The impact of psychopathology, social adversity and stress-relevant DNA methylation on prospective risk for post-traumatic stress: A machine learning approach

Agaz H. Wani[a], Allison E. Aiello[b], Grace S. Kim[c], Fei Xue[d], Chantel L. Martin[b], Andrew Ratanatharathorn[e], Annie Qu[f], Karestan Koenen[g,h], Sandro Galea[i], Derek E. Wildman[a], Monica Uddin[a,*]

[a] *Genomics Program, College of Public Health, University of South Florida, Tampa, FL, United States*
[b] *Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill and Carolina Population Center, University of North Carolina at Chapel Hill, United States*
[c] *Medical Scholars Program, University of Illinois College of Medicine, United States*
[d] *Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, United States*
[e] *Department of Epidemiology, Columbia University, United States*
[f] *Department of Statistics, University of California Irvine, United States*
[g] *Department of Epidemiology, Harvard T.H. Chan School of Public Health, United States*
[h] *Psychiatric and Neurodevelopmental Genetics Unit & Department of Psychiatry, Massachusetts General Hospital, United States*
[i] *Boston University School of Public Health, United States*

ARTICLE INFO

Keywords:
DNA methylation
Machine learning
PTSD
Psychopathology

ABSTRACT

Background: A range of factors have been identified that contribute to greater incidence, severity, and prolonged course of post-traumatic stress disorder (PTSD), including: comorbid and/or prior psychopathology; social adversity such as low socioeconomic position, perceived discrimination, and isolation; and biological factors such as genomic variation at glucocorticoid receptor regulatory network (GRRN) genes. This complex etiology and clinical course make identification of people at higher risk of PTSD challenging. Here we leverage machine learning (ML) approaches to identify a core set of factors that may together predispose persons to PTSD.
Methods: We used multiple ML approaches to assess the relationship among DNA methylation (DNAm) at GRRN genes, prior psychopathology, social adversity, and prospective risk for PTS severity (PTSS).
Results: ML models predicted prospective risk of PTSS with high accuracy. The Gradient Boost approach was the top-performing model with mean absolute error of 0.135, mean square error of 0.047, root mean square error of 0.217, and R^2 of 95.29%. Prior PTSS ranked highest in predicting the prospective risk of PTSS, accounting for >88% of the prediction. The top ranked GRRN CpG site was cg05616442, in *AKT1*, and the top ranked social adversity feature was loneliness.
Conclusion: Multiple factors including prior PTSS, social adversity, and DNAm play a role in predicting prospective risk of PTSS. ML models identified factors accounting for increased PTSS risk with high accuracy, which may help to target risk factors that reduce the likelihood or course of PTSD, potentially pointing to approaches that can lead to early intervention.
Limitation: One of the limitations of this study is small sample size.

1. Introduction

Post-traumatic stress disorder (PTSD) is a common and severe psychiatric disorder that develops following exposure to life-threatening or terrifying events (Bisson et al., 2015; Shalev, 2001). It can develop following exposure to a single horrifying event or a prolonged exposure to a series of traumatic events such as physical or sexual assault, or combat. Not all people develop PTSD following exposure to trauma.

[*] Corresponding author.
E-mail addresses: monica43@usf.edu, monica43@health.usf.edu (M. Uddin).

https://doi.org/10.1016/j.jad.2020.12.076
Received 24 September 2020; Received in revised form 18 December 2020; Accepted 21 December 2020

Many people show the ability to recover from trauma exposure (Bonanno, 2004). Nevertheless, in a subset of individuals, PTSD can be severe and debilitating and show a chronic course over time (Perkonigg et al., 2005; Zvolensky et al., 2015). In addition, individuals with PTSD show different symptom presentations (Galatzer-Levy and Bryant, 2013) and often exhibit comorbid psychopathology, including depression and anxiety (Ginzburg et al., 2010). Comorbid psychopathology is likely to affect recovery and treatment outcome (Bradley et al., 2005; Dalenberg et al., 2012). Importantly, despite many decades of research, it remains challenging to predict individuals at high risk of prospective post-traumatic psychopathology.

It is well-established that social adversity, such as low socioeconomic position (SEP), isolation, and discrimination, have a significant impact on mental and physical health. Multiple studies have shown a strong association between SEP and increased risk of psychopathology (Koenen et al., 2007; Uddin et al., 2011; Ward-Caviness et al., 2020). People with low SEP are at a higher risk of mental health problems (Kiely et al., 2015; Ramos-Lima et al., 2019) and low SEP can adversely impact access to treatment and prevention of mental health conditions. In addition to SEP, research shows that parental education, and gender are associated with a prospective or higher risk of psychopathology (Park et al., 2013). Similarly, social relationships or isolation/loneliness profoundly affect mental and physical health including risk for PTSD (Cacioppo et al., 2015; Hyland et al., 2019; Kuwert et al., 2014; Link and Phelan, 1995). Also, the contribution of perceived discrimination, which can engender feelings of isolation, has been significantly associated with many health conditions, including PTSD (Brooks Holliday et al., 2018; Kessler et al., 1999), with people experiencing perceived discrimination more likely to show symptoms of the disorder (Bogart et al., 2011).

PTSD is thus shaped by a complex mix of stressful and traumatic life events that shape how our body responds to stress. Our central stress response system, the hypothalamic pituitary adrenal (HPA) axis plays a key role here. The HPA axis involves a complex set of instructions and feedback interactions between the hypothalamus, pituitary, and adrenal gland. Its main job is distributing glucocorticoid hormones, such as cortisol released by the adrenal gland (Whitnall, 1993). These hormones are vital for life and play a key role in mediating the stress response of the HPA axis. Numerous studies have shown that dysregulation in HPA axis function involving glucocorticoids, e.g., cortisol, plays a crucial role in PTSD pathophysiology (Yehuda et al., 2004; Yehuda et al., 1993). Dysregulation in the HPA axis has also been associated with major depressive disorder (Anacker et al., 2011; Yehuda et al., 2004). Genomic variation at glucocorticoid receptor regulatory network (GRRN) genes has been associated with PTSD (Binder et al., 2008; Labonté et al., 2014), childhood maltreatment (Bustamante et al., 2016), and depression (Bustamante et al., 2016; Hodes et al., 2016). These studies demonstrate that stress-relevant genomic and epigenomic variation, including DNAm variation, may play a role in predicting stress-related psychopathology.

As previously noted, predicting prospective risk of psychopathology remains a challenging task. Recent work, however, has seen a burgeoning interest in the application of machine learning (ML) methods for predicting PTSD risk (Karstoft et al., 2015; Wshah et al., 2019). ML refers to those approaches that do not use explicit programming but instead learn from data and experience to make decisions or predictions. For example, ML has recently been used to differentiate between combat-related PTSD and trauma-exposed controls (Zhang et al., 2020), and for diagnosing PTSD (Dean et al., 2019). With respect to PTSD, identifying factors that prospectively predict PTSD can potentially help target such factors to reduce the likelihood of PTSD following a traumatic event, and/or reduce the likelihood of a more chronic course of PTSD. To date, however, existing work has not considered social adversity-related factors, including loneliness and perceived discrimination, in developing ML-based models predicting PTSD, or combined such data with DNAm measures of relevance to stress-related psychopathology. To address these gaps, here we build ML models that incorporate factors previously associated with elevated risk of PTSD, including prior psychopathology, social adversity exposure, and DNAm variation in GRRN genes to predict prospective risk of PTSD symptom severity (PTSS).

2. Materials and methods

In this study, we used data from the Detroit Neighborhood Health Study (DNHS), a prospective population-based longitudinal cohort of individuals living in Detroit, Michigan (Goldmann et al., 2011; Uddin et al., 2010). All participants in this study were 18 years or older and predominantly self-identified as African American (AA). The main aim of the DNHS was to identify how genetic variation, stressful and traumatic life experiences, and features of the environment predict psychopathology and behavior. Participants were recruited for a structured telephone interview each year between 2008 and 2013 to assess perceptions of participant's neighborhoods, mental and physical health status, social support, exposure to traumatic events, post-traumatic stress disorder symptoms, depression symptoms, generalized anxiety symptoms and alcohol, and tobacco use. Informed consent was obtained at the beginning of each interview and again at specimen collection. The Institutional Review Board of the University of Michigan and the University of North Carolina-Chapel Hill reviewed and approved this study.

2.1. Measures

Demographic information, such as age, sex, race, education, marital status, and employment was self-reported by the participants. PTSD was assessed according to DSM-IV criteria using the PTSD Checklist Civilian Version (PCL-C) as described in (Uddin et al., 2010). Exposure to lifetime traumatic events was assessed using a survey of 19 item traumatic events, as in previous work (Breslau et al. 1998). Cumulative traumatic burden was estimated by summing the scores of lifetime traumatic event types (Uddin et al., 2010). Similarly, major depressive disorder (MDD) and generalized anxiety disorder (GAD) were measured using Patient Health Questionnaire (PHQ-9) and GAD-7 scales, respectively, as described earlier (Uddin et al., 2010). In addition to the categorical diagnosis of PTSD, we used continuous measures of PTSD, based on summing scores for 17 symptoms from the worst lifetime trauma. Also, severity measures for PTSD symptom clusters (intrusion, avoidance and hyperarousal) were used, by summing symptoms related to each symptom cluster. Depression and anxiety symptom severity measures based on respective scales (PHQ-9 and GAD-7) were used as well. Perceived discrimination was measured using the Everyday Discrimination Scale (EDS), a nine-item self-report scale (Williams et al., 1997). Loneliness was measured using a three-item scale (Hughes et al., 2004). Emotional mistreatment, financial problems, legal issues, drug and alcohol-related problems, job loss, unemployment, divorce were measured as standalone stressors.

2.2. DNAm collection, quality control, and pre-processing

Biospecimens were collected from DNHS participants who consented to give a sample. In this study, DNAm data were collected from DNHS participants exposed to one or more traumas, via DNA isolated from venipuncture blood draws as described in (Uddin et al., 2010).

