Journal of Psychiatric Research 143 (2021) 239–245

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/jpsychires

Artificial intelligence language predictors of two-year trauma-related outcomes

Joshua R. Oltmanns [a,*], H. Andrew Schwartz [a], Camilo Ruggero [b], Youngseo Son [a], Jiaju Miao [a], Monika Waszczuk [c], Sean A.P. Clouston [a], Evelyn J. Bromet [a], Benjamin J. Luft [a], Roman Kotov [a]

[a] Stony Brook University, USA
[b] University of North Texas, USA
[c] Rosalind Franklin University, USA

ARTICLE INFO

Keywords:
artificial Intelligence
Natural language processing
First responders
Assessment
Trauma

ABSTRACT

Background: Recent research on artificial intelligence has demonstrated that natural language can be used to provide valid indicators of psychopathology. The present study examined artificial intelligence-based language predictors (ALPs) of seven trauma-related mental and physical health outcomes in responders to the World Trade Center disaster.

Methods: The responders ($N = 174$, $M_{age} = 55.4$ years) provided daily voicemail updates over 14 days. Algorithms developed using machine learning in large social media discovery samples were applied to the voicemail transcriptions to derive ALP scores for several risk factors (depressivity, anxiousness, anger proneness, stress, and personality). Responders also completed self-report assessments of these risk factors at baseline and trauma-related mental and physical health outcomes at two-year follow-up (including symptoms of depression, post-traumatic stress disorder, sleep disturbance, respiratory problems, and GERD).

Results: Voicemail ALPs were significantly associated with a majority of the trauma-related outcomes at two-year follow-up, over and above corresponding baseline self-reports. ALPs showed significant convergence with corresponding self-report scales, but also considerable uniqueness from each other and from self-report scales.

Limitations: The study has a relatively short follow-up period relative to trauma occurrence and a limited sample size.

Conclusions: This study shows evidence that ALPs may provide a novel, objective, and clinically useful approach to forecasting, and may in the future help to identify individuals at risk for negative health outcomes.

Prediction of outcomes among trauma survivors remains challenging (Kotov et al., 2015; Lee and Park, 2018). Established risk factors for poor mental and physical health outcomes include personality predispositions (e.g., neuroticism) and life stress (DiGangi et al., 2013; Zvolensky et al., 2015). However, assessment of these variables relies heavily on time-consuming questionnaires. An alternative approach analyzes survivors' natural language using artificial intelligence-based language predictors (ALPs). Artificial intelligence models are now able to identify risk characteristics in spoken or written communications (Eichstaedt et al., 2018; Park et al., 2015). ALPs promise to provide objective and reliable evaluations of patients that are automatic, resulting in low cost, low effort, and scalability to large health care systems, which could benefit clinicians. The present study examines them in a longitudinal study of World Trade Center disaster responders, a sample with a high burden of trauma-related health symptoms.

Over the past 20 years, the study of natural language in psychiatry relied primarily on the Linguistic Inquiry and Word Count software (Pennebaker et al., 2007). LIWC is used to extract words from language samples and provides the user with count scores for more than 80 different grammatical, psychological, and topical word clusters. LIWC scores have been associated with mental health (M. P. Sasso et al., 2019) and were able to predict outcomes after a personal trauma (Kleim et al., 2018) and hurricane disaster (K. Marshall et al., 2020). However, LIWC uses only basic information about language such as simple word counts.

In an effort to overcome the limitations of the LIWC approach, machine learning techniques were applied to language used in social media messages. The findings indicate that these techniques substantially improved the ability of language analyses to assess mental health and

* Corresponding author.
E-mail address: joshua.oltmanns@stonybrookmedicine.edu (J.R. Oltmanns).

https://doi.org/10.1016/j.jpsychires.2021.09.015
Received 23 February 2021; Received in revised form 29 June 2021; Accepted 1 September 2021
Available online 6 September 2021

personality (Chancellor and De Choudhury, 2020; Park et al., 2015). The resulting ALPs have shown preliminary evidence for correct classification of people diagnosed with mental disorders and associations with other mental health variables; however, some studies have been limited by modest samples sizes and measurement issues (e.g., training models on self-disclosed diagnosis posted in status updates, rather than clinical assessments), cross-sectional designs, and non-clinical samples (Chancellor and De Choudhury, 2020).

In the present study, we employed ALPs (named for their AI, language, and prediction components) developed on gold standard measures using large discovery samples. ALPs use an *open-vocabulary* approach, which means that models recognize the meaning of two-to-three-word strings (called Ngrams) in addition to one-word counts (unigrams). Specifically, Schwartz et al. (2014) used Facebook status updates from 28,749 Facebook users to develop a depressivity ALP, which was successful correlating $r = .39$ with the self-reported depressivity. Further, topics most correlated with self-reported depressivity also included words used to describe major depressive disorder in the *Diagnostic and Statistical Manual of Mental Disorders—5*th *Edition's* (American Psychiatric Association, 2013) such as hopelessness, meaninglessness, and depressed mood. Park and colleagues created ALPs for the Five-Factor Model (FFM) personality domains and sub-components of neuroticism: Depressiveness, anxiousness, anger proneness (Park et al., 2015). They trained machine learning models on a large sample of Facebook users ($n = 66{,}732$) and cross-validated them in a separate sample of Facebook users ($n = 4824$). Cross-validation supported convergent and discriminant validity of these ALPs. Moreover, ALPs showed stronger associations with self-reported personality than other language-based approaches (e.g., LIWC scores, single-word models) (Eichstaedt et al., 2020). They were further validated through significant correlations with several external criteria including political orientation, number of Facebook friends, and satisfaction with life, and ALPs showed good test-retest stability across six months (mean $r = 0.70$).

