versus pain severity using a numeric rating scale. Journal of General Internal Medicine, 27(5), 555–560. [PubMed: 22081365]

Galatzer-Levy IR, Brown AD, Henn-Haase C, Metzler TJ, Neylan TC, & Marmar CR (2013). Positive and negative emotion prospectively predict trajectories of resilience and distress among high-exposure police officers. Emotion, 13(3), 545. [PubMed: 23339621]

Galatzer-Levy IR, & Bryant RA (2013). 636,120 ways to have posttraumatic stress disorder. Perspectives on Psychological Science, 8(6), 651–662. [PubMed: 26173229]

Galatzer-Levy IR, Karstoft K-I, Statnikov A, & Shalev AY (2014). Quantitative forecasting of PTSD from early trauma responses: A machine learning application. Journal of Psychiatric Research, 59, 68–76. [PubMed: 25260752]

Galatzer-Levy IR, Ma S, Statnikov A, Yehuda R, & Shalev AY (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. Translational Psychiatry, 7(3), e1070.

Galatzer-Levy IR, Steenkamp MM, Brown AD, Qian M, Inslicht S, Henn-Haase C, … Marmar CR (2014). Cortisol response to an experimental stress paradigm prospectively predicts long-term distress and resilience trajectories in response to active police service. Journal of Psychiatric Research, 56, 36–42. [PubMed: 24952936]

Goldstein RB, Smith SM, Chou SP, Saha TD, Jung J, Zhang H, … Grant BF (2016). The epidemiology of DSM-5 posttraumatic stress disorder in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions-III. Social Psychiatry and Psychiatric Epidemiology, 51(8), 1137–1148. 10.1007/s00127-016-1208-5 [PubMed: 27106853]

Hanley J, Brasel K, & others (2013). Efficiency of a four-item posttraumatic stress disorder screen in trauma patients. Journal of Trauma and Acute Care Surgery, 75(4), 722–727. [PubMed: 24064889]

Iniesta R, Stahl D, & McGuffin P (2016). Machine learning, statistical learning and the future of biological research in psychiatry. Psychological Medicine, 46(12), 2455–2465. [PubMed: 27406289]

Kearns MC, Ressler KJ, Zatzick D, & Rothbaum BO (2012). Early Interventions for Ptsd: A Review. Depression and Anxiety, 29(10), 833–842. 10.1002/da.21997 [PubMed: 22941845]

Kessler RC (2000). Posttraumatic stress disorder: the burden to the individual and to society. The Journal of Clinical Psychiatry. Retrieved from http://psycnet.apa.org/psycinfo/2000-15312-001

Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, & Mokdad AH (2009). The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders, 114(1), 163–173. [PubMed: 18752852]

Liberzon I, & Sripada CS (2007). The functional neuroanatomy of PTSD: a critical review. Progress in Brain Research, 167, 151–169. 10.1016/S0079-6123(07)67011-3

Lokshina Y, & Liberzon I (2017). Enhancing Efficacy of PTSD Treatment: Role of Circuits, Genetics, and Optimal Timing. Clinical Psychology: Science and Practice, 24(3), 298–301. 10.1111/cpsp.12203

Morris MC, Hellman N, Abelson JL, & Rao U (2016). Cortisol, heart rate, and blood pressure as early markers of PTSD risk: A systematic review and meta-analysis. Clinical Psychology Review, 49, 79–91. 10.1016/j.cpr.2016.09.001 [PubMed: 27623149]

Neria Y, DiGrande L, & Adams BG (2011). Posttraumatic stress disorder following the September 11, 2001, terrorist attacks: a review of the literature among highly exposed populations. American Psychologist, 66(6), 429. [PubMed: 21823772]

Neria Y, Nandi A, & Galea S (2008). Post-traumatic stress disorder following disasters: a systematic review. Psychological Medicine, 38(4), 467–480. [PubMed: 17803838]

Pietrzak RH, Goldstein RB, Southwick SM, & Grant BF (2011). Prevalence and Axis I comorbidity of full and partial posttraumatic stress disorder in the United States: Results from Wave 2 of the National Epidemiologic Survey on Alcohol and Related Conditions. Journal of Anxiety Disorders, 25(3), 456–465. 10.1016/j.janxdis.2010.11.010 [PubMed: 21168991]