DNAm was measured using Illumina's Infinium MethylationEPIC BeadChip in 500 samples from 190 unique participants following the manufacturer's recommended protocol. Metadata from trauma-exposed participants was randomized to minimize plate and chip mediated batch effects (Harper et al., 2013). Resulting DNAm data was then subjected to quality control (QC): Sex and genotype checks were performed to remove sex discordant samples using the *minfi* and *ewastools* R packages (Aryee et al., 2014; Heiss and Just, 2018). Raw DNAm β values were obtained using the *minfi* R package (Aryee et al., 2014). QC was performed to filter poorly performing samples and probes. Samples with low signal intensity (i.e. mean signal intensity <2000 arbitrary units), or <50% of the overall median were also removed (Barfield et al., 2012) as were samples and probes with >10% missing values. Probes with detection p-value > 0.01 were set to missing. Cross-reactive and polymorphic probes were removed (McCartney et al., 2016).

QC removed a total of 52 samples, leaving 448 samples from 179 participants (from four time points/waves) for subsequent analysis. For ML, we used samples from waves 1 and 2 to predict PTSS from wave 4, and included survey data from waves 1, 2, and 3 for model building (described below). From the set of 448 samples that passed QC, a total of 210 samples from 148 unique participants met the criteria for inclusion in ML analyses (Fig. 1). Normalization of the methylation data was performed using the *Noob* approach implemented in the *minfi* R package (Aryee et al., 2014; Triche et al., 2013). *ComBat* adjustment was performed to reduce the likelihood of bias due to known batch effects using an empirical Bayesian framework implemented in *SVA* R package (Johnson et al., 2006; Leek et al., 2012). Cell estimations were computed using the *IDOL* algorithm (Salas et al., 2018).

2.3. Pre-processing for ML

To prepare the data (DNAm, phenotype and cell proportions) for ML, we performed some additional pre-processing steps, including imputation of missing data, feature selection, and scaling. The overall workflow of the pre-processing and ML models is shown in Fig. 2.

General ML Approach:

In this study, we used multiple well-known and efficient ML algorithms for prediction. As described in more detail below, the goal is to compare and determine the best performing algorithms predicting PTSD with high accuracy.

Imputation:

As ML models require no missing data, we imputed both phenotype and DNAm data. We used Predictive Mean Matching (PMM), a semi-parametric approach to impute the phenotype information using the *mice* R package (van Buuren and Groothuis-Oudshoorn, 2011). DNAm data were imputed using the *k*-nearest neighbor approach ($k = 2$) implemented in the *Scikit-learn* Python framework (Pedregosa et al., 2011; Troyanskaya et al., 2001). All the data (DNAm, cell estimation, and phenotypes) were combined for ML.

Feature selection:

Feature selection is a critical pre-processing step used to remove redundant and irrelevant features; CpGs, phenotypes (including social adversity exposures) and estimated cell type proportions are all defined as features. This step helps to identify the significant features that are highly predictive for the outcome variable of interest. We used a univariate feature selection approach based on a univariate statistical test to select the top 150 features important for PTSS prediction, implemented in the *Scikit-learn* python framework (Pedregosa et al., 2011). This univariate approach looks at the strength of association between each feature and the response variable. A score (F-score) based on the F-test is given to each feature, and then *K* highest scoring features are selected.

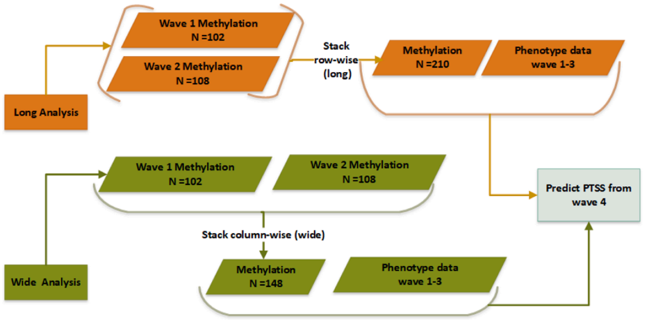

Fig. 1. Two sets of analyses showing methylation data in a long and wide format, including the phenotype information. Long Analyses: combines methylation data from two-time points (waves) in a row-wise format ($n = 210$). Wide Analyses: combines methylation data from two-time points in a column-wise format ($n = 148$).

Feature Scaling:

To standardize data, we performed feature scaling to standardize the data with mean = 0 and standard deviation = 1 using *Scikit-learn* framework (Pedregosa et al., 2011).

2.4. Machine learning approaches

We used multiple approaches such as Random Forest (RF) (Breiman, 2001), Adaboost (AB) (Drucker, 1997; Freund and Schapire, 1997), Gradient Boost (GB) (Friedman, 2001), Linear Regression (LR) (Lai et al., 1978), Support Vector Regression (SVR) (Chang and Lin, 2011; Vapnik, 1995)*,* Bagging Regression (BR) (Breiman, 1996) and Voting Regression (VR) (An and Meng, 2010). More information about these approaches is given in Supplementary Information.

2.5. Evaluation

Cross-validation:

Cross-validation is a technique used to estimate how well a model will generalize to a test data set. We performed *k*-fold ($k = 10$) cross-validation on the training data. It works by training the model on *k-1*-fold of the training data and validating on the *kth* fold. Each of the *k*-folds follows this approach, and average performance on *k*-folds is measured. Cross-validation is computationally intensive but it makes more data available for training than when apportioning a separate validation dataset which is not feasible when the sample size is small.

To evaluate the error rate and accuracy of the models, we used standard error reporting metrics: mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R Squared (R^2) (Supplementary Information). As we scaled the data, including the response variable, the error rates (MAE, MSE, and RMSE) are also scaled.

2.6. Analysis

In the current sample from the DNHS, we have data from a baseline wave (wave 1) and three follow-up waves (wave 2, wave 3, and wave 4). We performed two types of analyses using ML. For both types of analyses, we used DNAm data from waves 1 and 2, and phenotype data from waves 1, 2, and 3 to predict PTSS from wave 4 as shown in Fig. 1. In the first analysis, DNAm data from waves 1 and 2 were stacked together, such that participants with DNAm data from more than one wave were included as rows in the data matrix. This analysis included 210 samples from 148 unique participants. In the second analysis, we arranged participants with DNAm data from more than one wave as columns in the data matrix, to identify the CpGs and phenotypes that are significant in both waves. This analysis included 148 samples from 148 unique participants. In both analyses, we examined, using DNAm and phenotype data, how well we can predict the prospective risk of PTSS, and identify the CpG sites and phenotypes that are highly predictive for PTSS. From now onwards in this study, we will refer first analysis (DNAm row-wise) as “long” and the second as “wide” (DNAm column-wise) for the sake of simplicity.

As mentioned, feature selection is an important step in ML. To see if ML performs better on the data with a significant set of features as compared to the data with a full set of features, we used data in four different ways 1) All the DNAm features as long 2) All the DNAm features as wide 3) Important DNAm features as long and 4) Important DNAm features as wide. To implement the machine learning models, we first divided the data into training and testing sets (3:1). We then trained the models and performed hyperparameter tuning and cross-validation on the training dataset (75%), followed by testing on the training set. Final evaluation was done on the test dataset (25%).

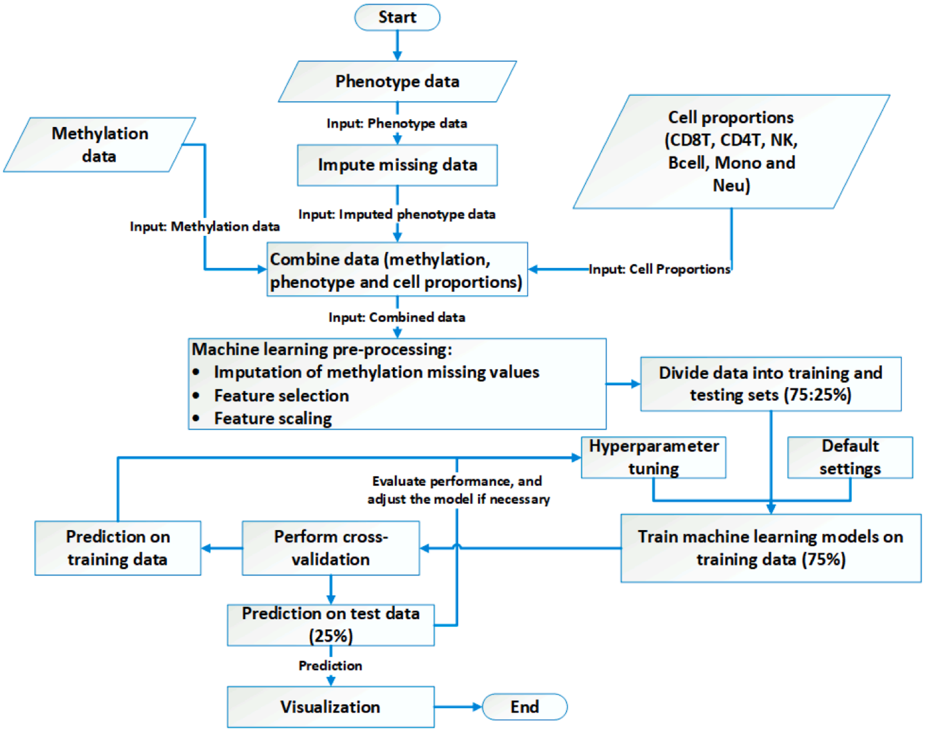
Start
Phenotype data
Input: Phenotype data
Methylation data
Impute missing data
Cell proportions (CD8T, CD4T, NK, Bcell, Mono and Neu)
Input: Imputed phenotype data
Input: Methylation data
Combine data (methylation, phenotype and cell proportions)
Input: Cell Proportions
Input: Combined data
Machine learning pre-processing:
• Imputation of methylation missing values
• Feature selection
• Feature scaling
Divide data into training and testing sets (75:25%)
Hyperparameter tuning
Default settings
Evaluate performance, and adjust the model if necessary
Prediction on training data
Perform cross-validation
Train machine learning models on training data (75%)
Prediction on test data (25%)
Prediction
Visualization
End

Fig. 2. Workflow of the ML process. It begins with combining data (DNAm, phenotype and cell proportions) and ends with the PTSS risk prediction and visualization of the results.

