Published in final edited form as:
Psychol Med. 2023 February ; 53(3): 918–926. doi:10.1017/S0033291721002294.

World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews

Youngseo Son[1], **Sean A. P. Clouston**[2,3], **Roman Kotov**[4], **Johannes C. Eichstaedt**[5], **Evelyn J. Bromet**[4], **Benjamin J. Luft**[6], **H. Andrew Schwartz**[1]

[1]Department of Computer Science, Stony Brook University, New York, USA

[2]Program in Public Health, Stony Brook University, New York, USA

[3]Department of Family, Population and Preventive Medicine, Stony Brook University, New York, USA

[4]Department of Psychiatry, Stony Brook University, New York, USA

[5]Department of Psychology & Institute for Human-Centered A.I., Stanford University, Stanford, California, USA

[6]Department of Medicine, Stony Brook University, New York, USA

Abstract

Background.—Oral histories from 9/11 responders to the World Trade Center (WTC) attacks provide rich narratives about distress and resilience. Artificial Intelligence (AI) models promise to detect psychopathology in natural language, but they have been evaluated primarily in non-clinical settings using social media. This study sought to test the ability of AI-based language assessments to predict PTSD symptom trajectories among responders.

Methods.—Participants were 124 responders whose health was monitored at the Stony Brook WTC Health and Wellness Program who completed oral history interviews about their initial WTC experiences. PTSD symptom severity was measured longitudinally using the PTSD Checklist (PCL) for up to 7 years post-interview. AI-based indicators were computed for depression, anxiety, neuroticism, and extraversion along with dictionary-based measures of linguistic and interpersonal style. Linear regression and multilevel models estimated associations of AI indicators with concurrent and subsequent PTSD symptom severity (significance adjusted by false discovery rate).

Results.—Cross-sectionally, greater depressive language ($\beta = 0.32$; $p = 0.049$) and first-person singular usage ($\beta = 0.31$; $p = 0.049$) were associated with increased symptom severity. Longitudinally, anxious language predicted future worsening in PCL scores ($\beta = 0.30$; $p = 0.049$),

Author for correspondence: Youngseo Son, yson@cs.stonybrook.edu.

Conflict of interest. None.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291721002294.

whereas first-person plural usage ($\beta = -0.36$; $p = 0.014$) and longer words usage ($\beta = -0.35$; $p = 0.014$) predicted improvement.

Conclusions.—This is the first study to demonstrate the value of AI in understanding PTSD in a vulnerable population. Future studies should extend this application to other trauma exposures and to other demographic groups, especially under-represented minorities.

Keywords

9/11; depression; disaster responders; language-based assessments; oral history interviews; posttraumatic stress disorder; risk factors; trajectories; World Trade Center

Introduction

The 9/11 attacks on the World Trade Center (WTC) left thousands of casualties and drastically affected the lives of hundreds of thousands of New Yorkers and others nearby (Bergen, 2019). Many affected were those dedicating their lives to the safety of others – police, firefighters, emergency medical personnel, and other responders to the crisis. There has been a significant physical and mental burden of the events that day which has left many struggling with their health as they age (Durkin, 2018; Luft et al., 2012). Many responders suffer from PTSD which has been either worsening, staying the same, or gradually improving over time (Cukor et al., 2011; Neria et al., 2010).

Massive disasters, such as the WTC attacks, can affect a large number of people at the same time and usually occur within a relatively short period. Illuminating the risk and protective factors that reliably predict future reductions or increases in PTSD symptoms can lead to improved understanding, more accessible in-clinic guidance on patient’s well-being, and more immediate care for those involved in catastrophic events. Previous work has made major headway in establishing longitudinal associations of exposure severity, demographic characteristics, and job duties with health trajectories of WTC responders (Bromet et al., 2016; Cone et al., 2015; Pietrzak et al., 2014). However, additional approaches to risk assessment are needed to more rapidly and thoroughly differentiate those at greatest risk in situations where structured approaches to data collection are not possible.

Recently, Artificial Intelligence (AI)-based techniques have begun to show promise for quickly and accurately assessing mental health from human behavioral data, such as language use patterns. For example, from social media language, researchers have predicted those more prone to post-partum depression (De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016), those more likely to receive a clinical diagnosis of depression (Eichstaedt et al., 2018) or those appearing at greatest risk of suicide (Matero et al., 2019; Zirikly, Resnik, Uzuner, & Hollingshead, 2019). For PTSD in particular, although studies have yet to validate models in a clinical setting, past work has shown that AI-based language techniques can distinguish Twitter users that have publicly disclosed a diagnosis of the condition from random selections of users (e.g. Coppersmith, Dredze, & Harman, 2014; Preotiuc-Pietro, Sap, Schwartz, & Ungar, 2015; Reece et al. 2017).

AI-based language analyses are strong candidates to improve risk assessments in a clinical setting because they enable a much wider range of responses (like an interview) whereby a score can be objectively determined (e.g. like standardized assessment). Once an AI-based technique is created (i.e. it is 'pre-trained'), it will always yield the same and robust score for a given input. While these language-based assessments were studied with social media texts for PTSD based on self-disclosures (Coppersmith et al., 2014; Preotiuc-Pietro et al., 2015), few works have investigated how effective these approaches are with the language outside social media for predicting PTSD severity evaluated in the clinical settings, especially in a longitudinal study context for PTSD future trajectories. In all such cases, modern machine learning techniques are used to automatically extract and quantify patterns of language from hundreds to thousands of words per individual, which are then used to automatically produce a mental health or risk score. As compared to traditional questionnaire-based assessments, such approaches seem to suffer from fewer self-report biases (Youyou, Kosinski, & Stillwell, 2015) and generally leverage a larger amount of information per person (Kern et al., 2016). However, using such approaches in a clinical setting requires patients to share private information from social media pages, and requires that each participant has a substantial amount of data to share in the first place.

