Pre-Print

Demonstrating High Validity of a New AI-Language Assessment of PTSD: A Sequential Evaluation with Model Pre-registration

Oscar Kjell, PhD*[1,2], Adithya V Ganesan, MSc[1], Ryan L. Boyd, PhD[1], Joshua Oltmanns, PhD[3], Alfredo Rivero, MSc[1], Scott Feltman, MSc[4], Melissa A. Carr, BA[5], Benjamin Luft, MD[5, 6], Roman Kotov, PhD[7], H. Andrew Schwartz, PhD[1]

[1] Department of Computer Science, Stony Brook University
[2] Department of Psychology, Lund University
[3] Department of Psychology, Southern Methodist University
[4] Department of Psychiatry, Stony Brook University
[5] World Trade Center Health and Wellness Program, State University of Stony Brook, Stony Brook, New York
[6] Division of Infectious Diseases, School of Medicine, State University of Stony Brook, Stony Brook, New York
[7]Department of Psychiatry and Behavioral Health, Stony Brook University

* Corresponding author: oscar.kjell@psy.lu.se

ABSTRACT

BACKGROUND: Modern Artificial Intelligence (AI) has shown promise in identifying psychopathology based on the language used by patients, providing a scalable method for obtaining relevant behavioral markers. However, no existing models for assessing posttraumatic stress disorder (PTSD) have successfully demonstrated out-of-sample replicability. We develop a language-based AI model for PTSD and rigorously evaluate replicability in a prospective sample.

METHODS: Participants from the *Stony Brook World Trade Center (WTC) Health and Wellness Program* described their lives in an automated interview during a clinical monitoring visit. The language was analysed using AI to assess PTSD CheckList (PCL) for total symptom severity score and four symptom subscales and validated against medical record PTSD diagnosis.

To yield realistic accuracy estimates in this cross-sectional study, we propose the *Sequential Evaluation with Model Pre-registration* design, consisting of an iterative, two-phase pre-registration paradigm. The first pre-registration specifies the data split, the model development, and the initial hypotheses. The second pre-registration specifies the exact pre-trained models, data cleaning procedures, and the refined hypotheses.

RESULTS: The data split included a development (*N*=1437) and a prospective (*N*=346) dataset. Within the prospective sample, the pre-registered models produced scores that significantly correlated with their targets: PCL total (*r*=.38, p-value<.001) and the four subscales (*r*=.28–.37, p-value<.001). The pre-registered model for PCL total showed a robust association with PTSD diagnosis (AUC=.76), significantly outperforming demographics (AUC=.61, p-value=.006), WTC attack exposures (AUC=.61, p-value=.007) and a validated depression language model (AUC=.60, p-value<.001).

CONSLUSIONS: We developed new AI-language assessments of PTSD symptom severity. Within a clinical setting and over prospectively collected participant data, the assessments replicated with high convergent validity with self-report and high external validity against diagnosis in medical records. Analyses of observable behavioral markers in automated clinical interview language can produce robust psychiatric assessments, overcoming limitations found in traditional assessments.

Keywords. posttraumatic stress disorder; depression; language-based assessments; oral history interviews; AI; machine learning

OSF including pre-registration 1 and 2

Post-traumatic stress disorder (PTSD) manifests in heterogeneous symptom severity patterns, which are not easily expressed by patients via the closed-ended response formats found in traditional, structured assessments (1,2). Diagnostic interviews and self-reports rely on subjective interpretations from either the clinician (interview) or the patient (self-report), leading to idiosyncratic biases (3–7). New assessment methods are needed to reduce these limitations and complement current methods.

Recent advances in AI-based language analyses have demonstrated promise for increasing the validity and scope of mental health assessment (3,8). However, for PTSD in particular, evaluations have typically been limited to either (a) models trained to assess other conditions (e.g., AI models for detecting depression, anxiety, and sentiment evaluated against PTSD symptom severity; 9–11) or (b) models trained for PTSD but evaluated against non-clinical outcomes (e.g., public self-disclosures of PTSD; 12–15). Moreover, many of these have been derived from social media language (e.g., 12–14) rather than data collected in clinical settings. The lack of comprehensive evaluations is underscored by recent work demonstrating that AI models can fail to perform when evaluated prospectively (16–18).