3. Results

We included a total of 210 samples from 148 unique DNHS participants for "long" ML analyses, and 148 samples from 148 unique participants for "wide" ML analyses. All participants were age 18 years or above and exposed to traumatic events. The participants predominantly self-identified as AAs (93.81%) and female (60%). The demographic characteristics of the cohort are shown in Table 1.

3.1. Best features and importance

First, to remove the irrelevant and redundant features, we performed feature selection and selected the top 150 features from both long and wide datasets, respectively. We performed a series of experiments to identify the set of features that most precisely predict PTSS. We tested sets of features that included between 10 - 590 features, increasing the size of feature set by increments of 20, until the error rate increased as we added more features. The models, RF, AB, and GB were used to find the errors (MAE and RMSE) of these different sized feature sets, and we evaluated the performance of these sets based on the mean of all three models. The feature set with the minimum mean error rate was used in the final analysis. Error rates of different feature sets are given in Table S1 in the supplementary information.

From the significant set of 150 features in the long data, 79 features were CpGs, and 71 were phenotype variables (Supplementary Figures S1 and S2). Similarly, from the wide data, 99 significant features were identified to be CpGs, while 51 were phenotype variables (Supplementary Figures S3 and S4). Between the long and wide data, 44 CpGs and 51 phenotypes were shared in common. Interestingly, three CpGs (cg04444450: *NCOA2*, cg20509117: *IL6; LOC541472* and cg05790989: *POU2F1*), shown in Table 2, were consistently significant, meaning that they were in both waves of the wide dataset (wave 1 and wave 2) and in the long dataset. The important CpGs and phenotypes identified in both the long and wide datasets are shown in Fig. 3. Prior PTSS ranked highest in predicting the prospective risk of PTSS, accounting for 88–89% of the prediction in the wide and long datasets, respectively. Psychopathology of depression and anxiety were also found to be significant in predicting PTSS risk. In addition to prior psychopathology, PTS symptom clusters (hyperarousal, intrusion, and avoidance) and social adversity factors (loneliness, perceived discrimination, financial problems, and emotional mistreatment) were significant predictors of PTSS risk. Using scores from both long and wide datasets, the top ranked GRRN CpG site was cg05616442 in the gene *AKT1* (the gene that encodes Protein kinase B, PKB) (see Table 2), and the top ranked social adversity feature was loneliness, followed by perceived discrimination and financial problems. The cumulative score of traumatic event types was also on the list of significant predictors for both datasets. In general, the importance of phenotypes ranked higher than the importance of most CpGs, although there were exceptions to this pattern (Fig. 3; Figures S1-S4). Correlation of DNAm data in CpGs common to both approaches is shown in Fig. 4. Most CpGs were positively correlated with each other; however, CpG cg26560981 in gene Mitogen-Activated Protein Kinase 10 (*MAPK10*) was negatively correlated with the greatest number of CpGs.

Table 1
Demographic characteristics, traumatic events, PTSS score and number of unique participants in the study sample.

	Mean (SD) or N (%)
Sex, Female	126 (60%)
Age[a]	54.57(12.79)
Race, AA	197 (93.81%)
Cumulative traumatic event types[b]	6.76(4.55)
PTSS[c]	43.91(16.26)
Unique participants	148
Participants with two time-points of DNAm data	62

[a] Age is taken from the baseline wave.

[b] Cumulative traumatic event types are from corresponding waves (e.g., if a sample is from wave 2, its cumulative trauma will be the sum of scores from wave 1 and 2).

[c] PTSS is from wave 4 (response variable).

Table 2

Glucocorticoid receptor regulatory network CpGs and associated genes identified as important features in both the long and wide datasets[*].

Illumina ID	UCSC_Ref Gene Name	Wide Importance	Long Importance
cg05616442	*AKT1*	0.00418	0.00212
cg18071894	*AKT1*	0.00281	0.00052
cg26495008	*FKBP5*	0.00213	0.00037
cg10913456	*FKBP5*	0.00196	0.00029
cg23462257	*NFKB1*	0.00190	0.00199
cg22237988	*BAX*	0.00175	0.00028
cg26049684	*AKT1*	0.00153	0.00085
cg25368824	*IL4*	0.00147	0.00356
cg02842899	*TP53;WRAP53*	0.00144	0.00069
cg02564102	*SMARCC1*	0.00137	0.00081
cg11916669	*AKT1*	0.00127	0.00036
cg13135255	*CREBBP*	0.00127	0.00054
cg04444450[+]	*NCOA2*	0.00110	0.00094
cg24307117	*SPI1*	0.00097	0.00068
cg23523922	*AFP*	0.00095	0.00028
cg15912732	*AKT1*	0.00091	0.00074
cg11540119	*SMARCC1*	0.00086	0.00040
cg20509117[+]	*IL6; LOC541472*	0.00084	0.00313
cg24026230	*NR3C1*	0.00083	0.00057
cg13986355	*AKT1*	0.00078	0.00065
cg20730067	*NFKB1*	0.00074	0.00066
cg14849556	*NCOA1*	0.00067	0.00146
cg03883275	*MAPK8*	0.00057	0.00108
cg19261497	*MDM2*	0.00057	0.00073
cg16224829	*NR3C1*	0.00056	0.00081
cg01277438	*NFATC1*	0.00047	0.00065
cg11321922	*NR1I3*	0.00046	0.00014
cg02521996	*MAPK3*	0.00039	0.00053
cg23751680	*SMARCA4*	0.00038	0.00031
cg13103915	*CSF2*	0.00038	0.00037
cg26560981	*MAPK10*	0.00035	0.00068
cg16569373	*CREBBP*	0.00033	0.00015
cg09566021	*CREBBP*	0.00027	0.00194
cg14825287	*AKT1*	0.00026	0.00134
cg16182267	*NFATC1*	0.00026	0.00039
cg20090430	*GSK3B*	0.00025	0.00041
cg20509117[+]	*IL6; LOC541472*	0.00023	0.00313
cg05790989[+]	*POU2F1*	0.00019	0.00028
cg17406386	*NR1I3*	0.00017	0.00043
cg04444450[+]	*NCOA2*	0.00011	0.00094
cg01312837	*CREBBP*	0.00010	0.00054
cg05121010	*POU2F1*	0.00010	0.00073
cg23624957	*HDAC2*	0.00008	0.00031
cg05790989[+]	*POU2F1*	0.00004	0.00028

[*] Shown are the Illumina ID for each CpG, UCSC gene name, and feature importance in the wide and long datasets.

[+] CpGs that are consistently important in long and both waves (wave 1 and 2) of wide analyses.

As mentioned, wave 3 PTSS had a prediction power of >88%, therefore we performed analyses to determine how the models work without this highly predictive variable and identified other significant features in the absence of wave-specific PTSS scores. We performed the analyses in two steps, first removing wave 3 PTSS only and then waves 2 and 3 PTSS together. In the first step, wave 2 PTSS had the most predictive power, as shown in Table S2, accounting for 65% and 74% prediction in wide and long analyses, respectively. In the second step, we found that wave 3 lifetime PTSD was the most predictive variable, with 36% and 35% prediction power in wide and long analyses, respectively (Table S3). In both steps, the error rate increased and R^2 decreased when the most predictive features were removed from the analyses. More information about the model performance under these modified conditions is given in the supplementary information (Tables S4 and S5).

3.2. *Prediction on training data and cross-validation*

Next, we used RF, AB, and GB to predict PTSS. These models were used in two different settings, i.e., the base model and tuned model, to search for the best set of parameters. In the base model, all the models used default settings provided by the *Scikit-learn* framework. After training the models in both the settings, we used 10-fold cross-validation on the training data to avoid overfitting of the models. The cross-validation error score of the base models on the training data is shown in Fig. 5. The base models performed very well; cross-validation R^2 values of the base models are shown in Fig. 6. The mean R^2 values for RF and AB were > 85%, and GB > 83%, which indicates the independent variables explain very well the variance of the dependent variable (PTSS). The cross-validation results using the base models were better on the significant set of features as compared to all the features, as shown in Figs. 5 and 6.

3.3. *Prediction using tuned models*

To potentially increase the performance of the base models, we ran 100 iterations to search for the best hyperparameters to tune the models. The results, RMSE and R^2, are shown in the supplementary Figures S5 and S6, respectively. Tuning the hyperparameters did not increase the accuracy of the models. R^2 for RF = 86.4% and GB = 85.1% in the long set; in fact, RF showed a small decrease in R^2, suggesting the base model parameters are optimal.

3.4. *Prediction on test data*

After confirming the scores of base and tuned models using cross-validation to avoid over-fitting, we used the models on test data to get the final prediction score of the PTSS. The base model test score is shown in Table 3, and the tuned model score is shown in Table S6. The highest R^2 values achieved on long test data using the base model and important set of features were GB = 95.29%, followed by RF = 95.10%. Models performed better on the important set of features compared to all features in wide data.

We also implemented many other models such as BR, LR, SVR, VR, the scores of which are shown in Table S7. More information about the methods and results is given in the supplementary information.