Eichstaedt et al. (2018) trained depression ALP on social media collected from a sample of patients ($n = 569$ non-depressed and $n = 114$ depressed) who made their electronic medical records available, including depression diagnoses. With moderate accuracy, the patients who became depressed could be identified via their social media language even before diagnosis (area under the curve = 0.69). Merchant et al. (2019) showed that ALPs from 999 social media users also predicted anxiety diagnoses, in addition to depression diagnoses in medical health records (area under the curve = 0.69 and 0.64, respectively), and other mental and physical health diagnoses. The AUC values for these prediction studies are comparable to those found for imaging methods used to classify PTSD (e.g., Ramos-Lima et al., 2020) and are superior to performance of other language-based models (Eichstaedt et al., 2018, 2020).

The present study tests ALPs of depressivity, anxiousness, anger proneness, stress, and personality for predicting trauma-related outcomes in responders to the World Trade Center (WTC) disaster. Despite 20 years passed since September 11th, 2001, responders continue to show high rates of both psychiatric and medical sequelae of trauma (Bromet et al., 2016; Wisnivesky et al., 2011). Most participants in the present sample suffer from chronic symptoms and two years is not sufficient length of time to observe substantial change (Waszczuk et al., 2018). Hence, we did not seek to predict change, but thought it important to measure outcomes at a different time point from predictors to avoid transient methodological confounds (e.g., state effects, response biases) and allow for a clear temporal sequence between predictors and outcomes. We build on the work of Schwartz, Park, Eichstaedt, and colleagues to develop personality and mental health ALPs (Park et al., 2015; Schwartz et al., 2014) and we extend an initial smaller scale study applying ALPs to oral histories of a different sample of WTC responders (Son et al., in press). Son et al. (2020) predicted only PTSD symptoms with four psychiatric ALPs in a smaller sample of 75 responders. We present novel analyses of nine psychiatric ALPs and their longitudinal prediction of a wider array of mental health symptoms (including depression and sleep disturbance, in addition to PTSD symptoms), *and* physical health symptoms (lower respiratory and GERD symptoms) in a larger sample of 174 responders, as well as test their incremental validity over self-report measures of corresponding constructs. Based on prior research showing that post-disaster stress and personality vulnerabilities such as neuroticism are significant risk factors for poor long-term mental and physical health, we expected that the ALPs would account for significant portion of variance in the two-year trauma-related outcomes (DiGangi et al., 2013; Zvolensky et al., 2015).

1. Method

1.1. Procedure

Data were collected as part of the longitudinal WTC Personality and Health Study, which began in 2017 (Waszczuk et al., 2019). Participants were recruited from the Stony Brook site of WTC Health Program (Dasaro et al., 2017), established by the Center for Diseases Control to monitor the medical and psychiatric health of responders to the WTC disaster. To qualify for the program, responders were required to have been on the site of the disaster and/or spent significant time in clean-up efforts. Patients were recruited following an annual health monitoring visit to the program. To obtain a sample representative of the program, the only exclusion was inability to complete study procedures due to either limited comprehension of English language or major cognitive impairment.

At the baseline and two-year follow-up assessments, participants completed questionnaires in the laboratory. Moreover, for two weeks directly following the baseline assessment, participants completed daily surveys and voicemails. Collection of voicemails began when enrollment into the parent study was half completed, resulting in a smaller but random subsample. The study was approved by the local Institutional Review Board and all participants provided informed consent.

1.2. Participants

WTC responders ($N = 211$) participated in the study. To ensure an adequate sample of language per participant, participants with fewer than 200 words total in their voicemails were excluded (Kern et al., 2016), leaving $n = 174$ responders. A majority ($n = 148$) completed the follow-up assessment two years later. Participants were 55.4 years old on average ($SD = 8.7$ years), 89% male, and 90.8% white (6.9% Black, 1.7% Asian, and 0.6% other). Six percent identified as Hispanic ethnicity. The majority of participants worked in law enforcement on 9/11 (65%), while the other responders were primarily construction workers, electricians, and paramedics. The responders were drawn from the WTC Personality and Health (P&H) Study ($N = 450$), who were recruited from the WTC Health Program Stony Brook Site ($N = >10{,}000$). The present sample was demographically similar both to the parent study and overall patient population in our clinic (92% male and 79% white) (Bromet et al., 2016).

1.3. Measures

Baseline ALPs. Over a two-week period, participants left voicemails answering the prompted questions, "What was the worst part of your day?" and "What was the best part of your day?" Participants were also asked, "how did you respond?" to each experience. The average number of voicemails was 10.47 ($SD = 3.65$ voicemails, range 2–16). Across all days, participants said 1015 words on average ($SD = 678$).

ALPs used in this work build on those selected by Son and colleagues (in press) for analyses of oral history interviews of a different group of WTC responders. The process consisted of three steps: transcription, conversion to linguistic features, and application of AI models. Audio

was transcribed using *TranscribeMe*, a HIPAA-approved transcription service. Features were extracted using Differential Language Analysis Toolkit (DLATK) (Schwartz et al., 2017), consisting of relative frequencies and binary indicators of words and phrases (1- to 3-word sequences), as well as topic prevalence. Topics were derived through latent Dirichlet allocation (Blei et al., 2003), consist of semantically-related words, and provide values indicating topic frequency. A standard set of 2000 topics was used, introduced by Schwartz et al. (2013) and also applied by Eichstaedt et al. (2020). Linguistic features were then mapped to pre-trained, social-media-derived algorithms for depressivity, anxiousness, anger proneness, perceived stress, and the domains of the Five-Factor Model of personality (Park et al., 2015; Schwartz et al., 2014). Descriptive statistics are provided in Table 1.[1] Park et al. (2015) and Schwartz et al. (2014) provide word clouds that illustrate top words and phrases contributing to the ALPs. Not all features are clearly face-valid representations of the constructs, but many are. For example, the words "party" and "excited" are among top contributors to the extraversion ALP (Park et al., 2015) and the words "lonely" and "depressed" are top contributors to the depressivity ALP (Schwartz et al., 2014).