Pole N, Neylan TC, Otte C, Henn-Hasse C, Metzler TJ, & Marmar CR (2009). Prospective prediction of posttraumatic stress disorder symptoms using fear potentiated auditory startle responses. Biological Psychiatry, 65(3), 235–240. [PubMed: 18722593]

Polusny MA, Erbes CR, Murdoch M, Arbisi PA, Thuras P, & Rath MB (2011). Prospective risk factors for new-onset post-traumatic stress disorder in National Guard soldiers deployed to Iraq. Psychological Medicine, 41(4), 687–698. [PubMed: 21144108]

Powers MB, Warren AM, Rosenfield D, Roden-Foreman K, Bennett M, Reynolds MC, … Smits JAJ (2014). Predictors of PTSD symptoms in adults admitted to a Level I trauma center: A prospective analysis. Journal of Anxiety Disorders, 28(3), 301–309. 10.1016/j.janxdis.2014.01.003 [PubMed: 24632075]

R. Core Team. (2000). R language definition. Vienna, Austria: R Foundation for Statistical Computing Retrieved from http://web.mit.edu/~r/current/arch/amd64_linux26/lib/R/doc/manual/R-lang.pdf

Rothbaum BO, Kearns MC, Price M, Malcoun E, Davis M, Ressler KJ, … Houry D (2012). Early Intervention May Prevent the Development of Posttraumatic Stress Disorder: A Randomized Pilot Civilian Study with Modified Prolonged Exposure. Biological Psychiatry, 72(11), 957–963. 10.1016/j.biopsych.2012.06.002 [PubMed: 22766415]

Rowley G, & Fielding K (1991). Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users. The Lancet, 337(8740), 535–538.

Ruglass LM, Lopez-Castro T, Cheref S, Papini S, & Hien DA (2014). At the crossroads: the intersection of substance use disorders, anxiety disorders, and posttraumatic stress disorder. Current Psychiatry Reports, 16(11), 1–9.

Santiago PN, Ursano RJ, Gray CL, Pynoos RS, Spiegel D, Lewis-Fernandez R, … Fullerton CS (2013). A Systematic Review of PTSD Prevalence and Trajectories in DSM-5 Defined Trauma Exposed Populations: Intentional and Non-Intentional Traumatic Events. PLOS ONE, 8(4), e59236 10.1371/journal.pone.0059236 [PubMed: 23593134]

Selim AJ, Rogers W, Fleishman JA, Qian SX, Fincke BG, Rothendler JA, & Kazis LE (2009). Updated US population standard for the Veterans RAND 12-item Health Survey (VR-12). Quality of Life Research, 18(1), 43–52. [PubMed: 19051059]

Smoller JW (2016). The Genetics of Stress-Related Disorders: PTSD, Depression, and Anxiety Disorders., The Genetics of Stress-Related Disorders: PTSD, Depression, and Anxiety Disorders. Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology, Neuropsychopharmacology, 41, 41(1, 1), 297, 297–319. https://doi.org/10.1038/npp.2015.266 , https://doi.org/10.1038/npp.2015.26610.1038/npp.2015.266, 10.1038/npp.2015.266 [PubMed: 26321314]

Telch MJ, Beevers CG, Rosenfield D, Lee H-J, Reijntjes A, Ferrell RE, & Hariri AR (2015). 5-HTTLPR genotype potentiates the effects of war zone stressors on the emergence of PTSD,

depressive and anxiety symptoms in soldiers deployed to iraq. World Psychiatry, 14(2), 198–206. [PubMed: 26043338]

Telch MJ, Rosenfield D, Lee H-J, & Pai A (2012). Emotional reactivity to a single inhalation of 35% carbon dioxide and its association with later symptoms of posttraumatic stress disorder and anxiety in soldiers deployed to Iraq. Archives of General Psychiatry, 69(11), 1161–1168. [PubMed: 23117637]

van der Velden PG, & Wittmann L (2008). The independent predictive value of peritraumatic dissociation for PTSD symptomatology after type I trauma: A systematic review of prospective studies. Clinical Psychology Review, 28(6), 1009–1020. [PubMed: 18406027]

Varma S, & Simon R (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1), 91. [PubMed: 16504092]

Walsh K, Nugent NR, Kotte A, Amstadter AB, Wang S, Guille C, … Resnick HS (2013). Cortisol at the emergency room rape visit as a predictor of PTSD and depression symptoms over time. Psychoneuroendocrinology, 38(11), 2520–2528. [PubMed: 23806832]