In this study, we present the first evaluation of AI-based mental health assessments from language (henceforth language-based assessments) to predict future PTSD symptom trajectories of patients monitored in a clinical setting. Rather than social media, we utilize transcripts of oral history interviews from responders to the 9/11 attacks. We first examine whether existing ('pre-trained') predictive models (most of which were trained on social media) produce assessments associated with PTSD symptoms scores close to the time of interview. We then compare these language assessments to other information available within a mental health clinical cohort (e.g. age, gender, occupation) to evaluate the additional benefit of the AI-based assessments. Lastly, we seek to quantify the predictive power of language-based indicators, in part to assess their potential suitability for informing personalized therapeutic approaches.

Methods

Participants

The sample was derived from Stony Brook University's WTC Health & Wellness program, funded by the Centers for Disease Control and Prevention, that provides ongoing monitoring of WTC responders. A total of $N = 124$ responders underwent an oral history interview and agreed to allow researchers to merge data from the transcript of the oral history with information in their health monitoring records. Hammock et al. (2019) provide an extensive summary of data collection methods. Briefly, oral history participants were primarily recruited *via* word of mouth and by flyers posted in the Stony Brook WTC Wellness Program.

Each interview lasted approximately 1 h. It covered the responders' memory of 9/11 attacks and disaster relief efforts, their work activities at the site, experiences and sensations over the days and weeks that followed, and how the WTC disaster ultimately impacted their lives since. Interviews were conducted by clinical staff with diverse healthcare backgrounds after a comprehensive orientation in conducting guided interviews and eliciting details relevant to the key topics to be covered. Responders were encouraged to discuss what was most important to them. Interviews were completed between 2010 and 2018.

In order to restrict our sample responders who were not new to the WTC Health Program, the analysis sample was restricted to participants who had at least one valid score on the PTSD Checklist (PCL; Blanchard, Jones-Alexander, Buckley, & Forneris, 1996) within 2 years of their interview, and at least one pre-interview PCL yielding an analysis sample of $N = 113$ responders. The few newer health program enrollees who were excluded from this study were qualitatively different, having only had just begun care (and potential PTSD treatment) at interview time. Furthermore, to study longitudinal trajectories post-interview, we focused on the subset of individuals with at least three post-interview mental health assessments at least 2 years after the interview ($N = 75$). The demographic characteristics of the study samples are listed in Table 1. The demographic ratio of gender and police remained similar (<4% difference) after we limited the sample to responders who met the criteria for our language analysis; 92% of the subset group were male and 49% were police; their mean age at interview was 53.

Ethics

This study was approved by the Stony Brook University Institutional Review Board. The participants provided written informed consent.

Language-based assessments

We automatically derived nine variables assessing the responders' language during the interviews: four AI-based assessments of psychological traits (expression of anxiety, depression, neuroticism, and extraversion), three lexicon-based assessments of language style (first-person singular pronouns, plural pronouns, and use of articles), and two meta variables describing counts of words and lengths of words. The process to get these variables consisted of three steps: text transcription, conversion of text to linguistic features, and application of AI-based models or *lexica*.

Audio of each interview was transcribed into text using *TranscribeMe*, a HIPAA-approved transcription service. Each time the responders spoke, transcribers labelled the time and the words mentioned. The text of each interview was converted into 'features' – quantitative values describing the content of the interview language – and then input into: (a) four AI-based assessments of psychological traits, (b) three lexicon-based assessments of language style, and (c) two meta-variable extractions describing counts of words and lengths of words. All analyses, described below, were performed using the Differential Language Analysis ToolKit (DLATK) (Schwartz et al., 2017).

Conversion into linguistic features

The models we used required up to three types of linguistic features: (1) relative frequencies of words and phrases, (2) binary indicators of words and phrases, and (3) topic prevalence scores. Words and phrases are sequences of 1–3 words in a row. Their relative frequency was recorded by *DLATK* by counting each word or phrase mentioned and dividing by the total number of words or phrases mentioned by the responder. The binary indicator for words and phrases simply indicated whether each word or phrase shows up (1) or not (0). The tokenizer built into the *DLATK* package was used to extract words per interview.

Topics are weighted groups of semantically-related words, often derived through a statistical process called latent Dirichlet allocation (Blei, Ng, & Jordan, 2003). Once derived, topics can be applied to textual data to scoring, ranging from 0 to 1, indicating how frequently each group of words was mentioned (Kern et al., 2016). We use a standard set of 2000 topics introduced by Schwartz et al. (2013a, 2013b), which has frequently been applied in the psychological domain including most recently in Eichstaedt et al. (2020). Once extracted, features were mapped to nine coarse-grained scores as described below and used for analyses herein.

AI-based psychological traits (4)

The AI-based assessments input linguistic features such as words, phrases, and topics, and map them to psychological constructs (Kern et al., 2016; Schwartz & Ungar, 2015). We focused on existing pre-trained models for constructs known to be related to our mental health outcomes: (1) neuroticism and (2) extraversion (Park et al., 2015; Schwartz et al., 2013b) – the two factors of the five-factor model known to relate negatively to depression and anxiety-related mental health conditions (Farmer et al., 2002; Jorm et al., 2000; Jylhä & Isometsä, 2006), as well as (3) degree of depression and (4) anxiousness (Schwartz et al., 2014) – subfacets of emotional stability which correspond to negative high arousal language (anxiousness) and negative low arousal language (depressive). These models were trained on large and diverse populations (approximately sample sizes of $N = 65\ 000$ for neuroticism and extraversion and $N = 29\ 000$ for degrees of depression and anxiousness). They utilize the linguistic features of previously mentioned words and phrases as well as topics as input and output continuous scores for each of the four constructs. They have been validated against standard questionnaire-based measures as well as convergent factors and external criteria under a range of situations (Kern et al., 2016; Matero et al., 2019; Park et al., 2015; Schwartz et al., 2014). However, the predictive validity of these models has yet to be assessed in clinical interview settings. Importantly, to guard against overfitting, no adjustments were made to the models, and thus this can be considered an evaluation of the models exactly as they were presented in their respective papers (Park et al., 2015; Schwartz et al., 2014).