A central concern surrounding the validity of AI-based models for detecting psychopathology is that standard assessment evaluation practices do not directly translate directly to AI-based approaches. Whereas standard pre-registration practices are suitable for many research goals, they fall short in meeting the needs for developing and evaluating "robust models" for translational research. For example, pre-registration practices that require specifying exact data processing steps when larger and more complex datasets (e.g., open-ended language) needed for AI model training often require unexpected bug fixes, several preprocessing steps, and hyper-parameter tuning that only come to light during model development stages (19,20). Further, it is best practice in AI to evaluate models over held-out samples to control for overfit among these complex models (i.e., cross-validation; (21)), which is not typical in pre-registered studies. In short, AI model development often involves an *iterative refinement process* that does not fit well within the standard, *a priori* pre-registration paradigms that have been developed for the purpose of traditional hypothesis testing. In fact, commonly used pre-registration paradigms can be understood to work against research aims more commonly found in machine learning, namely, developing models with incredibly high accuracy.

Here, we develop and evaluate new AI clinical models for assessing the severity of PTSD and related symptom domains. We propose a new evaluation paradigm, *Sequential Evaluation with Model Pre-registration (SEMP)*, that aims to combine good scientific practices (i.e., pre-registration) with robust AI model development practices (i.e., bug fixes, hyper-parameter tuning, and out-of-sample test). SEMP calls for first developing the model and then *registering it* (i.e., literally the model code and weights). We first register the steps for training the models and then register the models before testing them on new prospective data (see Figure 1). By registering the models before testing, the approach mitigates the overestimation of accuracies for clinical prediction models by (1) preventing overfitting of hyper-parameters (whether purposeful or accidental); (2) mitigating the risk of test data leaks – when data from the test set inadvertently influences decisions in the training process, thereby yielding overestimated accuracy for new, unseen data; and (3) when using a prospective test set, more realistically testing the model being applied to a sample later in time as would happen in practice.

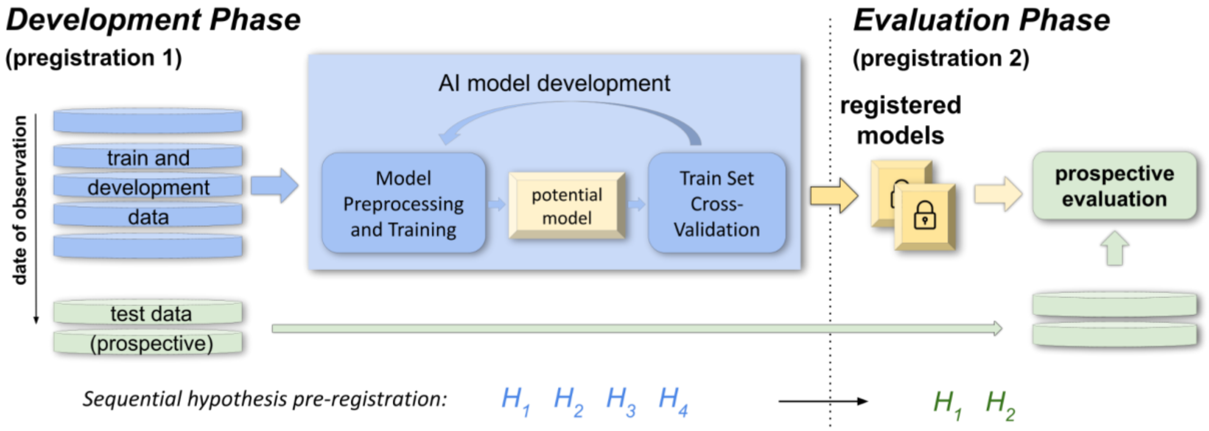


Figure 1. Overview of the Sequential Evaluation with Model Pre-registration (SEMP). *The stages include data split (i.e., prospective data), pre-registering the prediction model development (e.g., ridge regression), n-fold cross-validation (e.g., 10-folds), hypothesizes (H_{1-4}; where hypotheses can change across stages, as visualized with the different colors), registered prediction models (i.e., the exact models to validate), pre-processing (e.g., removal of parts where research assistants speak) and held-out perspective set evaluation (e.g., control variables).*

We apply SEMP when analyzing language from automated spoken interviews from a clinical visit to assess overall and domain-specific PTSD symptom severity over a cohort of 1783 World Trade Center emergency responders 20 years after the traumatic event. By using language-based assessments, responders are free to express the heterogeneous ways in which PTSD symptoms manifest during their everyday lives. The language is thus an open-ended behavioral, observable marker for assessing PTSD severity. We hypothesized that pre-registered AI-based models produce robust PTSD severity assessments, which are further evaluated against external criteria (i.e., medical records). To the best of our knowledge, this is the first study to prospectively evaluate AI models developed specifically for PTSD and its symptom domains.