4. Discussion

While previous research suggests that prior psychopathology, social adversity, and genomic variation at GRRN genes shape risk of traumatic stress, it remains unclear which of these factors figure most prominently in increasing prospective risk of PTSD. It is also clinically challenging to identify individuals at higher risk of PTSD because of its complex etiology and clinical course. Here we applied ML approaches to identify the features that associate with elevated prospective risk of PTSD. We used DNAm data from GRRN genes and multiple survey measures to determine the role of prior psychopathology, GRRN CpGs and social adversity as risk factors of future traumatic stress. ML identified prior PTSS as the most important predictor for prospective risk of PTSD. Many additional factors including PTS symptom clusters, loneliness, perceived discrimination and GRRN CpGs were identified as significant predictors for PTSD risk. Using these identified factors, ML models predicted the prospective risk of PTSD with high accuracy. We could not compare the accuracy of models in this study with existing approaches for PTSD (Dean et al., 2019; Karstoft et al., 2015; Schultebraucks et al., 2020; Wshah et al., 2019; Zhang et al., 2020) because this prior work used classification approaches, whereas we applied a regression task on a continuous outcome, PTSS, and different performance measures are used for classification and regression.

To apply our ML approach, we evaluated multiple models with different settings, i.e. base, and tuned models, on both long and wide datasets with different sets of features. We found that, in general, both base and tuned models perform very well, as confirmed using cross-validation on the training set and prediction on the final test set. Mean values on 10-fold cross-validation on the training sets showed the

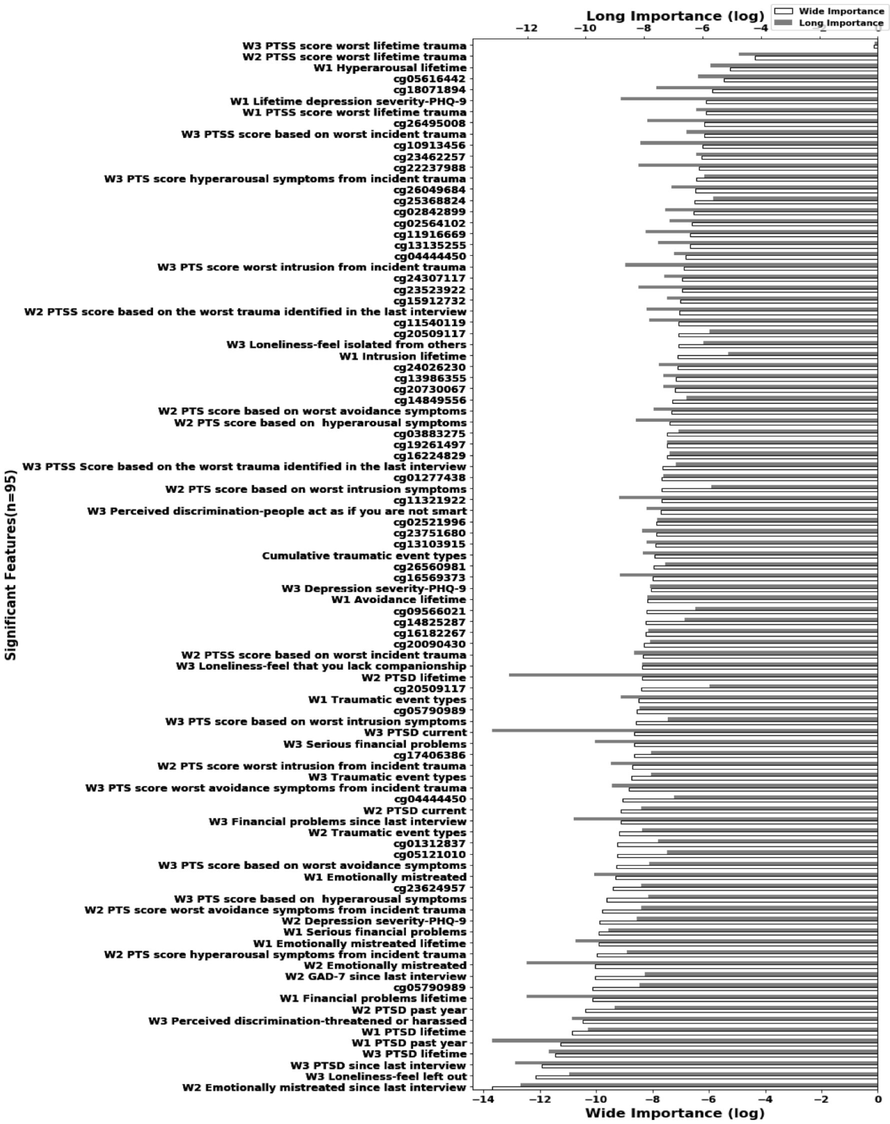

Fig. 3. Features identified as important in both the long and wide datasets. Feature importance is in log scale, and the values near 0 indicate higher importance. The importance of a significant feature can be between 0 and 1 and importance of all features sums to 1. Filled bar: significance levels identified in the long dataset. White bar: significance levels identified in the wide dataset.

RF model to be the best performing model on the training set, followed by AB and GB. However, GB outperformed other methods using base model on test set. Results show that both GB and RF performed better, showing a lower error rate and higher R^2 compared to other approaches, and the difference in prediction between the two is small. In the tuned model on test data, RF was the best model, but it was still less than the base model. Overall, all models showed better accuracy on the long vs. the wide dataset, both in terms of models applied using important and all features. One potential reason for this is the difference in sample size (210 in long vs. 148 in wide datasets). Another reason may be the missing values that are created when arranging the data in wide format. These missing values were imputed, but there is always some difference between the observed and imputed values. Two methods, SVR and VR, performed better using important features on the wide set as compared to the all the features on the long set but still less than the score on the important features on the long set. In general, SVR and LR performed poorly as compared to other approaches, whether used for long or wide data, full, or important features.

The results showed that not all features from the GRRN genes and phenotypes are needed to predict the risk of PTSS, or in other words, using only significant features increases the accuracy of prediction. The sets of 150 significant features produced better results when compared with other feature sets or all the features. It is evident that the significant sets of features both for the long and the wide datasets decreases the error rate and improves the prediction of the models. With this set, we identified significant CpGs and phenotypes as risk factors that predict the prospective risk of PTSS. From features identified as important in the long and wide datasets, we identified common features, where phenotypes were more numerous compared to CpGs.

From the list of significant features in phenotypes, PTSS in prior waves was highly predictive of prospective risk of PTSS. In particular, PTSS from wave 3 was of the highest importance in predicting wave 4 PTSS, confirming the importance of recency of symptom burden to prediction of future symptoms. This feature alone contributed >88% of the significance in both long and wide approaches, and all 149 other features contributed to the remaining percentage. One interesting note is that the feature had almost the same score in both long and wide datasets. The symptom severity at wave 2 is also highly significant.

These results are consistent with results of previous studies. For example, in a prospective study that followed adolescents and young adults for up to 50 months, 52% of the PTSD cases remitted during the follow-up period, whereas the remaining 48% showed no significant remission of PTSD (Perkonigg et al., 2005). The authors concluded that PTSD is usually a persistent and chronic disorder, and that symptom clusters might also be associated with the chronic course of PTSD (Perkonigg et al., 2005). Another study looking at PTSD, anxiety, and depression after the Spitak earthquake and subsequent political violence in Armenia showed no remission of PTSD over a 3-year interval, but that depression symptoms subsided over time (Goenjian et al., 2000). A longitudinal study showed that trauma-related psychopathology increases the risk of PTSD and significant impairment over time (Lewis et al., 2019). Our results also showed that prior post-traumatic psychopathology is highly important in predicting the prospective risk of PTSS in the subsequent year. We saw a decrease in the predictive significance of PTSS in relation to more distant symptoms, with wave 3 symptoms showing the highest importance in predicting wave 4 PTSS, followed by wave 2 and wave 1 symptoms. The higher importance of recent vs. more distant PTS symptoms in predicting future PTSS is consistent with a previous study that reported a general diminution in PTSD symptom severity over time (Yehuda et al., 2009). We also found prior depression and anxiety as significant predictors for PTSS risk, consistent with a prior ML study (Schultebraucks et al., 2020), which showed pre-deployment depression and anxiety as risk factors for PTSD in army personnel deployed to Afghanistan. In addition, our results showed that all three PTSD symptom clusters (intrusion, avoidance, and hyperarousal), appear on the significant feature list. The symptom cluster, hyperarousal, that is detected as very significant in predicting PTSS by ML, has been associated and reported as the first symptoms to occur with chronic PTSD, followed by avoidance and intrusion symptoms (Bremner et al., 1996). Our results suggest that symptom clusters contribute to the chronic course of PTSD and that people with higher symptom severities, both overall and across multiple symptom domains, are more likely to be at high risk of PTSD in the future.

We also found that wave 2 PTSS was the most predictive feature to predict prospective PTSD risk in the absence of wave 3 PTSS, and wave 3 PTSD life was most significant when both wave 2 and 3 PTSS variables were removed from the analyses. The error rate increased, and R^2 decreased for all the models when excluding these significant features (wave 2 and 3 PTSS). Removing the most significant features showed a clear trend of reducing model performance; however, it is interesting to note that in both updated models, some form of prior PTSS remained the most significant feature (i.e. wave 2 PTSS in the first model, and lifetime PTSD at wave 3 in the second model). These updated results support our conclusion that prior PTSS is highly important for predicting prospective PTSD risk.