Function word lexicon features (3)

We extracted word frequencies of terms in LIWC 2015 categories (Pennebaker, Boyd, Jordan, & Blackburn, 2015) and calculated categories for an interview with each responder. Due to the relatively low sample size, we focused on the function word categories that were most prevalent and then selected those that had a literature-suggested association with mental health:

- First-person singular: depressed, low status, personal, emotional, informal. Previously correlated positively with neuroticism, depression, and anxiety (Baddeley & Singer, 2008; Holtzman, 2017; Rude, Gortner, & Pennebaker, 2004) and negatively with life satisfaction (Schwartz et al., 2013a).

- First-person plural: high status, socially connected to group. Previously correlated negatively with depression and anxiety (Ramirez-Esparza, Chung, Kacewicz, & Pennebaker, 2008) and positively correlated with life satisfaction (Schwartz et al., 2013a) along with the cognition and psychological well-being variables of our interest (Tausczik & Pennebaker, 2010).
- Articles: use of concrete nouns, interest in objects and things (Tausczik & Pennebaker, 2010).

Language meta features (2)

- Average word length is known to be associated with higher cognitive (Khawaja, Chen, & Marcus, 2010), conceptual complexity (Lewis & Frank, 2016), education, and social class (Hartley, Pennebaker, & Fox, 2003; Tausczik & Pennebaker, 2010). PTSD is known to impair cognitive processing and impose a cognitive burden (e.g. through intrusive memories and thought suppression) (Nixon, Nehmy, & Seymour, 2007).
- Word counts: We also recorded total word counts, the number by which all lexica above were normalized. Given the interviews were all an hour long, this is a proxy for the rate of speech from each participant.

Mental health outcomes

The PTSD Symptom Checklist for DSM-IV PTSD (PCL) was used to assess PTSD severity in the past month (Bromet et al., 2016; Cone et al., 2015; Pietrzak et al., 2014). We chose the PCL closest to the interview date (all within 2 years) for concurrent analyses (average initial PCL score = 33.7; S.D. = 16.2). Following previous work which suggests that a fixed cutoff might not be optimally established for all cases (Andrykowski, Cordova, Studts, & Miller, 1998; Bovin et al., 2016), we focused on continuous values. Post-interview PCL scores were used to create trajectories as described under *trajectory prediction* below.

Statistical analysis

We used linear regression coefficient of the target explanatory variable (PCL score) as its correlation strength and multivariable adjustment for possible confounders (age, gender, occupation, and years after 9/11) to acquire the unique effects of language-based assessments. On average, the interviews were conducted 10.31 years (S.D. = 1.43) after the event. We controlled for days since 9/11 in the analyses. Since we explored many language assessments at once, we considered coefficients significant if their Benjamini–Hochburg adjusted p values were <0.05.

Concurrent evaluation

We processed the interviews of responders who had PTSD assessments three or more interviews after the closest dates to interviews and at least one assessment before the closest dates to interviews for the stable future trajectory modeling. For our cross-sectional correlation analysis linking language-based assessments with PTSD, we selected PCL scores of WTC responders as their cross-sectional PTSD symptom severity at the time of the interview (Interview PCL), and it is controlled for future PTSD trajectories as a baseline.

Trajectory prediction

For modeling the trajectory of PCL scores of each responder, we fit an ordinary least squares regression model with an intercept to the post-interview PCL scores as a function of time t:

$$\mathrm{PCL}_{it} = \beta_{0i} + \beta_{1i}t + \epsilon_{it}^{(t)} \tag{1}$$

where PCL scores were measured at (t) years after the interviews, then use the β_{1i} coefficient as a future PCL score trajectory of a responder (i). Then, for the person-level prediction over β_{1i} using the language-based assessments controlling the age, gender, occupation, and years between the interview and 9/11 of the responder i as following:

$$\beta_{1i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_6 x_{6i} + \epsilon_i^{(i)} \tag{2}$$

where x_1: language-based assessments, x_2: baseline PCL, $x_{3\ldots6}$: age, gender, occupation, years after 9/11 (all valuables standardized). Using equation (1) and (2), we use the following joint model:

$$\mathrm{PCL}_{it} = \beta_{0i} + (\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_5 x_{5i})t + \epsilon_{it}^{(t)} \tag{3}$$

and evaluate an effect size of each language-based assessment as its predictive power for future PCL trajectories of the responders (Fig. 1).

For the longitudinal trajectories post-interview, we focused on the subset of individuals with at least three post-interview PCL assessments occurring at least 2 years following the interview (N = 75). Sample demographics are reported in Table 1. Counting the interview, these criteria allowed the trajectories to be derived from at least four data points per participant, with the last assessment occurring on average (mean) 5.5 years (S.D. = 1.3) after the interview. By using this trajectory-based approach, all assessments available were used and aligned with their dates of administration.

Results

Most responders were male (90%) and half (48%) were police (see Table 1 for sample characteristics). Their median age at their interviews was 55 (53 for the longitudinal cohort). The median number of words across the interviews was 10 254.

Associations between language-based assessments and PTSD severity

Table 2 shows the linear regression analyses linking language-based assessments and PCL scores among responders around their interview dates. Higher PCL scores were significantly associated with language-based assessments consistent with anxious, depressive, and neuroticism. High scores were also associated with greater use of first-person singular and more total count of words in their interviews (r > 0.22). Conversely, higher scores were also associated with less extraversion language patterns, first-person plurals, and articles. Results remained unchanged after adjusting for age, gender, occupation, and years after 9/11 despite some effects from covariates (<0.07).

Trajectory analysis

Table 3 shows that language-based assessments of the oral histories significantly predicted responders' PCL trajectories during the follow-up period. First, we calculated linear regression coefficient effect sizes when we modeled PCL score trajectories with language features only (first column of Table 3). Then we add the control variables into the model (the second column). Although the general directions of correlations were the same both with and without controls, suppression effects of control variables increased the effect sizes for anxiety and first-person plural usage (Fig. 2).