Watts BV, Schnurr PP, Mayo L, Young-Xu Y, Weeks WB, & Friedman MJ (2013). Meta-analysis of the efficacy of treatments for posttraumatic stress disorder. The Journal of Clinical Psychiatry, 74(6), e541–50. [PubMed: 23842024]

Yarkoni T, & Westfall J (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12(6), 1100–1122. 10.1177/1745691617693393 [PubMed: 28841086]

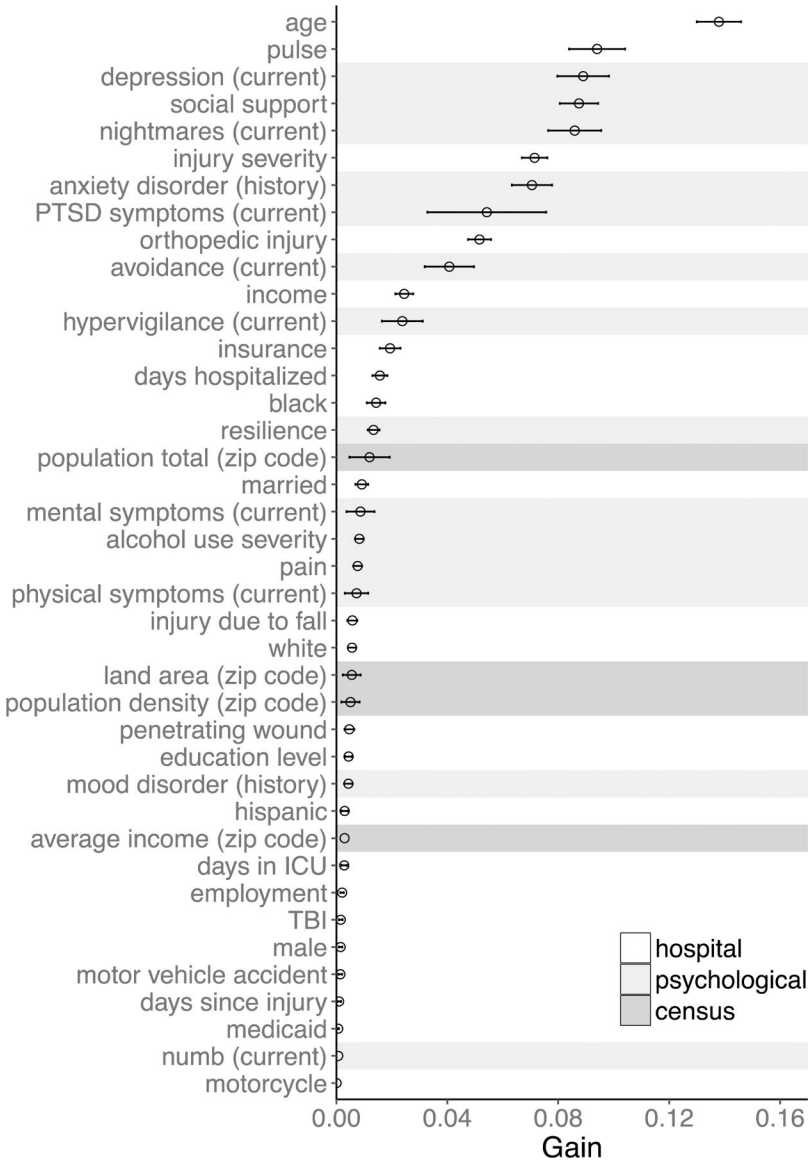

Figure 1.

Average gain across cross-validated tests of the full model indicates the proportion that a feature contributed toward the ensemble's prediction of PTSD screening status. Shading indicates the source of each feature, and the error bars show 95% CI. Hospital features were extracted from the hospital intake. Census features were extracted from publicly available census data based on participant zip codes. Psychological features were extracted from the following: patients self-reported history of mood and anxiety diagnoses, current depression was assessed with the Patient Health Questionnaire-8, physical and mental health functioning was assessed with the Veterans RAND 12-item Health survey, social support was assessed with the Social Provisions Scale, resilience was assessed with the Connor Davidson Resilience Scale, and alcohol use was assessed with the Alcohol Use Disorder Identification Test-Consumption.