Baseline Self-Report Predictors. Self-report predictors were selected to match the constructs captured by the ALPs. Participants completed two personality inventories. The Faceted Inventory of the Five-Factor Model (FI-FFM) (Watson et al., 2019) was used to assess neuroticism, extraversion, conscientiousness, and agreeableness. From within the neuroticism domain, we included the personality traits Depression, Anger Proneness, and Anxiety. We refer to these constructs as depressivity, anger proneness, and anxiousness from here forward to avoid confusion with symptom outcomes. The Big Five Inventory-2 (Soto and John, 2017) was used to assess openness. Items are rated on a Likert-type scale from 1 (disagree strongly) to 5 (agree strongly).

Perceived stress was assessed every evening for two weeks using three items drawn from the Perceived Stress Scale (PSS) (Cohen et al., 1983), the most widely used instrument measuring the perception of stress. An example item is, "Today, I felt I was unable to control important things in my life." The three items were adapted for daily diary and rated on a 5-point Likert scale (i.e., 1-none at all to 5-extremely). Scores were averaged across the two weeks of surveys.

Two-Year Outcomes. Outcomes were assessed at the two-year follow-up using the following inventories: PTSD symptoms in the past month were assessed with the PTSD Checklist for DSM-5 (PCL-5) (Weathers et al., 2013), a reliable and widely-used measure of PTSD severity. It consists of 20 items rated in reference to WTC events as 1 (not at all) to 5 (extremely).

The Inventory of Depression and Anxiety Symptoms, expanded version (IDAS-II) was used to assess General Depression (i.e., clinical depression, as opposed to personality trait depressivity; 20 items), Suicidality (6 items), and Well-Being (8 items), and has shown strong evidence of reliability and validity (Watson et al., 2012). The IDAS items are rated for the past 2 weeks on a 5-point Likert scale from 1 (not at all) to 5 (extremely).

Lower Respiratory Symptoms (LRS) in the past week were assessed with the LRS questionnaire that demonstrated excellent reliability and validity in WTC population (Waszczuk et al., 2017, 2019). It consists of six items (e.g., "How often did your chest feel tight?") rated on a scale from 1 (e.g., none) to 5 (e.g., 6–7 days). GERD symptoms in the past week were assessed with the Reflux Disease Questionnaire (RDQ) (Shaw et al., 2001). This version included six items (e.g., "a pain in the center of the upper stomach") that rated symptom severity from 1 (did not have) to 6 (severe).

Over the two-week period after the follow-up visit, responders completed daily diaries, which included sleep quality assessed in the morning with a questionnaire based on the Pittsburgh Assessment Conference consensus sleep diary (Natale et al., 2015). Participants also rated the quality of sleep from 1 (very poor) to 5 (very good).

Table 1
Descriptive statistics for study variables.

Scales	*N*	Min	Max	Mean	*SD*
Baseline ALP					
Neuroticism	174	−0.57	0.34	−0.15	0.16
Extraversion	174	−0.46	0.35	−0.06	0.15
Openness	174	−0.31	0.60	0.11	0.15
Agreeableness	174	−0.57	0.58	0.03	0.14
Conscientiousness	174	−0.41	0.51	−0.05	0.14
Depressivity	174	2.30	3.34	2.74	0.20
Anger Proneness	174	2.25	3.40	2.81	0.23
Anxiousness	174	2.50	3.72	3.10	0.23
Stress	174	2.26	3.35	2.84	0.17
Baseline Self-Report Predictors					
Neuroticism	172	1.00	4.07	2.32	0.72
Extraversion	172	1.43	4.75	3.39	0.60
Openness	173	1.92	5.00	3.83	0.61
Agreeableness	172	2.40	4.81	3.73	0.48
Conscientiousness	172	2.48	4.93	3.92	0.53
Depressivity	172	1.00	4.60	2.09	0.88
Anger Proneness	171	1.00	5.00	2.28	0.87
Anxiousness	172	1.00	4.20	2.59	0.78
Daily Stress	171	3.21	13.25	6.11	1.63
Two-Year Self-Report Outcomes					
General Depression	148	23.00	73.00	37.11	11.39
Suicidality	148	6.00	12.00	6.68	1.14
Well-Being	148	8.00	40.00	23.30	6.90
PCL	147	20.00	72.00	32.03	12.37
LRS	148	6.00	24.00	9.79	4.23
GERD	148	6.00	30.00	9.97	5.46
Daily Sleep Quality	109	1.35	5.00	3.40	0.72

Note. PCL = PTSD symptoms, LRS = lower respiratory symptoms, GERD = gastroesophageal reflux disease symptoms. ALPs are expressed in arbitrary units and are scaled based on performance of AI models adapted from social media. Unfortunately, this precludes meaningful interpretation of absolute values, but they are presented for archival purposes.

[1] ALPs are expressed in arbitrary units and are scaled based on performance of AI models adapted from social media. Unfortunately, this precludes meaningful interpretation of absolute values, and hence the present paper focused on rank-order effects, but we include ALPs into Table 1 for archival purposes.

1.4. Analyses

Missing data on the self-report questionnaires were imputed with ipsative mean imputation if less than 20% of a respective questionnaire's data was missing. Correlations were used to examine bivariate relationships among the variables. Multiple regressions were completed with each outcome as the DV and each ALP and self-report construct as the IVs. For example, one multiple regression model included ALP stress and self-report stress as independent variables predicting two-year follow-up self-report depression as the dependent variable. Thus, as there were 9 pairs of predictors (i.e., ALP and corresponding self-report scale) and seven outcomes, there were 63 multiple regression analyses. Alpha was set at $p < .01$ to balance Type 1 errors. The data that support the findings of this study are available from the corresponding author upon reasonable request.

2. Results

The median absolute value intercorrelation among the ALPs was $r = 0.23$ and $r = 0.44$ among outcomes (see Supplemental Tables S1 and S2 for individual intercorrelations). The median absolute value intercorrelation among the self-report predictors was $r = 0.49$. This indicates that ALPs were distinct from each other, even more so than self-reports and the outcomes.