Discussion

The goal of the present study was to examine whether AI-based language assessments developed in non-clinical contexts were reliably (1) associated with PTSD on self-reported questionnaires, and (2) able to predict the extent to which one's symptoms would get better or worse (trajectory) within a long-term clinical setting. The study found support for the view that language-based assessments could be reliably used in a clinical setting when processing naturalistic interviews: specifically, we found that language-based features were indicative of current functioning (supporting aim 1) and that language-based features could predict future PTSD symptom trajectories (supporting aim 2). This study, for the first time, suggested that AI assessments of interviews from a clinical sample not focused specifically on the topic of mental health could be used to identify features indicative of a person's current and future mental well-being.

Implications

There are three major implications from this work. First, AI-based assessments of interviews were associated with the assessment of mental health scores concurrently, supporting the first aim. Specifically, depressed language was associated with greater PTSD symptom severity, as is self-focused language. This corroborates the clinical conceptualization of PTSD as involving self-focused rumination that maintains PTSD symptoms over time (Michael, Halligan, Clark, & Ehlers, 2007).

Second, the use of more anxious language predicted increased PTSD symptoms in the future, even when adjusting for age, gender, occupation, and days since the 9/11 disaster. This suggests that while immediate PTSD severity is associated with low mood, a worsening of PTSD is determined by anxiety, rather than depression. These results may suggest that while immediate PTSD severity is reflected in affective experience, it may be the cognitive processes associated with anxiety (worry, rumination) that underlie future increases in PTSD symptoms. This dovetails with the accounts of PTSD that understand it to be maintained through rumination and worry (Michael et al., 2007).

Third, the use of more first-person plural pronouns ('we', 'us', 'our') predicted decreased PTSD symptoms in the future when adjusting for the confound variables. This supports research showing that social support is an important affordance that can buffer against and help alleviate the psychopathological load of a traumatic life event. Previous findings have suggested that processes associated with chronic sympathetic arousal (which include the chronic activation of the HPA-axis in states of hypervigilance) may be ‘buffered against’ by social interactions with kin and close others (e.g. McGowan, 2002).

Depressive language and current PTSD severity

Depressive language ($\beta = 0.32$; $p = 0.049$) and high usage of first-person singulars ($\beta = 0.31$; $p = 0.049$) were most highly correlated with high PCL scores even after accounting for age, gender, days since 9/11, and responder occupation. These associations were consistent with findings from prior studies showing an association of PTSD symptoms with increased risk of depression (Breslau, Davis, Peterson, & Schultz, 2000; Stander, Thomsen, & Highfill-McRoy, 2014). Similarly, high usage of first-person singular in messages is negatively correlated with life satisfaction (Schwartz et al., 2013a). Anxious and neurotic language patterns had strong positive correlations with PCL scores, which align with a previous study that identified avoidance and hyperarousal symptoms as frequently reported symptoms (Bromet et al., 2016). For the associations between personality traits and PTSD symptoms, previous studies found that low extraversion and high neuroticism are associated with an increased risk of PTSD (Breslau, Davis, & Andreski, 1995; Fauerbach, Lawrence, Schmidt, Munster, & Costa, 2000), and we observed the same patterns of our language-based extraversion and neuroticism with PTSD severity.

Predictors of PTSD symptom trajectories

We examined language-based assessments as a predictor of responders’ PCL trajectories after their interviews. Usage of first-person plurals and longer average word lengths were most highly correlated with improvement in PTSD in all cases, whether adjusting for baseline PCL-score and demographics or not. For other language-based assessments, coefficient effect sizes increased when we accounted for confounding due to the suppression effects mainly attributable to PCL scores, age at interview, and gender (see online Supplementary Table S1).

Furthermore, we analyzed potential mediation effects of marital status to address whether differences in the use of pronouns were merely reflecting marital status although previous work does not suggest such a relationship (Simmons, Gordon, & Chambless, 2015). Our results showed that these two types of language-based assessments predicted beyond marital status as their correlations remained statistically significant after adjusting for both controls and marital status (see online Supplementary Tables S2 and S3). This demonstrates that these linguistic markers capture an orientation toward the self and others over and above marital status.

Social support

In line with an extensive literature in psychology, we observed the use of ‘I’ *v.* ‘we’ pronouns to mark classes of psychological processes that determined adjustment to and recovery from trauma. Previous work has related higher use of first-person singular pronouns (‘I’-talk) self-focus (Carey et al., 2015); we found it correlated with high cross-sectional PTSD severity. On the other hand, we found high usage of first-person plural pronouns (‘we’) to be associated with a decrease of PTSD symptoms in the future. Self-focused thinking has been identified as a transdiagnostic factor of PTSD and depressive symptoms marking an often maladaptive preoccupation with the self and negative experience (Birrer & Michael, 2011; Ingram, 1990; Martin, 1985). The use of ‘I’ pronouns, in turn, has previously been found to be a dependable marker of self-focus in natural language (Carey et al., 2015; Watkins & Teasdale, 2001; Wegner & Giuliano, 1980). Beyond mere self-focus, depression and negative affectivity have also been robustly associated with higher use of first person singular pronouns (Holtzman, 2017; Rude et al., 2004) and PTSD (Miragoli, Camisasca, & Di Blasio, 2019); PTSD also with few ‘we’ pronouns (Papini et al., 2015). Our study showed further evidence for these patterns: greater use of ‘I’ pronouns positively correlated with severe cross-sectional PTSD symptoms, and high usage of ‘we’ pronouns predicted decreasing PTSD symptoms in the future.

Limitations

This was the first study to evaluate the relationship between automatic language-based assessments from interviews and PTSD symptoms of a trauma population, and there were several limitations. First, our sample covered a particular cohort of trauma survivors, those responding to the WTC disaster. WTC responders are predominantly male, and members of the monitoring population eligible to participate in this study were predominantly police officers. As such, this study relied on a sample that is similar to the rest of the WTC responder population (Bromet et al., 2016). Nevertheless, this may limit the generalizability of present findings to other occupations and demographic groups. Future research would need to investigate whether the results replicate to additional populations. Second, language-based assessment predicted future change in PTSD and suggested that cognitive and social risk processes may be involved, but mechanisms underpinning these predictive effects were not tested directly. Third, while our feature-based identification process was completed in a large database with ample capacity to train robust AI models, the present analysis had a relatively small sample size that could only be reliably used for application and was too small to retrain models for the current population. Future work in larger samples will be able to tailor AI-based assessments to specific populations and clinical questions substantially enhancing their predictive power.