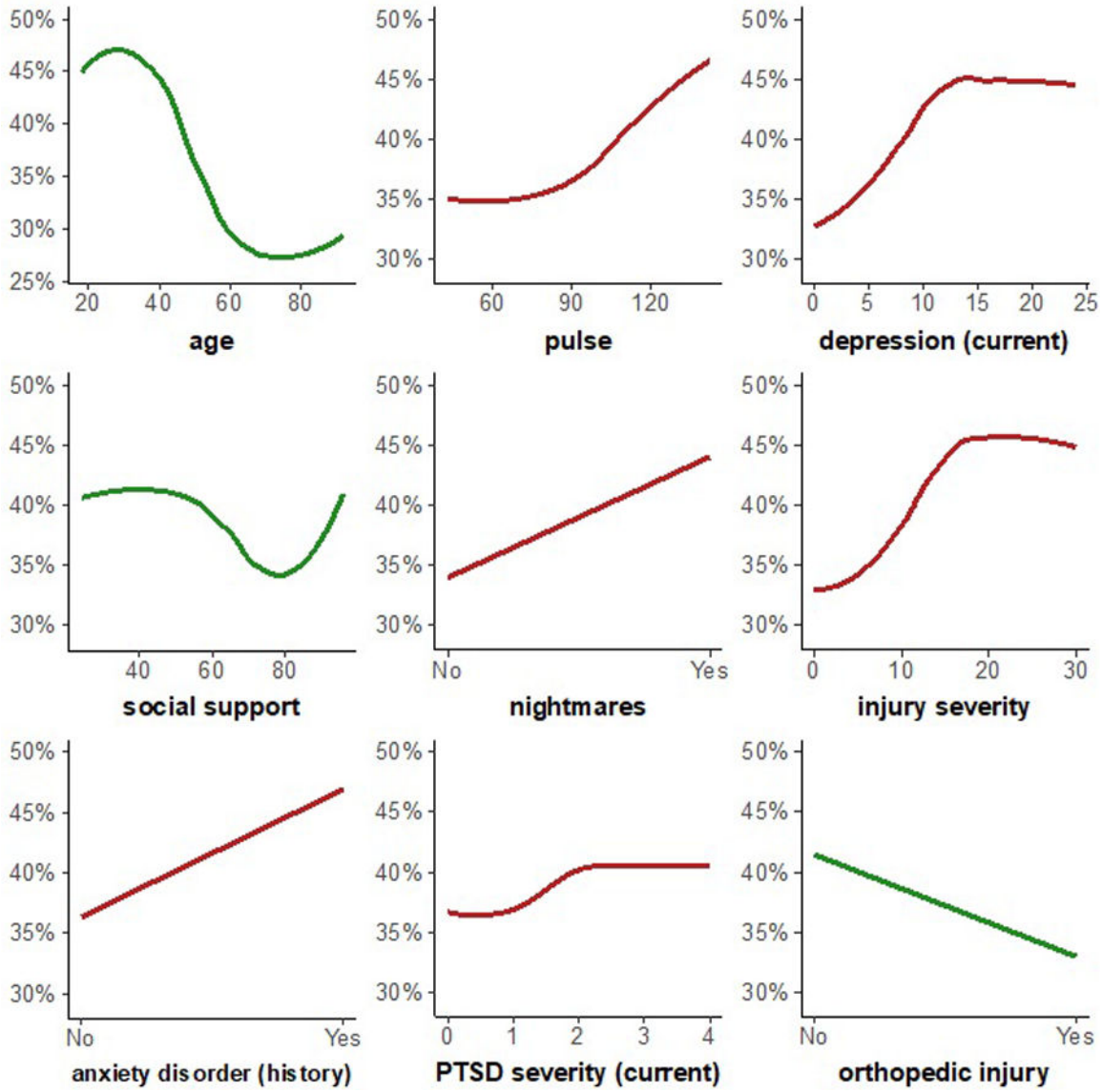

Figure 2.

Nine features with gain > 0.05. Locally weighted smoothing (loess) curves show how the probability of PTSD+ prediction (y-axis) varies as a function of a feature (x-axis) after controlling for all other features. Note the y-axis range, which was selected to illustrate the narrow range of influence that single features have on prediction.