Correlations between the ALPs and self-report variables are presented in Table 2. ALPs converged significantly with the corresponding

Table 2
Intercorrelations among predictor variables.

ALP	Self-Report Predictors								
	1	2	3	4	5	6	7	8	9
1. Neuroticism	**.28**	**.30**	.10	**.32**	**-.31**	-.08	-.06	**-.38**	**.33**
2. Depressivity	**.31**	**.39**	.11	**.29**	**-.28**	-.07	-.07	**-.40**	**.48**
3. Anger Proneness	**.21**	**.27**	.05	**.22**	**-.17**	-.02	-.06	**-.23**	**.30**
4. Anxiousness	**.16**	**.24**	.00	**.18**	-.14	-.03	-.02	**-.23**	**.30**
5. Extraversion	**-.18**	**-.22**	-.08	**-.16**	**.26**	.12	.00	.13	**-.20**
6. Openness	.11	**.16**	.00	.12	-.14	.08	.04	-.08	**.19**
7. Agreeableness	**-.18**	**-.22**	-.10	-.14	**.18**	.05	**.15**	.14	**-.28**
8. Conscientiousness	**-.30**	**-.38**	**-.16**	**-.22**	**.25**	.05	**.17**	**.31**	**-.22**
9. Stress	**.29**	**.34**	.12	**.29**	**-.20**	-.05	-.10	**-.43**	**.40**

Note. Bold indicates correlations significant at $p < .05$. Gray shade = expected convergent association.

self-reports, except for openness and anger proneness. Convergence was particularly high ($r > 0.30$) for depressivity, conscientiousness, and stress. ALP anger proneness and openness showed little convergence with their corresponding self-reports, but ALP anger proneness did show significant relationships with other constructs (e.g., ALP anger proneness with self-report stress, and ALP openness with self-report dysphoria). These may reflect limitations of self-reports, as openness and anger proneness scales did not correlate with any ALPs, except for one weak association with ALP conscientiousness. In terms of discriminant validity, most ALPs correlated with several other self-report variables in addition to their corresponding self-report, indicating that the ALPs tended to associate with a general domain rather than a specific scale.

The ALPs significantly predicted all two-year trauma-related outcomes (Fig. 1). Individually, ALP depressivity and ALP stress were significantly correlated with all outcomes. ALP depressivity predicted PTSD and depression severity most strongly, and ALP stress predicted LRS and depression the most. ALP agreeableness, ALP neuroticism, ALP anger proneness, and ALP anxiousness related to somewhat fewer outcomes, but also showed highest prediction of depression. ALP conscientiousness and ALP openness were most predictive of well-being. ALP extraversion was protective against future PTSD symptoms.

To determine whether ALPs convey prognostic information not captured already by self-report assessments of the same constructs, they were entered in pairs as predictors for each outcome in turn. The regression models in which an ALP was statistically significant at $p < .01$ are presented in Table 3. Seven of the 9 ALPs showed evidence of predictive power over and above their corresponding self-report measures. ALP anger proneness uniquely predicted four outcomes: Depression, PTSD, LRS, and GERD symptoms. ALP openness contributed to prediction of depression, well-being, PTSD, and GERD symptoms. ALP neuroticism uniquely predicted three outcomes: depression, well-being, and PTSD symptoms. ALPs for agreeableness, depressivity, and stress also showed unique predictive information from their corresponding self-report scales. The effect sizes for unique predictive effects were moderate across all outcomes. When controlling for the respective trauma-related outcome at baseline in all analyses, two ALPs remained statistically significant at $p < .001$: ALP anger still predicted depression, ALP openness still predicted GERD symptoms, indicating that these ALPs unexpectedly predicted increases in depression and GERD symptoms, despite the short time span relative to time since trauma.

3. Discussion

Artificial intelligence assessments of risk factors are becoming increasingly more refined and accessible, but little evidence is available regarding their validity for clinical populations, especially trauma survivors. As evidence of clinical and prognostic utility accumulates, this technology could improve standard psychiatric assessment with relatively low cost and effort for both patients and clinicians. The present study applied machine learning-derived algorithms developed in social media language to create ALPs from voicemails of WTC responders. Results support the translational value of these ALPs in a primary care setting, especially with regard to prognosis for trauma-related outcomes.

We found that ALPs can predict diverse trauma-related mental health outcomes (e.g., PTSD symptoms, depression, suicidality, and low well-being) and trauma-related physical health (e.g., LRS, GERD symptoms, and sleep disturbance). The ALPs significantly correlated with all two-year outcomes and each ALP showed significant effects. Predictive effects for conscientiousness, neuroticism, depressivity, anger proneness, and perceived stress ALPs reached correlations of .30, and the largest was 0.40. These effects are particularly impressive because predictors and outcomes were in entirely different modalities. In contrast, the majority of prior disaster studies relied on self-report to assess both predictors and outcomes, which inflates effects due to common assessment method. Moreover, predictors were scored from a brief sample of natural language (mean of 1015 words or less than 7 min of speech), underscoring how quickly substantial predictive power can be acquired from language.

For PTSD, the strongest predictors were stress, depressivity, and

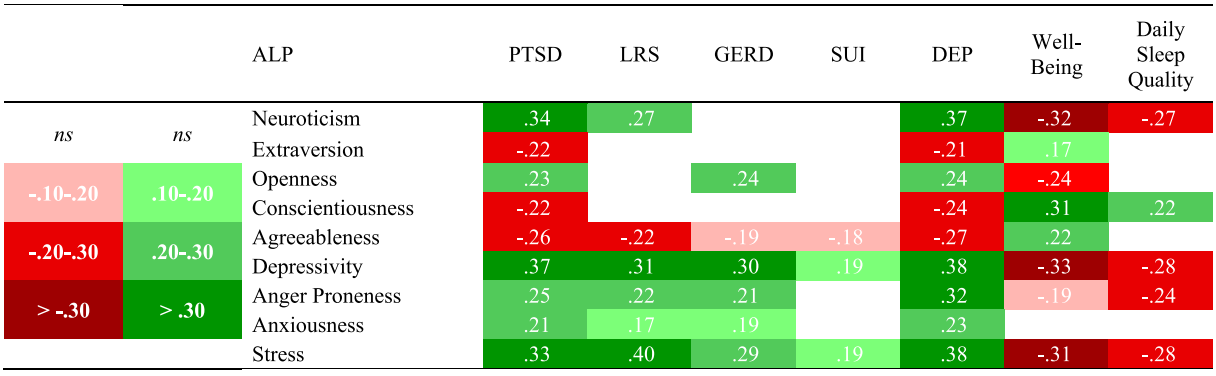

Fig. 1. Correlations between baseline ALPs and two-year outcomes.