In addition to these symptom-related findings, our ML results showed other phenotypes are significantly associated with predicting prospective risk of PTSD. For example, number of traumatic event types was identified as an important feature both overall and at waves 1 and 2, with similar levels of importance in both the long and wide datasets. More importantly, our analyses showed that the cumulative traumatic event type is a better predictor for PTSD risk as compared to traumatic

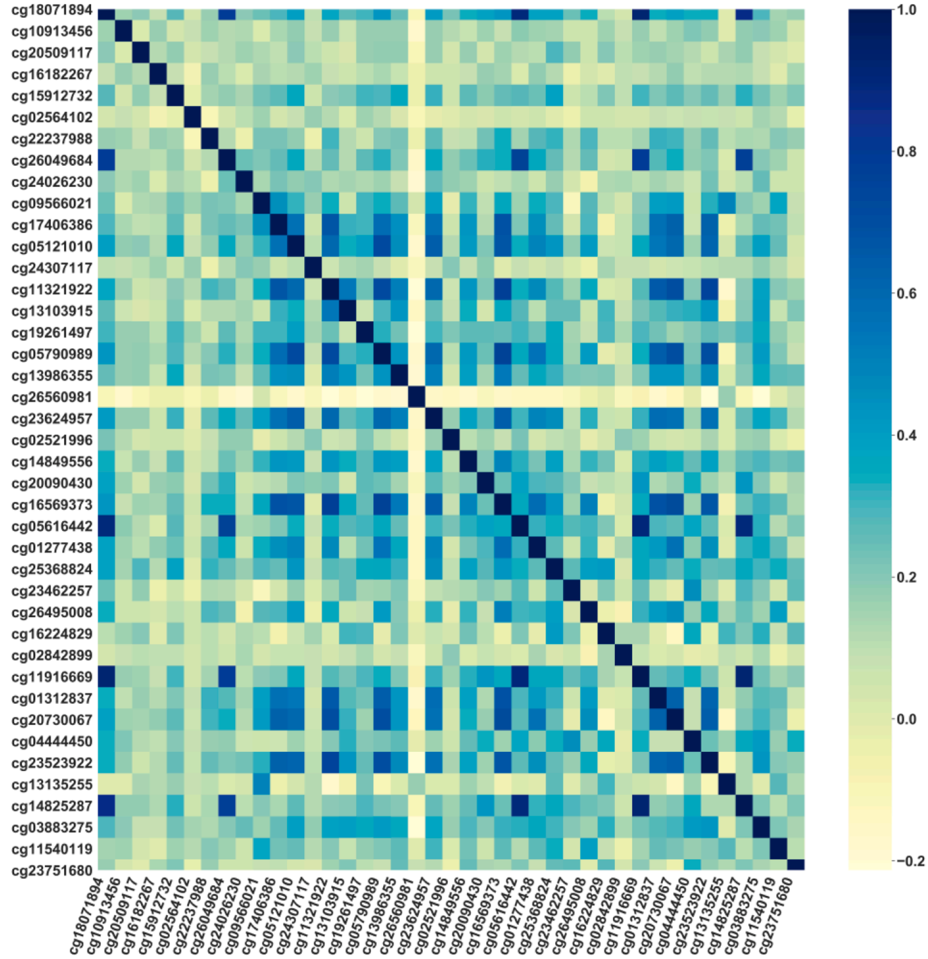

Fig. 4. Correlation plot of DNAm values in CpGs identified as important in both long and wide analysis. Almost all the CpGs are positively correlated with each other. One CpG (cg26560981) shows a negative correlation with the largest number of CpGs.

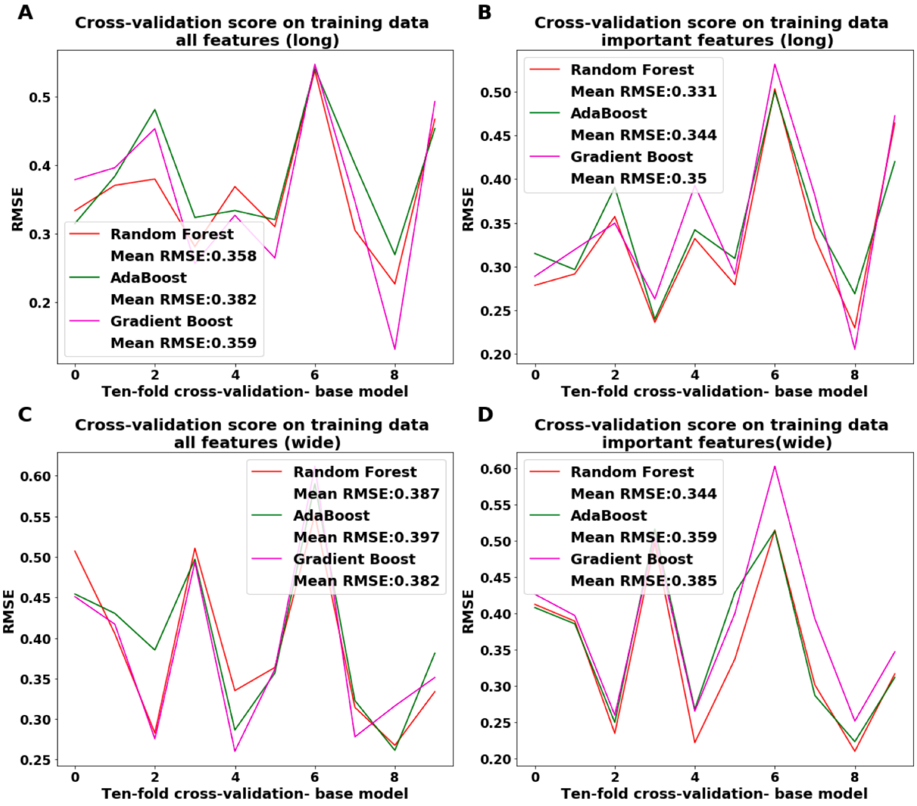

Fig. 5. Ten-fold cross-validation on training data showing RMSE score on each of the ten-folds and mean RMSE on all ten-folds for each model. A) All features (GRRN genes, cell proportions and phenotype), and using DNAm features in long data. B) Important features (150) using DNAm in long. C) All features, using methylation in wide data. D) Important features, methylation in wide. The legend in each plot shows the mean RMSE on ten-fold cross-validation for each model. Using important features proved a better score (less RMSE) for both long and wide data.

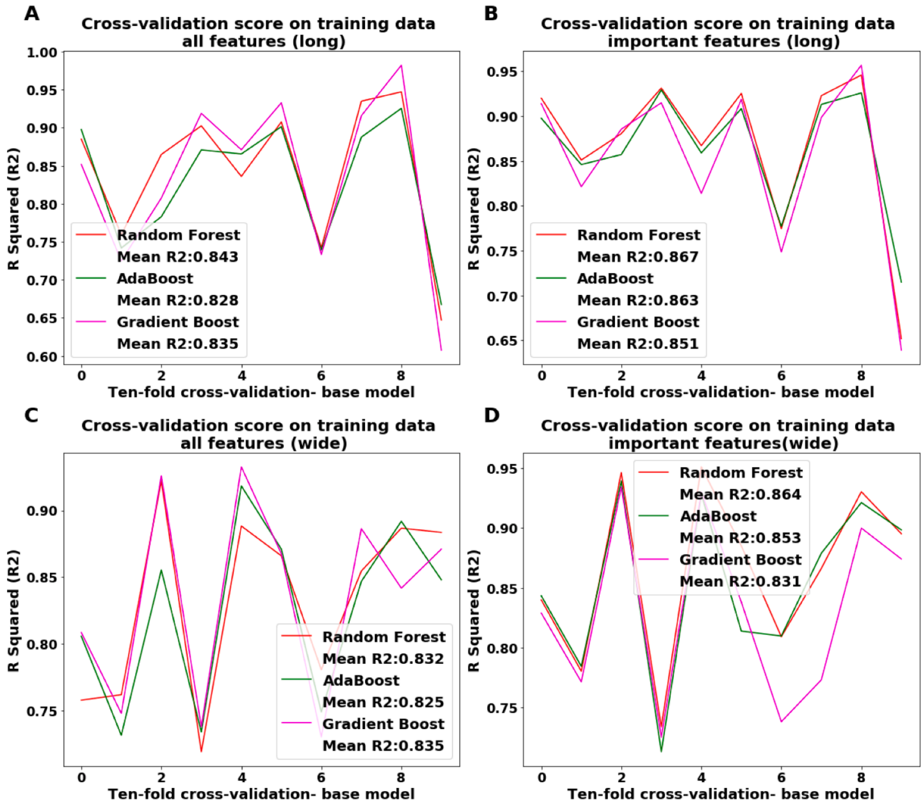

Fig. 6. R squared (R^2) values on each of the ten-folds and mean R^2 for each model. The order of subplots A, B, C, D is the same as described in Fig. 5. Results show better R^2 value on an important set of features, both long and wide data, for all three models as compared to full data.

Table 3
Performance measures on the test set using the base model.

Model	Data	Mean Absolute Error	Mean Square Error	Root Mean Square Error	R Squared
Adaboost	Full long	0.2128	0.0871	0.2952	0.9129
Gradient Boost		**0.1830**	**0.0752**	**0.2742**	**0.9248**
Random Forest		0.1779	0.0715	0.2674	0.9285
Adaboost	Full wide	0.2728	0.1420	0.3768	0.8580
Gradient Boost		**0.2639**	**0.1204**	**0.3469**	**0.8796**
Random Forest		0.2638	0.1248	0.3532	0.8752
Adaboost	Important long	0.2161	0.0834	0.2888	0.9166
Gradient Boost		**0.1353**	**0.0471**	**0.2171**	**0.9529**
Random Forest		0.1403	0.0490	0.2214	0.9510
Adaboost	Important wide	0.2452	0.1204	0.3470	0.8796
Gradient Boost		**0.2059**	**0.0951**	**0.3084**	**0.9049**
Random Forest		0.2205	0.1070	0.3271	0.8930

Full long: All the methylation features in the long data. Full wide: All methylation features in wide data. Important long: Important features in the long data. Important wide: important features in the wide data. All three models performed best on the important set of features in long data. Results show GB to be the best performing approach.

event types from a single wave. Previous work has shown that cumulative traumatic burden is associated with increased risk of PTSD (Copeland et al., 2007; Ogle et al., 2014), and additional work (Perkonigg et al., 2005) has reported that participants with a chronic course of PTSD are more likely to experience new traumatic events. Many additional social adversity factors, including loneliness, perceived discrimination, emotional mistreatment, were also predictive of prospective risk of PTSD in our analyses. Chronic PTSD has been associated with reduced social support, a higher frequency of social phobia, and greater avoidance symptoms (Davidson et al., 1991). These results show that social adversity and trauma exposure contribute to the increased risk of PTSS even when biologic factors are taken into account, suggesting a multiplicity of mechanisms that explain the pathogenesis of PTSD.