Potential use in clinical care

Clinical evaluation of PTSD symptoms in trauma-exposed patients is time-consuming and burdensome. Moreover, primary care providers often lack expertise to complete this assessment. Our results show that natural language can provide clinically useful information both for the detection of PTSD and the prediction of future symptom escalation. These methods can be applied to routine clinical interviews completed by staff without mental health expertise. Although oral history interviews used in this project were lengthy, previous research has shown that interactions as brief as 5 min (e.g. 200 words) can be sufficient to obtain reliable AI-based assessments (Kern et al., 2016). These assessments would not replace a psychiatric evaluation, but can be useful for screening in primary care and as an aid to psychiatrists, picking-up on diagnostic and prognostic features in language that may be missed clinically. Specific language-based risk factors could inform treatment selection, such as low social support, and may suggest group therapy or peer support interventions, whereas maladaptive cognitive styles suggest cognitive behavioral therapy.

Conclusion

We found automated AI-based assessments utilizing the language of WTC responders in their oral history interviews predicted their PTSD symptoms in both cross-sectional and longitudinal trajectory analyses. The patterns and the correlations from these studies should be examined cautiously, and may require independent confirmations from other WTC cohorts and across different types of exposures before general applications for PTSD treatments. Still, the patterns of language-based assessments consistent with previous findings in other settings and their strong statistical correlations provided unique insights and explanations beyond commonly known confounds or risk factors such as age, gender, occupation, marital status, or even questionnaire-based depression measures, suggesting support for clinicians toward more precise decisions. More generally, language-based assessments that capture individual digital phenotypes and distinctive linguistic markers from transcripts of interviews are veiy useful for investigating underlying causes of PTSD and may play a critical role as a supplement for enhancing personalized preventive care (Hamburg & Collins, 2010) and more effective treatments for PTSD; they may even enable real-time screening or preventive measures with reduced costs and less therapist time for helping a large number of people exposed to large-scale traumatic events (e.g. natural disasters, WTC PTSD) similar to a previous online PTSD treatment (Lewis et al., 2017). Nevertheless, future studies with applying language-based assessment on larger samples will be required in order to more precisely validate their statistical significance and correlations, and even further studies into subphenotypes and more detailed categorizations of language-based assessments will lead to more diverse analysis with rich high-dimensional digital phenotypes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

The authors are extremely grateful to the WTC rescue and recovery workers, who gave of themselves so readily in response to the WTC attacks and agreed to participate in this ongoing research effort. We also thank the clinical staff of the World Trade Center Medical Monitoring and Treatment Programs for their dedication and the labor and community organizations for their continued support. Son and Schwartz were supported, in part, by NIH R01 AA028032-01 and, in part, by DARPA via grant #W911NF-20-1-0306 to Stony Brook University; the conclusions and opinions expressed are attributable only to the authors and should not be construed as those of DARPA or the U.S. Department of Defense. Clouston was supported, in part, by NIH R01 AG049953.

References

Andrykowski MA, Cordova MJ, Studts JL, & Miller TW (1998). Posttraumatic stress disorder after treatment for breast cancer: Prevalence of diagnosis and use of the PTSD Checklist – Civilian Version (PCL-C) as a screening instrument. Journal of Consulting and Clinical Psychology, 66(3), 586. [PubMed: 9642900]

Baddeley JL, & Singer JA (2008). Telling losses: Personality correlates and functions of bereavement narratives. Journal of Research in Personality, 42 (2), 421–438.

Bergen PL (2019). September 11 attacks. [Online; posted 10-September-2019].

Birrer E, & Michael T (2011). Rumination in PTSD as well as in traumatized and non-traumatized depressed patients: A cross-sectional clinical study. Behavioural and Cognitive Psychotherapy, 39(4), 381–397. [PubMed: 21457604]

Blanchard EB, Jones-Alexander J, Buckley TC, & Forneris CA (1996). Psychometric properties of the PTSD Checklist (PCL). Behaviour Research and Therapy, 34(8), 669–673. [PubMed: 8870294]

Blei DM, Ng AY, & Jordan MI (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993–1022.

Bovin MJ, Marx BP, Weathers FW, Gallagher MW, Rodriguez P, Schnurr PP, & Keane TM (2016). Psychometric properties of the PTSD checklist for diagnostic and statistical manual of mental disorders–fifth edition (PCL-5) in veterans. Psychological Assessment, 28(11), 1379. [PubMed: 26653052]

Breslau N, Davis GC, & Andreski P (1995). Risk factors for PTSD-related traumatic events: A prospective analysis. The American Journal of Psychiatry, 152(4), 529–535. [PubMed: 7694900]

Breslau N, Davis GC, Peterson EL, & Schultz LR (2000). A second look at comorbidity in victims of trauma: The posttraumatic stress disorder – major depression connection. Biological Psychiatry, 48(9), 902–909. [PubMed: 11074228]

Bromet E, Hobbs M, Clouston S, Gonzalez A, Kotov R, & Luft B (2016). DSM-IV post-traumatic stress disorder among World Trade Center responders 11–13 years after the disaster of 11 September 2001 (9/11). Psychological Medicine, 46(4), 771–783. [PubMed: 26603700]

Carey AL, Brucks MS, Küfner AC, Holtzman NS, Back MD, Donnellan MB, … Mehl MR (2015). Narcissism and the use of personal pronouns revisited. Journal of Personality and Social Psychology, 109(3), e1. [PubMed: 25822035]

Cone JE, Li J, Kornblith E, Gocheva V, Stellman SD, Shaikh A, … Bowler RM (2015). Chronic probable PTSD in police responders in the world trade center health registry ten to eleven years after 9/11. American Journal of Industrial Medicine, 58(5), 483–493. [PubMed: 25851164]

Coppersmith G, Dredze M, & Harman C (2014). Quantifying mental health signals in Twitter. In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, pp. 51–60.