Table 3
Regression of two-year outcomes on predictors.

Two-Year Follow-Up Outcome	Baseline Self-Report Predictor	β	*p*	Baseline ALP	β	*p*	R^2
Depression	Neuroticism	**.62**	.000	Neuroticism	**.20**	.002	.49
Well-Being	Neuroticism	**-.47**	.000	Neuroticism	**-.20**	.005	.31
PTSD	Neuroticism	**.60**	.000	Neuroticism	**.17**	.008	.44
Depression	Openness	-.12	.137	Openness	**.25**	.002	.07
Well-Being	Openness	**.23**	.004	Openness	**-.26**	.001	.11
PTSD	Openness	-.08	.354	Openness	**.23**	.005	.06
GERD	Openness	-.05	.506	Openness	**.24**	.003	.06
Depression	Agreeableness	**-.28**	.001	Agreeableness	**-.22**	.006	.14
PTSD	Agreeableness	**-.32**	.000	Agreeableness	**-.21**	.009	.17
LRS	Depressivity	**.28**	.001	Depressivity	**.22**	.009	.17
Depression	Anger Proneness	**.38**	.000	Anger Proneness	**.29**	.000	.24
PTSD	Anger Proneness	**.37**	.000	Anger Proneness	**.23**	.003	.20
LRS	Anger Proneness	.13	.107	Anger Proneness	**.22**	.008	.07
GERD	Anger Proneness	.07	.374	Anger Proneness	**.22**	.009	.05
LRS	Stress	**.32**	.000	Stress	**.27**	.001	.25

Note. Bold = significant at $p < .01$. ALP = artificial intelligence-based language predictor, PCL = PTSD Symptom Checklist, LRS = lower respiratory symptoms, GERD = gastro-esophageal reflux disease, DEP = depression, WB = well-being.

neuroticism ALPs, which is consistent with prior research consistently finding that these risk factors contribute prominently to PTSD (DiGangi et al., 2013). Anxiety and hostility are also established predictors of PTSD symptoms (DiGangi et al., 2013; Olatunji et al., 2010), and in the present study, we found significant moderate effects for both. ALP neuroticism predicted LRS, replicating a link between LRS and self-reported neuroticism observed in another sample of WTC responders (Waszczuk et al., 2018). Suicidality was predicted by depressivity and stress ALPs, which also replicates prior associations (Liu et al., 2006; R. D. Marshall et al., 2001). Sleep disturbance was predicted by neuroticism, conscientiousness, depressivity, anger proneness, and stress ALPs, and all have been reported previously in the literature (Morin and Jarrin, 2013). In sum, the associations uncovered in the present study between ALPs and trauma-related outcomes converge with those that have been found previously, bolstering support for their validity.

Importantly, ALPs showed significant levels of convergence with self-report assessments of the same constructs. Convergent validity evidence provides support that the ALPs capture meaningful variance in target constructs. Moreover, ALPs are more distinct than self-report assessments of the same constructs, as correlations among the ALPs had median $r = 0.23$ compared to $r = 0.49$ for self-reports. This may indicate a reduction in confounding due to limitations of self-report and improved precision due to the objective behavioral input (language) used to score the ALPs, but this possibility requires further investigation.

In particular, ALPs predicted certain differences in outcomes that self-reports could not. For instance, ALP anger proneness was associated with depression, PTSD, lower respiratory symptoms, and GERD symptoms over and above self-report anger proneness. This is consistent with prior research indicating that anger contributes to many poor health outcomes (Maan Diong et al., 2005) and is important and sometimes central feature in PTSD (Jakupcak et al., 2007; Olatunji et al., 2010). ALP anger proneness even predicted *increases* in depression over the relatively short two-year period over and above self-report anger proneness and depression symptoms at time 1. ALP neuroticism also showed incremental prediction over self-reports for mental health outcomes, consistent with the literature on the predictive power of neuroticism (Ormel et al., 2013). ALP depressivity and stress uniquely contributed to future LRS, consistent with prior literature on these risk factors in respiratory health (Kotov et al., 2015; Waszczuk et al., 2017, 2019). ALP agreeableness indicated lower risk of depression and PTSD symptoms, aligned with prior evidence of its protective role (Ozer and Benet-Martínez, 2006). The relationship of ALP openness to negative outcomes was inconsistent with prior literature, as openness has been shown to act as a buffer against symptom severity in PTSD (Caska and Renshaw, 2013; Knaevelsrud et al., 2010). Indeed, ALP and self-report openness had little in common, suggesting that scoring lower in ALP openness was more protective against health problems in the responders rather than higher. In the derivation study by Park and colleagues (Park et al., 2015), low openness ALP did display more positive words than high openness, and the openness ALP did correlate negatively with extraversion, agreeableness, and conscientiousness in that study, which is consistent with results in the present study.

Overall, though, the regression analyses indicate that important information about trauma-related outcomes can be obtained from ALPs that is not captured by self-reports. These results are especially impressive given that the ALPs were the only variables not measured by self-report. In the future, ALPs may be able to obtain information that self-reports cannot provide due to limitations with self-reports such as demand characteristics, over- and underreporting, acquiescence, or lack of insight. They may also be beneficial in complementarity with self-reports.