Our work also identified GRRN CpG sites whose DNAm measures were predictive of prospective PTSD risk. There are 44 CpGs that were consistently significant in both long and wide analyses. Out of 44 CpGs, 3 CpGs (*cg04444450: NCOA2, cg20509117: IL6; LOC541472, cg05790989: POU2F1*) were significant in the long dataset and both the waves (wave 1 and wave 2) of the wide dataset. The significance indicates that the CpGs are highly important and consistent. All three of the genes related to these CpGs have been associated with PTSD. For example, Nuclear Receptor Coactivator 2 (*NCOA2)* has been identified as a biomarker for PTSD (Breen et al., 2019), and Interleukin 6 (*IL6)* has been associated with PTSD (Haxhibeqiri et al., 2019; Lima et al., 2019; Pervanidou et al., 2007; Somvanshi et al., 2020). POU Class 2 Homeobox 1 (*POU2F1)* has been implicated as a transcription factor located proximal to a SNP associated with emotional memory formation in PTSD (Wilker et al., 2018). *IL6* has been associated with depression and anxiety as well (Crawford et al., 2018; Ryan et al., 2017; Sanada et al., 2020). The CpGs *cg05616442* and *cg18071894* in gene AKT Serine/Threonine Kinase 1 (*AKT1*) showed highest importance score when looking at long and wide scores together. This gene has been associated with depression previously (Starnawska et al., 2019). The other CpGs on the significant list was *cg26495008* and *cg10913456*, both in the FKBP Prolyl Isomerase 5 (*FKBP5*) gene, previously associated with childhood maltreatment and depression (Klinger-König et al., 2019; Mehta et al., 2013); and risk of PTSD (Binder et al., 2008; Pape et al., 2018). The role of other significant genes detected by ML include: Nuclear Factor Kappa B Subunit 1 (*NFKB1*), which has been associated with personality disorders (Gescher et al., 2018); Interleukin 4 (*IL4*), which has been associated with PTSD (Smith et al., 2011); Nuclear Receptor Subfamily 3 Group C Member 1 (*NR3C1*), previously associated with emotion dysregulation, psychopathology (Cicchetti and Handley, 2017), depression (Borçoi et al., 2020; Efstathopoulos et al., 2018), and risk of PTSD (Schechter et al., 2015; Vukojevic et al., 2014); and Nuclear Factor Of Activated T Cells 1 (*NFATC1*), which has been previously associated with PTSD and depression (Kuan et al., 2017). It is clear from the literature that GRRN genes play a crucial role in mediating the stress response, and many genes identified here by our ML analyses have been previously associated with traumatic stress and stress-related psychopathology.

This study has several limitations: First, the sample size used in this study is limited, and a larger sample size could increase prediction accuracy and may help to design better generalized models. Second, many phenotype variables have missing data, a common issue in many domains. We imputed the missing values, but some differences may remain between imputed and the observed values. Missing values often represent hidden patterns in the data, and complete data may have provided more insights in this study. Future work involving a more complete dataset may address this issue. Third, we could not assess the generalizability of the ML models due to the lack of an available independent, external dataset. In the future, the availability of such datasets will help to validate the ML models. Finally, in this dataset, we could not demonstrate the causal relationship of the identified predictors with the increased risk of PTSD, due to the limitations inherent to working human study participants.

Despite these limitations, this study has many strengths. The data used in this study is rich in terms of data collected about social adversity factors and biology, and there are very few large samples with the richness of data we have here and none, to our knowledge, that have leveraged such data for ML approaches in PTSD. In addition, we used standard pipeline and approaches for our analyses and implemented multiple ML models; these ML models are very efficient in predicting the prospective risk, confirmed by cross-validation, and using the test dataset. Our applied approach identified a specific list of DNAm features and phenotypes that play a significant role as predictors for prospective PTSD risk with high accuracy. These features chosen by ML as significant have been previously associated with PTSD. Our results suggest that ML can effectively use a wide variety of data such as DNAm, psychopathology, social adversity, and cell proportions to address the challenging task of identifying the associated risk factors and predicting PTSS risk.

In conclusion, it is both challenging and crucial to identify people at higher risk of PTSD, and to understand the factors associated with elevating the prospective risk of traumatic stress. Our results show that ML approaches are efficient in identifying the factors that predict the prospective risk of PTSD with high accuracy. Many of the factors identified by ML as risk predictors are consistent with previous studies exploring the determinants of PTSD; our results extend this prior work, however, by assigning relative importance of these determinants from a wide range of social, psychopathological, and genomic features that has not been previously examined in the context of ML-based risk prediction of PTSD. Results from this study suggest that ML approaches may be further developed to detect elevated post-traumatic risk efficiently to assist early intervention.

Authors contribution

Agaz H Wani: Conceptualization, Methodology, Analysis, Interpretation, Writing - original draft, review & editing. **Allison E. Aiello:** Conceptualization, Interpretation, Writing - review & editing. **Grace S. Kim**: Writing - review & editing. **Fei Xue**: Writing - review & editing. **Chantel L. Martin:** Writing - review & editing. **Andrew Ratanatharathorn:** Writing - review & editing. **Annie Qu:** Conceptualization, Writing - review & editing. **Karestan Koenen:** Conceptualization, Interpretation, Writing - review & editing. **Sandro Galea:** Conceptualization, Interpretation, Writing - review & editing. **Derek E. Wildman:** Conceptualization, Interpretation, Writing - review & editing. **Monica Uddin:** Conceptualization, Supervision, Interpretation, Writing - original draft, review & editing.

Funding source

This work was supported by the National Institutes of Health [R01DA022720, R01DA022720-S1, RC1MH088283, R01 MD011728, 3R01MD011728-02S1], and additionally supported by grants (1U01MH115485 and R00MD012808).

Declaration of Competing Interest

The authors have no competing interests to declare.

Acknowledgments

We thank Mr. Zachary Graham and Ms. Sarah Burgan, who kindly helped in sample plating to generate data analyzed in this manuscript. We appreciate the time and effort of study participants, staff and volunteers of the Detroit Neighborhood Health Study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2020.12.076.

References

An, K., & Meng, J. (2010). *Voting-averaged combination method for regressor ensemble*, Berlin, Heidelberg.

Anacker, C., Zunszain, P.A., Carvalho, L.A., Pariante, C.M., 2011. The glucocorticoid receptor: pivot of depression and of antidepressant treatment? Psychoneuroendocrinology 36 (3), 415–425. https://doi.org/10.1016/j.psyneuen.2010.03.007.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., Irizarry, R.A., 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30 (10), 1363–1369. https://doi.org/10.1093/bioinformatics/btu049.

Barfield, R.T., Kilaru, V., Smith, A.K., Conneely, K.N., 2012. CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinformatics 28 (9), 1280–1281. https://doi.org/10.1093/bioinformatics/bts124.

Binder, E.B., Bradley, R.G., Liu, W., Epstein, M.P., Deveau, T.C., Mercer, K.B., Ressler, K. J., 2008. Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. JAMA 299 (11), 1291–1305. https://doi.org/10.1001/jama.299.11.1291.

Bisson, J.I., Cosgrove, S., Lewis, C., Robert, N.P., 2015. Post-traumatic stress disorder. BMJ 351, h6161. https://doi.org/10.1136/bmj.h6161 h6161.

Bogart, L.M., Wagner, G.J., Galvan, F.H., Landrine, H., Klein, D.J., Sticklor, L.A., 2011. Perceived discrimination and mental health symptoms among Black men with HIV. Cult. Divers. Ethn. Minor. Psychol. 17 (3), 295–302. https://doi.org/10.1037/a0024056.

Bonanno, G.A., 2004. Loss, trauma, and human resilience: have we underestimated the human capacity to thrive after extremely aversive events? Am. Psychol. 59 (1), 20–28. https://doi.org/10.1037/0003-066x.59.1.20.

Borçoi, A.R., Mendes, S.O., Gasparini dos Santos, J., Mota de Oliveira, M., Moreno, I.A. A., Freitas, F.V., Álvares-da-Silva, A.M., 2020. Risk factors for depression in adults: NR3C1 DNA methylation and lifestyle association. J. Psychiatr. Res. 121, 24–30. https://doi.org/10.1016/j.jpsychires.2019.10.011.

Bradley, R., Greene, J., Russ, E., Dutra, L., Westen, D., 2005. A multidimensional meta-analysis of psychotherapy for PTSD. Am. J. Psychiatry 162 (2), 214–227. https://doi.org/10.1176/appi.ajp.162.2.214.

Breen, M.S., Bierer, L.M., Daskalakis, N.P., Bader, H.N., Makotkine, I., Chattopadhyay, M., Yehuda, R., 2019. Differential transcriptional response following glucocorticoid activation in cultured blood immune cells: a novel approach to PTSD biomarker development. Transl. Psychiatry 9 (1), 201. https://doi.org/10.1038/s41398-019-0539-x.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., 2001. Random forests. Mach, Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Bremner, J.D., Southwick, S.M., Darnell, A., Charney, D.S., 1996. Chronic PTSD in Vietnam combat veterans: course of illness and substance abuse. Am. J. Psychiatry 153 (3), 369–375. https://doi.org/10.1176/ajp.153.3.369.

Brooks Holliday, S., Dubowitz, T., Haas, A., Ghosh-Dastidar, B., DeSantis, A., Troxel, W. M., 2018. The association between discrimination and PTSD in African Americans: exploring the role of gender. Ethn. Health 1–15. https://doi.org/10.1080/13557858.2018.1444150.

Bustamante, A.C., Aiello, A.E., Galea, S., Ratanatharathorn, A., Noronha, C., Wildman, D. E., Uddin, M., 2016. Glucocorticoid receptor DNA methylation, childhood maltreatment and major depression. J. Affect. Disord. 206, 181–188. https://doi.org/10.1016/j.jad.2016.07.038.