METHOD

Design: Sequential Evaluation with Model Pre-registration

SEMP is a two-phase procedure to pre-register models before using held-out test-data. The *Development Phase* comprises developing the models; the *Evaluation Phase* involves a pre-registration locking in the exact models and preprocessing (data cleaning) code. The two-phase procedure enables iteratively developing the preprocessing and modeling choices (i.e., "hyperparameter optimization") without using the test-data such that final accuracies will be established more robustly on held-out, optionally prospective data. Instead of testing whether models produce assessment accuracies better than chance, the development phase enables one to register specific effect-size intervals for the evaluation phase.

Within the SEMP procedure, we first developed the AI-based PTSD symptoms models using automatically transcribed language from video-recorded automated clinical interviews over a development "training-set" sample. This was done for pre-registration phase 1 (the development phase) to produce and pre-register pre-trained models. In the second phase (the evaluation phase), we applied the pre-registered models to a "prospective held-out test" sample consisting of new participants. The base hypothesis for the evaluation phase is that the models trained during the development phase will continue to predict their intended outcomes on the unseen data. To strengthen hypotheses, we included expected correlational accuracy ranges:

> *The registered language-based assessments of PTSD produce scores that:*
>
> ***a)** are positively associated with PTSD symptom severity,*
>
> ***b)** achieve a correlation $r \geq .35$ for overall symptom severity and $r \geq .18$ for the four individual symptom domain dimensions (based on training-set cross-validated r=.41, 99% CI [.35, .46], N=1437 for the combined symptom severity and r=.25, 99% CI [.18, .31] for the individual symptom dimensions, N=1422), and*
>
> ***c)** are predictive above and beyond the pre-trained models using baseline demographics: age, gender, occupation, and race.*

In our case, we also had hypotheses beyond testing an AI model (i.e., those based on pre-existing language-based models) that also accompanied the SEMP process (depicted at the bottom of Figure 1 as "sequential hypothesis pre-registration"; see section 3 in the Supplement for these hypotheses).

Participants

Participants were recruited from the *Stony Brook WTC Health and Wellness Program*, where their health has been monitored over several years. Participant data was split into two non-temporally overlapping parts: The "development" data (2021-09-09 – 2022-07-29) and the held-out prospective "evaluation" data (2022-08-01 – 2022-09-30). The development data totaled 1437 participants (Female=7%, Male=93%; Mean age=57.9, SD=8.0 years; 14.5% with reported PTSD diagnosis in their medical record) – the prospective data includes an additional 346 participants (Female=9%, Male=91%; Mean age=58.5, SD=7.8 years; 15.6% with reported PTSD diagnosis in their medical record) enrolled prospectively relative to the participants in the development dataset (see Supplement section 1).

Measures and Material

Automated clinical interviews: Video-recorded answers about life

Participants were recorded while answering questions automatically shown on a screen in a private room during a clinical visit (i.e., an automated clinical interview). Questions probed respondents to describe positive (e.g., *What are the three things in your life that you look forward to the most right now?)* and negative aspects of their (past, present, and future) life in general (e.g., nicest and worst things, challenges, support network) and in relation to serious events (e.g., COVID-19 and 9/11; e.g., *How does 9/11 affect you now?*; see Supplement section 2 for all questions). To maximize the generalizability of content, the questions aimed at being broad using layman's terms (rather than, e.g., asking about specific clinical symptoms). The questions were presented on a screen with instructions on how not to read the questions out loud and to try spending at least 60 seconds answering each question. While the questions were updated and changed over three iterations of the development phase to increase engagement and more detailed answers, the questions were the same for everyone in the evaluation data test. The recording for those responding with at least 150 words (the pre-registered threshold) took, on average, 7.5 minutes (*SD*=4.1; range=1.1–43.0).