3.1. Limitations

The present study supports use of ALPs to predict trauma-related outcomes in a primary care sample. It used models developed in other samples, which provided a truly independent replication of past research and also extended it by applying these models to the novel task of predicting future outcomes. Nevertheless, it is limited in several respects. First, the sample was assessed many years after trauma exposure, and the ability of ALPs to predict risk immediately following traumatization remains to be tested. ALPs measured at baseline may have detected larger and more frequent long-term changes in trauma-related outcomes. Second, the present study relied on one modality to assess ALPs—voicemails. Alternative sources of language data should be evaluated in the future, especially language collected during routine clinical interviews to improve scalability of ALPs. Nevertheless, voicemails already offer a modality that is feasible to collect in clinical settings, unlike social media, which many patients may not engage in or may be unwilling to share with healthcare providers. Indeed, voicemails give patients control over information that they disclose to providers, and the present study shows that even a small sample of such language is very informative. Third, the present study had limited sample size, limiting power to detect small effects, which are still very consequential for psychological-physical health connections. It is possible that more effects would have been illuminated with a larger sample size. It was sufficient to detect bivariate effects, but too small to employ multivariate models with numerous predictors, an important direction for future research. Fourth, the sample was mostly white and male. Demographics of this sample are similar to the Stony Brook WTC responder population drawn from Long Island (Bromet et al., 2016). However, the present

findings may not generalize to other responder samples with greater sex and race/ethnicity diversity. More research is necessary to ensure findings generalize across race/ethnicity and gender. Further, more research is necessary to compare ALP utility across clinical and community samples, and perhaps use other methods to validate ALPs such as comparisons with thin slice ratings made by humans (Oltmanns et al., 2004; K. E. Sasso and Strunk, 2013) and different sources of language samples. Fifth, outcomes considered here were obtained by self-report. They included the most important concerns of WTC responders (e.g., cough, PTSD symptoms, insomnia), but future research should consider a broader range of outcomes, such as service utilization, neuropsychological functioning, and biomarkers. In sum, there are several areas for methodological improvement that could provide even more impressive results for the validity of ALPs.

4. Conclusions

Artificial intelligence can be used to improve prognosis, but has not been implemented in psychiatry practice. Such technology could reduce demands on the time of clinicians and patients and simultaneously increase predictive validity of clinical assessments. The present study found that ALPs derived from analyses of social media performed well when applied to voicemails and contributed substantially to prediction of trauma-related outcomes two years later. The effects were moderate and present a proof of concept at this stage. However, the predictive power of ALPs is expected to become stronger as machine learning models are fine-tuned and trained on increasing larger datasets. Meanwhile, these findings suggest that collection of natural language data, to which clinicians may have access, and scoring of ALPs, can be automated for potential integration into clinical care in the future. ALPs offer a promising avenue for clinical assessment and artificial intelligence based on natural language might be developed into a powerful prognostic tool.

Declaration of competing interest

This authors declare no conflicts of interest.

Acknowledgements

The authors wish to sincerely thank the WTC responders who contributed their time and effort to participating in this project.

Dr. Joshua R. Oltmanns, Stony Brook University, Department of Psychiatry and Behavioral Health; Dr. Andrew Schwartz, Stony Brook University, Department of Computer Science; Dr. Camilo Ruggero, University of North Texas, Department of Psychology; Dr. Youngseo Son, Stony Brook University, Department of Computer Science; Jiaju Miao, Stony Brook University, Department of Applied Math and Statistics; Dr. Monika Waszczuk, Rosalind Franklin University, Department of Psychology; Dr. Sean A. P. Clouston, Stony Brook University, Department of Public Health; Dr. Evelyn J. Bromet, Stony Brook University, Department of Psychiatry and Behavioral Health; Dr. Benjamin J. Luft, Stony Brook University School of Medicine; Dr. Roman Kotov, Stony Brook University, Department of Psychiatry and Behavioral Health.

This research was supported by the National Institute for Occupational Safety and Health under Award Number U01OH011321 (PI: Roman Kotov) and the National Institute on Alcohol Abuse and Alcoholism Award Number R01AA028032-01 (PI: H. Andrew Schwartz). The funding agencies had no role in the conduct of the study or preparation of the manuscript. The findings and conclusions in this article are those of the authors and do not represent the official positions of NIOSH or NIAA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpsychires.2021.09.015.

References

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bromet, E.J., Hobbs, M.J., Clouston, S.A.P., Gonzalez, A., Kotov, R., Luft, B.J., 2016. DSM-IV post-traumatic stress disorder among World Trade Center responders 11–13 years after the disaster of 11 September 2001 (9/11). Psychol. Med. 46 (4), 771–783. https://doi.org/10.1017/S0033291715002184.

Caska, C.M., Renshaw, K.D., 2013. Personality traits as moderators of the associations between deployment experiences and PTSD symptoms in OEF/OIF service members. Hist. Philos. Logic 26 (1), 36–51. https://doi.org/10.1080/10615806.2011.638053.

Chancellor, S., De Choudhury, M., 2020. Methods in predictive techniques for mental health status on social media: a critical review. Npj Digital Medicine 3 (1), 43. https://doi.org/10.1038/s41746-020-0233-7.

Cohen, S., Kamarck, T., Mermelstein, R., 1983. A global measure of perceived stress. J. Health Soc. Behav. 24 (4), 385. https://doi.org/10.2307/2136404.

Dasaro, C.R., Holden, W.L., Berman, K.D., Crane, M.A., Kaplan, J.R., Lucchini, R.G., Luft, B.J., Moline, J.M., Teitelbaum, S.L., Tirunagari, U.S., Udasin, I.G., Weiner, J.H., Zigrossi, P.A., Todd, A.C., 2017. Cohort profile: World trade center health program general responder cohort. Int. J. Epidemiol. 46 (2) https://doi.org/10.1093/ije/dyv099 e9–e9.