Cukor J, Wyka K, Mello B, Olden M, Jayasinghe N, Roberts J, … Difede J (2011). The longitudinal course of PTSD among disaster workers deployed to the world trade center following the attacks of September 11th. Journal of Traumatic Stress, 24(5), 506–514. [PubMed: 22095774]

De Choudhury M, Kiciman E, Dredze M, Coppersmith G, & Kumar M (2016). Discovering shifts to suicidal ideation from mental health content in social media. In Proceedings of the 2016 CHI conference on human factors in computing systems, pp. 2098–2110.

Durkin E (2018). September 11: nearly 10 000 people affected by ‘cesspool of cancer’. [Online; posted 11-September-2018].

Eichstaedt JC, Kern ML, Yaden DB, Schwartz HA, Giorgi S, Park G, … Ungar LH (2020). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. Psychological Methods. The preprint of this article is available at https://psy-arxiv.com/t52c6/. The DOI of this preprint is 10.31234/osf.io/t52c6.

Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, … Schwartz HA (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences, 115(44), 11203–11208.

Farmer A, Redman K, Harris T, Mahmood A, Sadler S, Pickering A, & McGuffin P (2002). Neuroticism, extraversion, life events and depression: The Cardiff Depression Study. The British Journal of Psychiatry, 181(2), 118–122. [PubMed: 12151281]

Fauerbach JA, Lawrence JW, Schmidt CW Jr, Munster AM, & Costa PT Jr. (2000). Personality predictors of injury-related posttraumatic stress disorder. The Journal of Nervous and Mental Disease, 188(8), 510–517. [PubMed: 10972570]

Hamburg MA, & Collins FS (2010). The path to personalized medicine. New England Journal of Medicine, 363(4), 301–304. [PubMed: 20551152]

Hammock AC, Dreyer RE, Riaz M, Clouston SA, McGlone A, & Luft B (2019). Trauma and relationship strain: Oral histories with World Trade Center disaster responders. Qualitative Health Research, 29(12), 1751–1765. [PubMed: 30920915]

Hartley J, Pennebaker JW, & Fox C (2003). Abstracts, introductions and discussions: How far do they differ in style?. Scientometrics, 57(3), 389–398.

Holtzman NS (2017). A meta-analysis of correlations between depression and first person singular pronoun use. Journal of Research in Personality, 68, 63–68.

Ingram RE (1990). Self-focused attention in clinical disorders: Review and a conceptual model. Psychological Bulletin, 107(2), 156. [PubMed: 2181521]

Jorm AF, Christensen H, Henderson AS, Jacomb PA, Korten AE, & Rodgers B (2000). Predicting anxiety and depression from personality: Is there a synergistic effect of neuroticism and extraversion?. Journal of Abnormal Psychology, 109(1), 145. [PubMed: 10740946]

Jylhä P, & Isometsä E (2006). The relationship of neuroticism and extraversion to symptoms of anxiety and depression in the general population. Depression and Anxiety, 23(5), 281–289. [PubMed: 16688731]

Kern ML, Park G, Eichstaedt JC, Schwartz HA, Sap M, Smith LK, & Ungar LH (2016). Gaining insights from social media language: Methodologies and challenges. Psychological Methods, 21(4), 507. [PubMed: 27505683]

Khawaja MA, Chen F, & Marcus N (2010). Using language complexity to measure cognitive load for adaptive interaction design. In Proceedings of the 15th international conference on Intelligent user interfaces, pp. 333–336.

Lewis CE, Farewell D, Groves V, Kitchiner NJ, Roberts NP, Vick T, & Bisson JI (2017). Internet-based guided self-help for posttraumatic stress disorder (PTSD): Randomized controlled trial. Depression and Anxiety, 34 (6), 555–565. [PubMed: 28557299]

Lewis ML, & Frank MC (2016). The length of words reflects their conceptual complexity. Cognition, 153, 182–195. [PubMed: 27232162]

Luft B, Schechter C, Kotov R, Broihier J, Reissman D, Guerrera K, … Bromet E (2012). Exposure, probable PTSD and lower respiratory illness among world trade center rescue, recovery and clean-up workers. Psychological Medicine, 42(5), 1069–1079. [PubMed: 22459506]

Martin M (1985). Neuroticism as predisposition toward depression: A cognitive mechanism. Personality and Individual Differences, 6(3), 353–365.

Matero M, Idnani A, Son Y, Giorgi S, Vu H, Zamani M, … Schwartz HA (2019). Suicide risk assessment with multi-level dual-context language and bert. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 39–44.

McGowan S (2002). Mental representations in stressful situations: The calming and distressing effects of significant others. Journal of Experimental Social Psychology, 38(2), 152–161.

Michael T, Halligan SL, Clark DM, & Ehlers A (2007). Rumination in posttraumatic stress disorder. Depression and Anxiety, 24(5), 307–317. [PubMed: 17041914]

Miragoli S, Camisasca E, & Di Blasio P (2019). Investigating linguistic coherence relations in child sexual abuse: A comparison of PTSD and non-PTSD children. Heliyon, 5(2), e01163. [PubMed: 30828653]

Neria Y, Olfson M, Gameroff MJ, DiGrande L, Wickramaratne P, Gross R, … (2010). Long-term course of probable PTSD after the 9/11 attacks: A study in urban primary care. Journal of Traumatic Stress, 23(4), 474–482. [PubMed: 20690169]

Nixon RD, Nehmy T, & Seymour M (2007). The effect of cognitive load and hyperarousal on negative intrusive memories. Behaviour Research and Therapy, 45(11), 2652–2663. [PubMed: 17666185]

Papini S, Yoon P, Rubin M, Lopez-Castro T, & Hien DA (2015). Linguistic characteristics in a non-trauma-related narrative task are associated with PTSD diagnosis and symptom severity. Psychological Trauma: Theory, Research, Practice, and Policy, 7(3), 295. [PubMed: 25961121]

Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, Stillwell DJ, … Seligman ME (2015). Automatic personality assessment through social media language. Journal of Personality and Social Psychology, 108(6), 934. [PubMed: 25365036]

Pennebaker JW, Boyd RL, Jordan K, & Blackburn K (2015). The development and psychometric properties of LIWC 2015. Technical report.