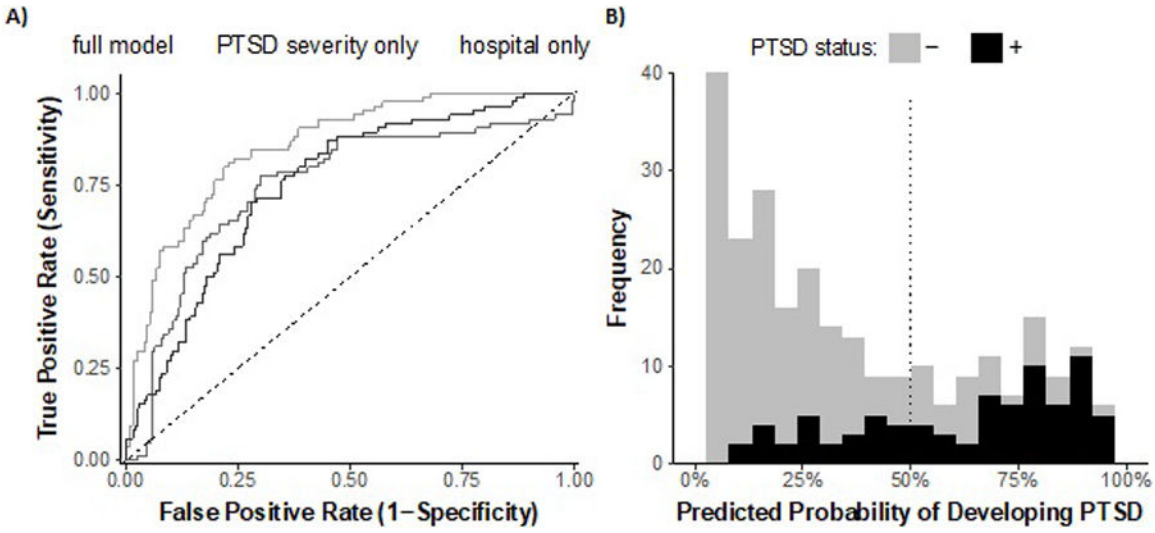

Figure 3.

A) Sensitivity as a function of 1-specificity for the full model (all features) and comparison models including logistic regression with PTSD severity at the hospital as the only predictor, and machine learning with only features routinely collected at the hospital. AUC statistics refer to area under these curves, which is significantly higher for the full model. The non-overlapping shaded areas represent improvement in prediction. The dashed line corresponds to a no-information model, which yields AUC = 0.50. B) Distribution of predicted probabilities of developing PTSD from the full model. The x-axis refers to mean test-sample prediction for each patient, and frequency bars are shaded by actual PTSD screening status at 3 months. The gray bars to the left of the 50% threshold represent the number of accurate PTSD− predictions, and the black bars to the right represent accurate PTSD+ predictions. The models yield a wide arrange of personalized predictions that are highly accurate at the extreme ends but less accurate near the 50% threshold, which should be considered for guiding treatment decisions.

Table 1.

Descriptive statistics of model features.

Feature (% missing)	Mean (SD) or *n* (%)		
	Total ***N*** **= 271**	**PTSD + at 3 mo.** ***n*** **= 110**	**PTSD − at 3 mo.** ***n*** **= 231**
Hospital features			
Pulse at admission (3%)	89.15 (17.99)	95.82 (20.59)	86.14 (15.85)
Days in hospital (4%)	5.31 (3.94)	5.5 (4)	5.22 (3.92)
Days in intensive care unit (4%)	0.58 (1.42)	0.63 (1.11)	0.55 (1.55)
Medicaid (0%)	81 (30%)	22 (26%)	59 (32%)
Private insurance (0%)	123 (45%)	21 (25%)	102 (55%)
Injury type (4%)			
Fall	76 (29%)	14 (17%)	62 (35%)
Motor vehicle collision	58 (22%)	23 (28%)	35 (20%)
Motorcycle collision	30 (12%)	6 (7%)	24 (13%)
Traumatic brain injury (2%)	68 (26%)	27 (33%)	41 (23%)
Penetrating wound (4%)	31 (12%)	18 (22%)	13 (7%)
Orthopedic injury (2%)	149 (56%)	32 (38%)	117 (64%)
Injury Severity Scale total (4%)	9.85 (6.13)	11.09 (6.82)	9.28 (5.72)
Assessment (days after injury)	3.39 (1.64)	3.37 (1.59)	3.4 (1.66)
Demographic features			
Age (0%)	46.73 (17.35)	38.69 (13.66)	50.34 (17.65)
Male (0%)	173 (64%)	56 (67%)	117 (63%)
Race (1%)			
White	183 (68%)	47 (57%)	136 (73%)
Black	55 (20%)	25 (30%)	30 (16%)
Hispanic (0.4%)	64 (24%)	30 (37%)	34 (18%)
Married (0.4%)	94 (35%)	16 (19%)	78 (42%)
Employed (0%)	159 (59%)	47 (56%)	112 (60%)
Education level (0%)			
Middle school	6 (2%)	3 (4%)	3 (2%)
Some high school	33 (12%)	15 (18%)	18 (10%)
High school graduate	104 (38%)	34 (40%)	70 (37%)
Some college or 2-year degree	54 (20%)	19 (23%)	35 (19%)
Four-year college degree	45 (17%)	9 (11%)	36 (19%)
Graduate or professional degree	29 (11%)	4 (5%)	25 (13%)
Income level (26%)			
< $25k	71 (26%)	33 (39%)	38 (20%)
$25–49k	38 (14%)	11 (13%)	27 (14%)
$50–74k	37 (14%)	7 (8%)	30 (16%)
≥$75k	55 (20%)	8 (10%)	47 (25%)
Psychological features			
History of mood disorder (4%)	94 (36%)	41 (51%)	53 (29%)