DiGangi, J.A., Gomez, D., Mendoza, L., Jason, L.A., Keys, C.B., Koenen, K.C., 2013. Pretrauma risk factors for posttraumatic stress disorder: a systematic review of the literature. Clin. Psychol. Rev. 33 (6), 728–744. https://doi.org/10.1016/j.cpr.2013.05.002.

Eichstaedt, J.C., Kern, M.L., Yaden, D.B., Schwartz, H.A., Giorgi, S., Park, G., Hagan, C., Tobolsky, V., Smith, L.K., Buffone, A., Iwry, J., Seligman, M., Ungar, L.H., 2020. Closed and open vocabulary approaches to text Analysis: a review. Quantitative Comparison, and Recommendations. https://doi.org/10.31234/osf.io/t52c6 [Preprint]. PsyArXiv.

Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A., 2018. Facebook language predicts depression in medical records. Proc. Natl. Acad. Sci. Unit. States Am. 115 (44), 11203–11208. https://doi.org/10.1073/pnas.1802331115.

Jakupcak, M., Conybeare, D., Phelps, L., Hunt, S., Holmes, H.A., Felker, B., Klevens, M., McFall, M.E., 2007. Anger, hostility, and aggression among Iraq and Afghanistan war veterans reporting PTSD and subthreshold PTSD. J. Trauma Stress 20 (6), 945–954. https://doi.org/10.1002/jts.20258.

Kern, M.L., Park, G., Eichstaedt, J.C., Schwartz, H.A., Sap, M., Smith, L.K., Ungar, L.H., 2016. Gaining insights from social media language: methodologies and challenges. Psychol. Methods 21 (4), 507–525. https://doi.org/10.1037/met0000091.

Kleim, B., Horn, A.B., Kraehenmann, R., Mehl, M.R., Ehlers, A., 2018. Early linguistic markers of trauma-specific processing predict post-trauma adjustment. Front. Psychiatr. 9, 1–7. https://doi.org/10.3389/fpsyt.2018.00645.

Knaevelsrud, C., Liedl, A., Maercker, A., 2010. Posttraumatic growth, optimism and openness as outcomes of a cognitive-behavioural intervention for posttraumatic stress reactions. J. Health Psychol. 15 (7), 1030–1038. https://doi.org/10.1177/1359105309360073.

Kotov, R., Bromet, E.J., Schechter, C., Broihier, J., Feder, A., Friedman-Jimenez, G., Gonzalez, A., Guerrera, K., Kaplan, J., Moline, J., Pietrzak, R.H., Reissman, D., Ruggero, C., Southwick, S.M., Udasin, I., Von Korff, M., Luft, B.J., 2015. Posttraumatic stress disorder and the risk of respiratory problems in World trade center responders: longitudinal test of a pathway. Psychosom. Med. 77 (4), 438–448. https://doi.org/10.1097/PSY.0000000000000179.

Lee, S.Y., Park, C.L., 2018. Trauma exposure, posttraumatic stress, and preventive health behaviours: a systematic review. Health Psychol. Rev. 12 (1), 75–109. https://doi.org/10.1080/17437199.2017.1373030.

Liu, K.Y., Chen, E.Y.H., Chan, C.L.W., Lee, D.T.S., Law, Y.W., Conwell, Y., Yip, P.S.F., 2006. Socio-economic and psychological correlates of suicidality among Hong Kong working-age adults: results from a population-based survey. Psychol. Med. 36 (12), 1759–1767. https://doi.org/10.1017/S0033291706009032.

Maan Diong, S., Bishop, G.D., Enkelmann, H.C., Tong, E.M.W., Why, Y.P., Ang, J.C.H., Khader, M., 2005. Anger, stress, coping, social support and health: modelling the relationships. Psychol. Health 20 (4), 467–495. https://doi.org/10.1080/08870440405123313339 60.

Marshall, K., Abate, A., Venta, A., 2020. Houston strong: linguistic markers of resilience after hurricane harvey. Journal of Traumatic Stress Disorders & Treatment 9 (2). https://doi.org/10.37532/jtsdt.2020.9(2).199.

Marshall, R.D., Olfson, M., Hellman, F., Blanco, C., Guardino, M., Struening, E.L., 2001. Comorbidity, impairment, and suicidality in subthreshold PTSD. Am. J. Psychiatr. 158 (9), 1467–1473. https://doi.org/10.1176/appi.ajp.158.9.1467.

Merchant, R.M., Asch, D.A., Crutchley, P., Ungar, L.H., Guntuku, S.C., Eichstaedt, J.C., Hill, S., Padrez, K., Smith, R.J., Schwartz, H.A., 2019. Evaluating the predictability of medical conditions from social media posts. PloS One 14 (6), 1–12. https://doi.org/10.1371/journal.pone.0215476.

Morin, C.M., Jarrin, D.C., 2013. Epidemiology of insomnia. Sleep Medicine Clinics 8 (3), 281–297. https://doi.org/10.1016/j.jsmc.2013.05.002.

Natale, V., Léger, D., Bayon, V., Erbacci, A., Tonetti, L., Fabbri, M., Martoni, M., 2015. The consensus sleep diary: quantitative criteria for primary insomnia diagnosis.