Cacioppo, J.T., Cacioppo, S., Capitanio, J.P., Cole, S.W., 2015. The neuroendocrinology of social isolation. Annu. Rev. Psychol. 66, 733–767. https://doi.org/10.1146/annurev-psych-010814-015240.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol.). https://doi.org/10.1145/1961189.1961199. Article 27.

Cicchetti, D., Handley, E.D., 2017. Methylation of the glucocorticoid receptor gene, nuclear receptor subfamily 3, group C, member 1 (NR3C1), in maltreated and nonmaltreated children: associations with behavioral undercontrol, emotional lability/negativity, and externalizing and internalizing symptoms. Dev. Psychopathol. 29 (5), 1795–1806. https://doi.org/10.1017/S0954579417001407.

Copeland, W.E., Keeler, G., Angold, A., Costello, E.J., 2007. Traumatic events and posttraumatic stress in childhood. Arch. Gen. Psychiatry 64 (5), 577–584. https://doi.org/10.1001/archpsyc.64.5.577.

Crawford, B., Craig, Z., Mansell, G., White, I., Smith, A., Spaull, S., Murphy, T.M., 2018. DNA methylation and inflammation marker profiles associated with a history of depression. Hum. Mol. Genet. 27 (16), 2840–2850. https://doi.org/10.1093/hmg/ddy199.

Dalenberg, C.J., Glaser, D., & Alhassoon, O.M. (2012). Statistical support for subtypes in posttraumatic stress disorder: the how and why of subtype analysis. 29(8), 671–678. doi:10.1002/da.21926.

Davidson, J.R., Hughes, D., Blazer, D.G., George, L.K., 1991. Post-traumatic stress disorder in the community: an epidemiological study. Psychol. Med. 21 (3), 713–721. https://doi.org/10.1017/s0033291700022352.

Dean, K.R., Hammamieh, R., Mellon, S.H., Abu-Amara, D., Flory, J.D., Guffanti, G., The, P.S.B.C., 2019. Multi-omic biomarker identification and validation for diagnosing warzone-related post-traumatic stress disorder. Mol. Psychiatry. https://doi.org/10.1038/s41380-019-0496-z.

Drucker, H., 1997. Improving Regressors using Boosting Techniques. In: Paper Presented at the Proceedings of the Fourteenth International Conference on Machine Learning.

Efstathopoulos, P., Andersson, F., Melas, P.A., Yang, L.L., Villaescusa, J.C., Rȕegg, J., Lavebratt, C., 2018. NR3C1 hypermethylation in depressed and bullied adolescents. Transl Psychiatry 8 (1), 121. https://doi.org/10.1038/s41398-018-0169-8.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Galatzer-Levy, I.R., Bryant, R.A., 2013. 636,120 ways to have posttraumatic stress disorder. Perspect. Psychol. Sci. 8 (6), 651–662. https://doi.org/10.1177/1745691613504115.

Gescher, D.M., Kahl, K.G., Hillemacher, T., Frieling, H., Kuhn, J., Frodl, T., 2018. Epigenetics in personality disorders: today's insights. Front. Psychiatry 9, 579. https://doi.org/10.3389/fpsyt.2018.00579, 579.

Ginzburg, K., Ein-Dor, T., Solomon, Z., 2010. Comorbidity of posttraumatic stress disorder, anxiety and depression: a 20-year longitudinal study of war veterans. J. Affect. Disord. 123 (1), 249–257. https://doi.org/10.1016/j.jad.2009.08.006.

Goenjian, A.K., Steinberg, A.M., Najarian, L.M., Fairbanks, L.A., Tashjian, M., Pynoos, R. S., 2000. Prospective study of posttraumatic stress, anxiety, and depressive reactions after earthquake and political violence. Am. J. Psychiatry 157 (6), 911–916. https://doi.org/10.1176/appi.ajp.157.6.911.

Goldmann, E., Aiello, A., Uddin, M., Delva, J., Koenen, K., Gant, L.M., Galea, S., 2011. Pervasive exposure to violence and posttraumatic stress disorder in a predominantly African American urban community: the detroit neighborhood health study. J. Trauma Stress 24 (6), 747–751. https://doi.org/10.1002/jts.20705.

Harper, K.N., Peters, B.A., Gamble, M.V., 2013. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. Cancer Epidemiol. Biomark. Prev. 22 (6), 1052–1060. https://doi.org/10.1158/1055-9965.Epi-13-0114.

Haxhibeqiri, V., Haxhibeqiri, S., Topciu-Shufta, V., Agani, F., Goci Uka, A., Hoxha, B., Babić, D., 2019. The association of catechol-o-methyl-transferase and interleukin 6 gene polymorphisms with posttraumatic stress disorder. Psychiatr. Danub. 31 (2), 241–248. https://doi.org/10.24869/psyd.2019.241.

Heiss, J.A., Just, A.C., 2018. Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. Clin. Epigenet. 10 (1), 73. https://doi.org/10.1186/s13148-018-0504-1.

Hodes, G.E., Ménard, C., Russo, S.J., 2016. Integrating Interleukin-6 into depression diagnosis and treatment. Neurobiol. Stress 4, 15–22. https://doi.org/10.1016/j.ynstr.2016.03.003.

Hughes, M.E., Waite, L.J., Hawkley, L.C., Cacioppo, J.T., 2004. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res. Aging 26 (6), 655–672. https://doi.org/10.1177/0164027504268574.

Hyland, P., Shevlin, M., Cloitre, M., Karatzias, T., Vallières, F., McGinty, G., Power, J.M., 2019. Quality not quantity: loneliness subtypes, psychological trauma, and mental health in the US adult population. Soc. Psychiatry Psychiatr. Epidemiol. 54 (9), 1089–1099. https://doi.org/10.1007/s00127-018-1597-8.

Johnson, W.E., Li, C., Rabinovic, A., 2006. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8 (1), 118–127. https://doi.org/10.1093/biostatistics/kxj037. J Biostatistics.

Karstoft, K.I., Galatzer-Levy, I.R., Statnikov, A., Li, Z., Shalev, A.Y., 2015. Bridging a translational gap: using machine learning to improve the prediction of PTSD. BMC Psychiatry 15, 30. https://doi.org/10.1186/s12888-015-0399-8.

Kessler, R.C., Mickelson, K.D., Williams, D.R., 1999. The prevalence, distribution, and mental health correlates of perceived discrimination in the United States. J. Health Soc. Behav. 40 (3), 208–230.

Kiely, K.M., Leach, L.S., Olesen, S.C., Butterworth, P., 2015. How financial hardship is associated with the onset of mental health problems over time. Soc. Psychiatry Psychiatr. Epidemiol. 50 (6), 909–918. https://doi.org/10.1007/s00127-015-1027-0.

Klinger-König, J., Hertel, J., Van der Auwera, S., Frenzel, S., Pfeiffer, L., Waldenberger, M., Grabe, H.J., 2019. Methylation of the FKBP5 gene in association with FKBP5 genotypes, childhood maltreatment and depression. Neuropsychopharmacology 44 (5), 930–938. https://doi.org/10.1038/s41386-019-0319-6.

Koenen, K.C., Moffitt, T.E., Poulton, R., Martin, J., Caspi, A., 2007. Early childhood factors associated with the development of post-traumatic stress disorder: results from a longitudinal birth cohort. Psychol. Med. 37 (2), 181–192. https://doi.org/10.1017/s0033291706009019.

Kuan, P.F., Waszczuk, M.A., Kotov, R., Marsit, C.J., Guffanti, G., Gonzalez, A., Luft, B.J., 2017. An epigenome-wide DNA methylation study of PTSD and depression in World Trade Center responders. Transl. Psychiatry 7 (6), e1158. https://doi.org/10.1038/tp.2017.130 e1158.

Kuwert, P., Knaevelsrud, C., Pietrzak, R.H., 2014. Loneliness among older veterans in the United States: results from the national health and resilience in Veterans study. Am. J. Geriatr. Psychiatry 22 (6), 564–569. https://doi.org/10.1016/j.jagp.2013.02.013.

Labonté, B., Azoulay, N., Yerko, V., Turecki, G., Brunet, A., 2014. Epigenetic modulation of glucocorticoid receptors in posttraumatic stress disorder. Transl. Psychiatry 4 (3), e368. https://doi.org/10.1038/tp.2014.3 e368.

Lai, T.L., Robbins, H., Wei, C.Z., 1978. Strong consistency of least squares estimates in multiple regression. PNAS 75 (7), 3034–3036. https://doi.org/10.1073/pnas.75.7.3034.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28 (6), 882–883. https://doi.org/10.1093/bioinformatics/bts034.

Lewis, S.J., Arseneault, L., Caspi, A., Fisher, H.L., Matthews, T., Moffitt, T.E., Danese, A., 2019. The epidemiology of trauma and post-traumatic stress disorder in a representative cohort of young people in England and Wales. Lancet Psychiatry 6 (3), 247–256. https://doi.org/10.1016/S2215-0366(19)30031-8.

Lima, B.B., Hammadah, M., Wilmot, K., Pearce, B.D., Shah, A., Levantsevych, O., Vaccarino, V., 2019. Posttraumatic stress disorder is associated with enhanced interleukin-6 response to mental stress in subjects with a recent myocardial infarction. Brain Behav. Immun. 75, 26–33. https://doi.org/10.1016/j.bbi.2018.08.015.

Link, B.G., Phelan, J., 1995. Social conditions as fundamental causes of disease. J. Health Soc. Behav. 80–94. https://doi.org/10.2307/2626958.

McCartney, D.L., Walker, R.M., Morris, S.W., McIntosh, A.M., Porteous, D.J., Evans, K.L., 2016. Identification of polymorphic and off-target probe binding sites on the illumina infinium methylation EPIC BeadChip. Genom. Data 9, 22–24. https://doi.org/10.1016/j.gdata.2016.05.012.