Pietrzak RH, Feder A, Singh R, Schechter CB, Bromet EJ, Katz C, … (2014). Trajectories of PTSD risk and resilience in World Trade Center responders: An 8-year prospective cohort study. Psychological Medicine, 44(1), 205–219. [PubMed: 23551932]

Preotiuc-Pietro D, Sap M, Schwartz HA, & Ungar LH (2015). Mental illness detection at the world well-being project for the CLPsych 2015 Shared Task. In Proceedings of the second workshop on computational linguistics and clinical psychology, pp. 40–45.

Ramirez-Esparza N, Chung CK, Kacewicz E, & Pennebaker JW (2008). The psychology of word use in depression forums in English and in Spanish: Texting two text analytic approaches. In ICWSM.

Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, & Langer EJ (2017). Forecasting the onset and course of mental illness with Twitter data. Scientific Reports, 7(1), 1–11. [PubMed: 28127051]

Rude S, Gortner E-M, & Pennebaker J (2004). Language use of depressed and depression-vulnerable college students. Cognition & Emotion, 18(8), 1121–1133.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Lucas RE, Agrawal M, … Ungar L (2013a). Characterizing geographic variation in well-being using tweets. In Seventh International AAAI Conference on Weblogs and Social Media.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, … (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE, 8(9), e73791. [PubMed: 24086296]

Schwartz HA, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, … 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.

Schwartz HA, 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, pp. 55–60.

Schwartz HA, & Ungar LH (2015). Data-driven content analysis of social media: A systematic overview of automated methods. The ANNALS of the American Academy of Political and Social Science, 659, 78–94.

Simmons RA, Gordon PC, & Chambless DL (2005). Pronouns in marital interaction: What do “you” and “I” say about marital health?. Psychological Science, 16(12), 932–936. [PubMed: 16313655]

Stander VA, Thomsen CJ, & Highfill-McRoy RM (2014). Etiology of depression comorbidity in combat-related PTSD: A review of the literature. Clinical Psychology Review, 34(2), 87–98. [PubMed: 24486520]

Tausczik YR, & Pennebaker JW (2010). The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54.

Watkins ED, & Teasdale JD (2001). Rumination and overgeneral memory in depression: Effects of self-focus and analytic thinking. Journal of Abnormal Psychology, 110(2), 353. [PubMed: 11358029]

Wegner DM, & Giuliano T (1980). Arousal-induced attention to self. Journal of Personality and Social Psychology, 38(5), 719.

Youyou W, Kosinski M, & Stillwell D (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036–1040.

Zirikly A, Resnik P, Uzuner O, & Hollingshead K (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 24–33.

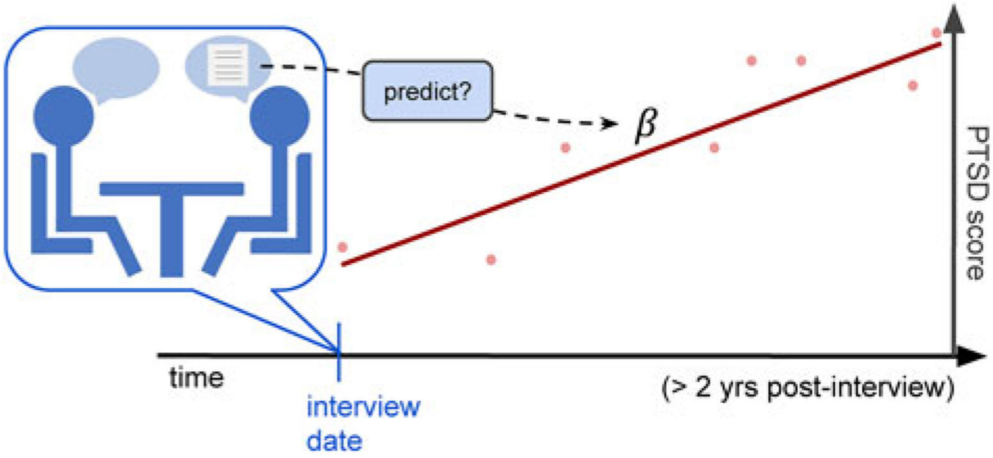

Fig. 1.

Evaluation setup for trajectory prediction. According to equation 3, we can then model the control-adjusted trajectory per user as $B_{1-\mathrm{cntrl},\,i} = (\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_5 x_{5i})$. Then, we used the slope of the fitted line as the PCL trajectory of the corresponding subject. Our main outcome was correlations between this trajectory slope and the subject's language patterns. The figure illustrates our trajectory modeling; dots in the figure represent the PTSD scores at the health assessments after the oral history interview of a responder and the red line represents the PTSD future trajectory line which is correlated with his/her language assessment from the interview.