Psychosom. Med. 77 (4), 413–418. https://doi.org/10.1097/PSY.0000000000000177.
Olatunji, B.O., Ciesielski, B.G., Tolin, D.F., 2010. Fear and loathing: a meta-analytic review of the specificity of anger in PTSD. Behav. Ther. 41 (1), 93–105. https://doi.org/10.1016/j.beth.2009.01.004.
Oltmanns, T.F., Friedman, J.N.W., Fiedler, E.R., Turkheimer, E., 2004. Perceptions of people with personality disorders based on thin slices of behavior. J. Res. Pers. 38 (3), 216–229. https://doi.org/10.1016/S0092-6566(03)00066-7.
Ormel, J., Jeronimus, B.F., Kotov, R., Riese, H., Bos, E.H., Hankin, B., Rosmalen, J.G.M., Oldehinkel, A.J., 2013. Neuroticism and common mental disorders: meaning and utility of a complex relationship. Clin. Psychol. Rev. 33 (5), 686–697. https://doi.org/10.1016/j.cpr.2013.04.003.
Ozer, D.J., Benet-Martínez, V., 2006. Personality and the prediction of consequential outcomes. Annu. Rev. Psychol. 57 (1), 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127.
Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.P., 2015. Automatic personality assessment through social media language. J. Pers. Soc. Psychol. 108 (6), 934–952. https://doi.org/10.1037/pspp0000020.
Pennebaker, J.W., Booth, R.J., Francis, M.E., 2007. Linguistic Inquiry and Word Count (LIWC): LIWC 2007. LIWC.net.
Ramos-Lima, L.F., Waikamp, V., Antonelli-Salgado, T., Passos, I.C., Freitas, L.H.M., 2020. The use of machine learning techniques in trauma-related disorders: a systematic review. J. Psychiatr. Res. 121, 159–172. https://doi.org/10.1016/j.jpsychires.2019.12.001.
Sasso, K.E., Strunk, D.R., 2013. Thin slice ratings of client characteristics in intake assessments: predicting symptom change and dropout in cognitive therapy for depression. Behav. Res. Ther. 51 (8), 443–450. https://doi.org/10.1016/j.brat.2013.04.007.
Sasso, M.P., Giovanetti, A.K., Schied, A.L., Burke, H.H., Haeffel, G.J., 2019. #Sad: twitter content predicts changes in cognitive vulnerability and depressive symptoms. Cognit. Ther. Res. 43 (4), 657–665. https://doi.org/10.1007/s10608-019-10001-6.
Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H., 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. PloS One 8 (9), e73791. https://doi.org/10.1371/journal.pone.0073791.
Schwartz, H.A., Eichstaedt, J., Kern, M.L., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L., 2014. Towards assessing changes in degree of depression through Facebook. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: from Linguistic Signal to Clinical Reality, pp. 118–125. https://doi.org/10.3115/v1/W14-3214.
Schwartz, H.A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., Eichstaedt, J., 2017. DLATK: Differential Language Analysis ToolKit. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, vols. 55–60. https://doi.org/10.18653/v1/D17-2010.
Shaw, M.J., Talley, N.J., Beebe, T.J., Rockwood, T., Carlsson, R., Adlis, S., Fendrick, A. M., Jones, R., Dent, J., Bytzer, P., 2001. Initial validation of a diagnostic questionnaire for gastroesophageal reflux disease. Am. J. Gastroenterol. 96 (1), 52–57. https://doi.org/10.1111/j.1572-0241.2001.03451.x.
Son, Y., Clouston, S. A. P., Kotov, R., Eichstaedt, J. C., Bromet, E. J., Luft, B. J., & Schwartz, H. A. (in press). World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews. Psychol. Med..
Soto, C.J., John, O.P., 2017. The next Big Five Inventory (BFI-2): developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. J. Pers. Soc. Psychol. 113 (1), 117–143. https://doi.org/10.1037/pspp0000096.
Waszczuk, M.A., Li, K., Ruggero, C.J., Clouston, S.A.P., Luft, B.J., Kotov, R., 2018. Maladaptive personality traits and 10-year course of psychiatric and medical symptoms and functional impairment following trauma. Ann. Behav. Med. 52 (8), 697–712. https://doi.org/10.1093/abm/kax030.
Waszczuk, M.A., Li, X., Bromet, E.J., Gonzalez, A., Zvolensky, M.J., Ruggero, C., Luft, B. J., Kotov, R., 2017. Pathway from PTSD to respiratory health: longitudinal evidence from a psychosocial intervention. Health Psychol. 36 (5), 429–437. https://doi.org/10.1037/hea0000472.
Waszczuk, M.A., Ruggero, C., Li, K., Luft, B.J., Kotov, R., 2019. The role of modifiable health-related behaviors in the association between PTSD and respiratory illness. Behav. Res. Ther. 115, 64–72. https://doi.org/10.1016/j.brat.2018.10.018.
Watson, D., Nus, E., Wu, K.D., 2019. Development and validation of the faceted inventory of the five-factor model (FI-FFM). Assessment 26 (1), 17–44. https://doi.org/10.1177/1073191117711022.
Watson, D., O'Hara, M.W., Naragon-Gainey, K., Koffel, E., Chmielewski, M., Kotov, R., Stasik, S.M., Ruggero, C.J., 2012. Development and validation of new anxiety and bipolar symptom scales for an expanded version of the IDAS (the IDAS-II). Assessment 19 (4), 399–420. https://doi.org/10.1177/1073191112449857.
Weathers, F.W., Litz, B.T., Keane, T.M., Palmieri, P.A., Marx, B.P., Schnurr, P.P., 2013. The PTSD checklist for DSM-5 (PCL-5)—standard [measurement instrument]. http://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp.
Wisnivesky, J.P., Teitelbaum, S.L., Todd, A.C., Boffetta, P., Crane, M., Crowley, L., de la Hoz, R.E., Dellenbaugh, C., Harrison, D., Herbert, R., Kim, H., Jeon, Y., Kaplan, J., Katz, C., Levin, S., Luft, B., Markowitz, S., Moline, J.M., Ozbay, F., Landrigan, P.J., 2011. Persistence of multiple illnesses in World Trade Center rescue and recovery workers: a cohort study. Lancet 378 (9794), 888–897. https://doi.org/10.1016/S0140-6736(11)61180-X.
Zvolensky, M.J., Farris, S.G., Kotov, R., Schechter, C.B., Bromet, E., Gonzalez, A., Vujanovic, A., Pietrzak, R.H., Crane, M., Kaplan, J., Moline, J., Southwick, S.M., Feder, A., Udasin, I., Reissman, D.B., 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.