Mehta, D., Klengel, T., Conneely, K.N., Smith, A.K., Altmann, A., Pace, T.W., Binder, E.B., 2013. Childhood maltreatment is associated with distinct genomic and epigenetic profiles in posttraumatic stress disorder. J. Proc. Natl. Acad. Sci. 110 (20), 8302–8307. https://doi.org/10.1073/pnas.1217750110.

Ogle, C.M., Rubin, D.C., Siegler, I.C., 2014. Cumulative exposure to traumatic events in older adults. Aging Ment. Health 18 (3), 316–325. https://doi.org/10.1080/13607863.2013.832730.

Pape, J.C., Carrillo-Roa, T., Rothbaum, B.O., Nemeroff, C.B., Czamara, D., Zannas, A.S., Binder, E.B., 2018. DNA methylation levels are associated with CRF1 receptor antagonist treatment outcome in women with post-traumatic stress disorder. Clin. Epigenet. 10 (1), 136. https://doi.org/10.1186/s13148-018-0569-x.

Park, A.L., Fuhrer, R., Quesnel-Vallée, A., 2013. Parents' education and the risk of major depression in early adulthood. Soc. Psychiatry Psychiatr. Epidemiol. 48 (11), 1829–1839. https://doi.org/10.1007/s00127-013-0697-8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830 (null).

Perkonigg, A., Pfister, H., Stein, M.B., Höfler, M., Lieb, R., Maercker, A., Wittchen, H.U., 2005. Longitudinal course of posttraumatic stress disorder and posttraumatic stress disorder symptoms in a community sample of adolescents and young adults. Am. J. Psychiatry 162 (7), 1320–1327. https://doi.org/10.1176/appi.ajp.162.7.1320.

Pervanidou, P., Kolaitis, G., Charitaki, S., Margeli, A., Ferentinos, S., Bakoula, C., Chrousos, G.P., 2007. Elevated morning serum interleukin (IL)-6 or evening salivary cortisol concentrations predict posttraumatic stress disorder in children and adolescents six months after a motor vehicle accident. Psychoneuroendocrinology 32 (8–10), 991–999. https://doi.org/10.1016/j.psyneuen.2007.07.001.

Ramos-Lima, L.F., Souza, P.R.A., Teche, S.P., Freitas, L.H.M, 2019. Trauma-related disorders in a low- to middle-income country: a four-year follow-up of outpatient trauma in Brazil. Psychiatry Res. 280, 112525 https://doi.org/10.1016/j.psychres.2019.112525.

Ryan, J., Pilkington, L., Neuhaus, K., Ritchie, K., Ancelin, M.L., Saffery, R., 2017. Investigating the epigenetic profile of the inflammatory gene IL-6 in late-life depression. BMC Psychiatry 17 (1), 354. https://doi.org/10.1186/s12888-017-1515-8.

Salas, L.A., Koestler, D.C., Butler, R.A., Hansen, H.M., Wiencke, J.K., Kelsey, K.T., Christensen, B.C., 2018. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 19 (1), 64. https://doi.org/10.1186/s13059-018-1448-7, 64.

Sanada, K., Montero-Marin, J., Barceló-Soler, A., Ikuse, D., Ota, M., Hirata, A., Iwanami, A., 2020. Effects of mindfulness-based interventions on biomarkers and low-grade inflammation in patients with psychiatric disorders: a meta-analytic review. Int. J. Mol. Sci. 21 (7), 2484. https://doi.org/10.3390/ijms21072484.

Schechter, D.S., Moser, D.A., Paoloni-Giacobino, A., Stenz, L., Gex-Fabry, M., Aue, T., Rusconi Serpa, S., 2015. Methylation of NR3C1 is related to maternal PTSD, parenting stress and maternal medial prefrontal cortical activity in response to child separation among mothers with histories of violence exposure. Front. Psychol. 6, 690. https://doi.org/10.3389/fpsyg.2015.00690.

Schultebraucks, K., Qian, M., Abu-Amara, D., Dean, K., Laska, E., Siegel, C., Marmar, C. R., 2020. Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: a machine-learning approach for analyzing multivariate predictors. Mol. Psychiatry. https://doi.org/10.1038/s41380-020-0789-2.

Shalev, A.Y., 2001. What is posttraumatic stress disorder? J. Clin. Psychiatry 62 (Suppl 17), 4–10.

Smith, A.K., Conneely, K.N., Kilaru, V., Mercer, K.B., Weiss, T.E., Bradley, B., Ressler, K. J., 2011. Differential immune system DNA methylation and cytokine regulation in post-traumatic stress disorder. Am. J. Med. Genet. Part B. 156B (6), 700–708. https://doi.org/10.1002/ajmg.b.31212.

Somvanshi, P.R., Mellon, S.H., Yehuda, R., Flory, J.D., Makotkine, I., Bierer, L., Doyle 3rd., F.J., 2020. Role of enhanced glucocorticoid receptor sensitivity in inflammation in PTSD: insights from computational model for circadian-neuroendocrine-immune interactions. Am. J. Physiol. Endocrinol. Metab. 319 (1), E48–e66. https://doi.org/10.1152/ajpendo.00398.2019.

Starnawska, A., Tan, Q., Soerensen, M., McGue, M., Mors, O., Børglum, A.D., Christiansen, L., 2019. Epigenome-wide association study of depression symptomatology in elderly monozygotic twins. Transl Psychiatry 9 (1), 214. https://doi.org/10.1038/s41398-019-0548-9.

Triche Jr., T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., Siegmund, K.D, 2013. Low-level processing of illumina infinium DNA methylation bead arrays. Nucleic Acids Res. 41 (7), e90. https://doi.org/10.1093/nar/gkt090 e90.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.

Uddin, M., Aiello, A.E., Wildman, D.E., Koenen, K.C., Pawelec, G., de Los Santos, R., Galea, S., 2010. Epigenetic and immune function profiles associated with posttraumatic stress disorder. Proc. Natl. Acad. Sci. U S A, 107 (20), 9470–9475. https://doi.org/10.1073/pnas.0910794107.

Uddin, M., Koenen, K.C., Aiello, A.E., Wildman, D.E., de los Santos, R., Galea, S., 2011. Epigenetic and inflammatory marker profiles associated with depression in a community-based epidemiologic sample. Psychol. Med. 41 (5), 997–1007. https://doi.org/10.1017/S0033291710001674.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. Mice: multivariate imputation by chained equations in R. *2011*. J. Stat. Softw. 45 (3), 67. https://doi.org/10.18637/jss.v045.i03.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag.

Vukojevic, V., Kolassa, I.-.T., Fastenrath, M., Gschwind, L., Spalek, K., Milnik, A., de Quervain, D.J.F., 2014. Epigenetic modification of the glucocorticoid receptor gene is linked to traumatic memory and post-traumatic stress disorder risk in genocide survivors. J. Neurosci. 34 (31), 10274–10284. https://doi.org/10.1523/JNEUROSCI.1526-14.2014.

Ward-Caviness, C.K., Pu, S., Martin, C.L., Galea, S., Uddin, M., Wildman, D.E., Aiello, A. E., 2020. Epigenetic predictors of all-cause mortality are associated with objective measures of neighborhood disadvantage in an urban population. Clin. Epigenet. 12 (1), 44. https://doi.org/10.1186/s13148-020-00830-8.

Whitnall, M.H., 1993. Regulation of the hypothalamic corticotropin-releasing hormone neurosecretory system. Prog. Neurobiol. 40 (5), 573–629. https://doi.org/10.1016/0301-0082(93)90035-Q.

Wilker, S., Schneider, A., Conrad, D., Pfeiffer, A., Boeck, C., Lingenfelder, B., Kolassa, I.-. T., 2018. Genetic variation is associated with PTSD risk and aversive memory: evidence from two trauma-Exposed African samples and one healthy European sample. Transl. Psychiatry 8 (1), 251. https://doi.org/10.1038/s41398-018-0297-1.

Williams, D.R., Yan, Y., Jackson, J.S., Anderson, N.B., 1997. Racial differences in physical and mental health: socio-economic status, stress and discrimination. J. Health Psychol. 2 (3), 335–351. https://doi.org/10.1177/135910539700200305.

Wshah, S., Skalka, C., Price, M., 2019. Predicting posttraumatic stress disorder risk: a machine learning approach. JMIR Ment. Health 6 (7), e13946. https://doi.org/10.2196/13946.

Yehuda, Halligan, S.L., Golier, J.A., Grossman, R., Bierer, L.M, 2004. Effects of trauma exposure on the cortisol response to dexamethasone administration in PTSD and major depressive disorder. Psychoneuroendocrinology 29 (3), 389–404. https://doi.org/10.1016/S0306-4530(03)00052-0.

Yehuda, Schmeidler, J., Labinsky, E., Bell, A., Morris, A., Zemelman, S., Grossman, R.A, 2009. Ten-year follow-up study of PTSD diagnosis, symptom severity and psychosocial indices in aging holocaust survivors. Acta Psychiatr. Scand. 119 (1), 25–34. https://doi.org/10.1111/j.1600-0447.2008.01248.x.

Yehuda, Southwick, S.M., Krystal, J.H., Bremner, D., Charney, D.S., Mason, J.W, 1993. Enhanced suppression of cortisol following dexamethasone administration in posttraumatic stress disorder. Am. J. Psychiatry 150 (1), 83–86. https://doi.org/10.1176/ajp.150.1.83.

Zhang, J., Richardson, J.D., Dunkley, B.T., 2020. Classifying post-traumatic stress disorder using the magnetoencephalographic connectome and machine learning. Sci. Rep. 10 (1), 5937. https://doi.org/10.1038/s41598-020-62713-5.

Zvolensky, M.J., Farris, S.G., Kotov, R., Schechter, C.B., Bromet, E., Gonzalez, A., Luft, B. J., 2015. World Trade Center disaster and sensitization to subsequent life stress: a longitudinal study of disaster responders. Prev. Med. 75, 70–74. https://doi.org/10.1016/j.ypmed.2015.03.017.