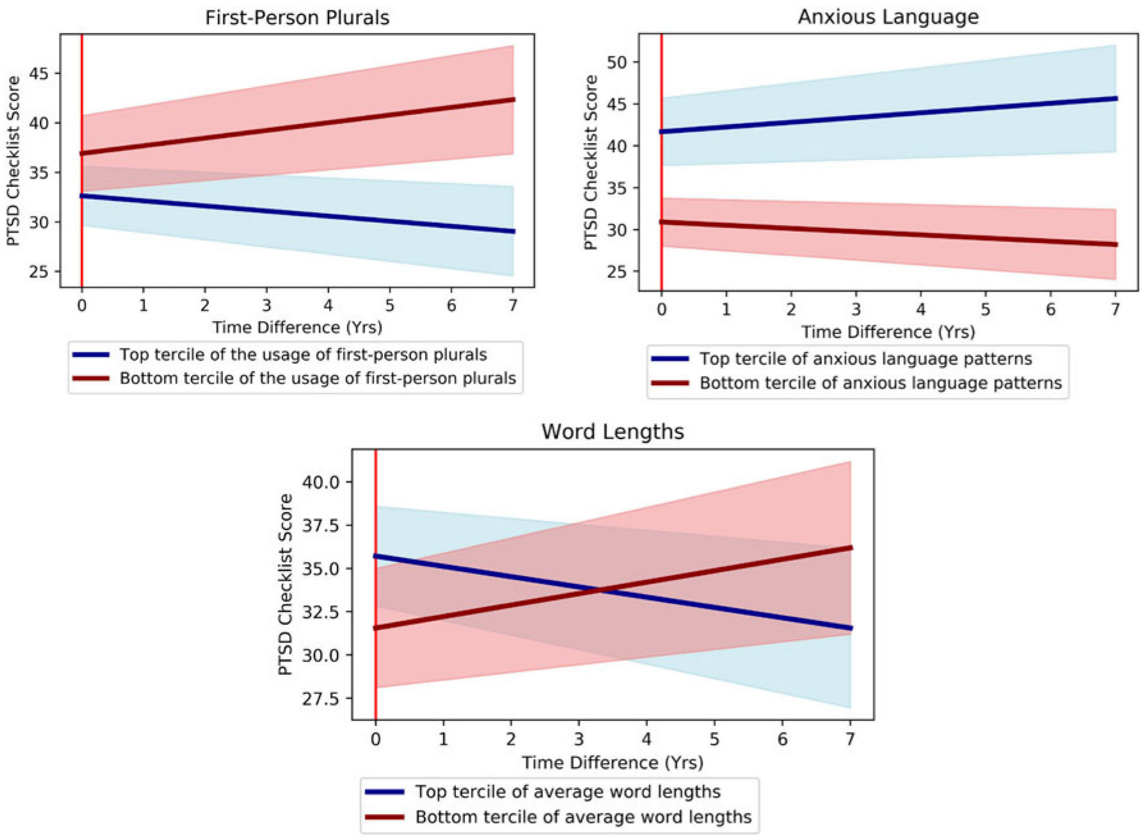

Fig. 2.

Average future PCL score trajectories of top (blue) and bottom (red) terciles of responders based on language-based assessments: word usages of first-person plurals (left), anxious language patterns (right), and average word lengths (bottom). All trajectories have been adjusted for interview (baseline) PCL scores, representing the residual after accounting for the expected trajectory at baseline. All differences are significant at $p < 0.05$ (see online Supplementary Table S1 for further analysis).

Table 1.

Data on subjects for health state correlation cross-sectional analysis and trajectory predictions

	N	Female (%)	Police (%)	Mean age at the interview (S.D.)	Median number of words
All participants	124	10	48	55.4 (9.8)	10 254
Meet inclusion criteria	75	8	49	53.4 (9.5)	9944

Table 2.

Cross-sectional association between language-based assessments and PCL PTSD Score

	PTSD symptoms	
Interview language features	***r*** **(direct correlation with symptom score)**	***β*** **(adjusted for age, gender, occupation, days since 9-11)**
Psychological traits		
Anxiety	0.26 (0.03–0.46)	0.20 (−0.03 to 0.41)
Depression	0.38 * (0.16–0.56)	0.32 * (0.10–0.51)
Neuroticism	0.32 * (0.10–0.51)	0.26 (0.04–0.46)
Extraversion	−0.10 (−0.32 to 0.13)	−0.15 (−0.37 to 0.08)
Linguistic style		
First-person singular	0.31 * (0.09–0.50)	0.31 * (0.09–0.51)
First-person plural	−0.05 (−0.28 to 0.18)	−0.10 (−0.32 to 0.13)
Articles	−0.09 (−0.31 to 0.14)	−0.06 (−0.28 to 0.17)
AVG word length	0.05 (−0.18 to 0.27)	0.03 (−0.20 to 0.25)
Word count	0.22 (−0.01 to 0.43)	0.20 (−0.02 to 0.41)

Associations are from ordinary least squares over standardized independent variable – the language-based assessment and the standardized dependent variable – PTSD Checklist scores (PCL scores). Without controls is equivalent to Pearson Product-Moment Correlation ($N = 75$). Square brackets indicate 95% confidence intervals. Controls included as covariates (right column) included age, gender, occupation, days between 9/11/01 and interview date.

* Indicates significant correlations (multi-test, Benjamini–Hochburg adjusted $p < 0.050$). Each row is color-coded separately, from red (negative correlations) to green (positive correlations); greyed values indicate non-significant.

Table 3.

Predicting PCL trajectories of the responders using language-based assessments

	PTSD symptoms future trajectories	
Interview language features	***r*** **(direct correlation with symptom slope)**	***β*** **(adjusted for age, gender, occupation, days since 9–11, Interview PCL)**
Psychological traits		
Anxiety	0.16 (−0.07 to 0.37)	0.30* (0.08–0.49)
Depression	−0.00 (−0.23 to 0.22)	0.16 (−0.07 to 0.37)
Neuroticism	0.07 (0.29 to −0.16)	0.20 (−0.03 to 0.40)
Extraversion	0.17 (−0.06 to 0.38)	0.18 (−0.05 to 0.39)
Linguistic style		
First-person singular	0.00 (−0.23 to 0.23)	0.13 (−0.10 to 0.35)
First-person plural	−0.36* (−0.54 to −0.14)	−0.36* (−0.54 to −0.15)
Articles	−0.16 (−0.37 to 0.07)	−0.23 (−0.43 to 0.00)
AVG word length	−0.36* (−0.54 to −0.14)	−0.35* (−0.53 to −0.13)
Word count	0.06 (−0.17 to 0.28)	0.14 (−0.09 to 0.36)

Associations are from ordinary least squares over standardized independent variable – the language-based assessment and the standardized dependent variable – PCL future trajectory. Without controls is equivalent to Pearson Product-Moment Correlation (*N* = 75) with controls: age, gender, occupation, days between 9/11/01 and interview date, and interview PCL score.