Feature (% missing)	Mean (SD) or *n* (%)		
	Total ***N*** **= 271**	**PTSD + at 3 mo.** ***n*** **= 110**	**PTSD − at 3 mo.** ***n*** **= 231**
History of anxiety disorder (7%)	63 (25%)	36 (49%)	27 (15%)
Current depression severity (0%)	6.7 (5.59)	9.93 (5.99)	5.26 (4.75)
Physical health (0.4%)	44.84 (11.24)	44.87 (11.29)	44.83 (11.25)
Mental health (0.4%)	51.07 (11.55)	45.58 (13.62)	53.54 (9.54)
Social support (12%)	78.85 (11.38)	76.54 (12.35)	79.79 (10.85)
Resilience (2%)	32.01 (6.38)	30.23 (7.49)	32.78 (5.68)
Alcohol use severity (0.4%)	2.79 (2.68)	3.1 (2.7)	2.65 (2.67)
Pain (0%)	6.91 (2.38)	7.38 (2.24)	6.7 (2.42)
PTSD Severity (0%)	1.42 (1.44)	2.44 (1.32)	0.96 (1.24)
Nightmares (0%)	102 (38%)	56 (67%)	46 (25%)
Avoidance (0%)	118 (44%)	60 (71%)	58 (31%)
Hypervigilance (0%)	93 (34%)	50 (60%)	43 (23%)
Dissociation (0%)	72 (27%)	39 (46%)	33 (18%)
Census features from zip codes			
Land area (1%)	36.82 (52.15)	30.45 (49.71)	39.69 (53.09)
Population total (1%)	35.00k (21.17k)	39.73k (24.55k)	32.87k (19.15k)
Population density (1%)	3.03k (2.49k)	3.44k (2.36k)	28.41k (25.34k)
Average annual income (2%)	18.73k (13.46k)	16.72k (12.92k)	19.65 (13.64k)
Health insurance coverage (1%)	79.59% (9.23%)	75.72% (9.06%)	79.89% (9.03%)

Note. For features with missing data, means and percentages are calculated from the subset of available data. Education and income level were coded as continuous features.

Table 2.

Mean performance metrics with bootstrapped 95% confidence intervals

Performance Metric	Full model (*N* features = 41)	PTSD severity at hospital only	Hospital features (*N* features = 22)
Area under curve	0.85 [0.83, 0.86]	0.78 [0.76, 0.80]	0.75 [0.73, 0.76]
Sensitivity	0.69 [0.66, 0.72]	0.51 [0.48, 0.55]	0.57 [0.53, 0.61]
Specificity	0.83 [0.80, 0.85]	0.87 [0.86, 0.88]	0.76 [0.73, 0.79]
Positive predictive value	0.65 [0.62, 0.69]	0.63 [0.61, 0.66]	0.53 [0.5, 0.56]
Negative predictive value	0.86 [0.84, 0.87]	0.80 [0.79, 0.81]	0.80 [0.79, 0.81]
Overall accuracy	0.78 [0.77, 0.80]	0.76 [0.74, 0.77]	0.70 [0.68, 0.72]

Note. These metrics reflect model performance at a 50% threshold (i.e., PTSD+ status was assigned when predicted probability was ≥ .5). The full model including hospital, psychological, and contextual features outperformed two benchmark comparisons: machine learning with hospital features only, and logistic regression with the strongest predictor only, PTSD severity at the hospital.