The PTSD CheckList

The PTSD CheckList (PCL; (22)) comprises 17 items assessing PTSD symptom severity based on DSM-IV. Respondents are asked to consider the previous months, answering using rating scales ranging from 1 (*not at all*) to 5 (*extremely*). We computed the total mean score and the four subscales (23,24), including Re-experiencing (e.g., *intrusive thoughts of trauma*), Avoidance (e.g., *avoiding thoughts of trauma*), Emotional Numbing (e.g., *inability to recall aspects of trauma*), and Hyperarousal (e.g., *sleep disturbance*). Cronbach alphas were acceptable for all scales in both datasets (≥ .70, see Supplement section 1).

PTSD Diagnosis in Medical Record

Diagnoses in the medical records are certifications that the participant has a WTC-related condition (i.e., a WTC-related PTSD diagnosis). The psychiatrists at the *Stony Brook WTC Health and Wellness Program* diagnosed based on clinical history in the medical records and the semi-structured *Diagnostic Interview Schedule* (see (25)).

WTC exposure

WTC exposure was assessed using a clinical interview at the initial monitoring visit (25). We use ten dichotomous (Yes/No) WTC exposure variables that were associated with increased risk of PTSD and other health outcomes in prior work (26–28; for more information, see supplement section 3).

Demographics

Self-reported age, gender, occupation, and race were collated from the monitoring data of the *Stony Brook WTC Health and Wellness Program*.

Procedure

The video recordings were collected in a clinical setting at the *Stony Brook WTC Health and Wellness Program*. All participants consented to participate and were informed about the study and their rights to withdraw at any time. A research assistant instructed the participants on how to conduct the automated interview. Last, the participants were debriefed.

The Institutional Review Board (IRB#604113) approved the study at Stony Brook University.

Statistical Analysis

The analyses were conducted using the *Differential Language Analysis Toolkit* (DLATK; 26)). Alpha was set at *P*<.05 with Benjamini-Hochberg adjustment to control for false discovery (type 1 error) rates. Analyses specific to pre-processing, the development of models, and pre-registration 1 are presented in the Supplement section 1.

Linguistic feature extractions

Two types of linguistic features were extracted for the mental health assessment: 1) word embeddings from a large language model (*RoBERTa-large, layer 23*; (30)) and 2) topics (*N*=300) prevalence scores based on the topic model created in the development dataset using Latent Dirichlet Allocation (31).

Pre-registered Models, Lexica, and Topics

Pre-registered PTSD Severity Models. We pre-registered models for PCL total score and the four subscales based on i) only language, ii) only the demographic controls, and iii) language and demographic controls (see Supplement for more details about how these were developed).

Pre-registered Pre-trained Ngram Models and Word Count Lexica of Theoretically Grounded Dimensions. To quantify the associations between PTSD severity and related theoretically grounded dimensions, we use pre-trained n-gram models (weighted lexica) trained on Facebook and Twitter language (32,33) to predict their self-reported neuroticism, depression, and anxiety.

We use word count lexica, including categories from the Linguistic Inquiry Word Count (34), including death, first-person singular and plural pronouns, and word lengths, as these were significantly related to PTSD in previous research (10). To assess respondents' re-experience of the WTC attack, we selected a lexicon combining five and seven LDA-topics relating to re-experiencing the attack.

Open Vocabulary Topics: Pre-registered Topics (Word Clouds). Topics significantly associated with PCL scores after controlling for demographics will be plotted using DLATK defaults. Three topics are pre-registered from analyses of the development dataset (for more details, see Supplement).

RESULTS

Participants' PCL scores did not significantly differ between the development (*M*=26.29; *SD*=11.7) and the prospective (*M* = 26.04; *SD*=10.5) datasets (*t*=.36, *df*=1781, *P*=.722), neither did the mean number of words in the development (*M*=838; *SD*=629) and prospective (*M*=776; *SD*=475) automated interviews (*t*=-1.74, *df*=1781, *P*=.082).

Language-based assessment of PTSD severity

Within the prospective language data, the pre-registered pre-trained language-based assessments of PTSD symptom severity produced scores that significantly correlated with the PCL total scores (*r*=.38; Table 1) and subscales (*r*=.28–.37). All correlations are above the pre-registered cut-offs based on the cross-validated correlations from the development set. Further, all the PCL models based on language and demographics, except for the Re-experience subscale, produced significantly less error than the pre-registered baseline models using only demographics (*r*s=.10–.15; age, gender, occupation, and race). Overall, the pre-registered models yielded correlations in the prospective dataset corresponding to the cross-validated correlations in the development training dataset (*r* difference range from .02–.08).

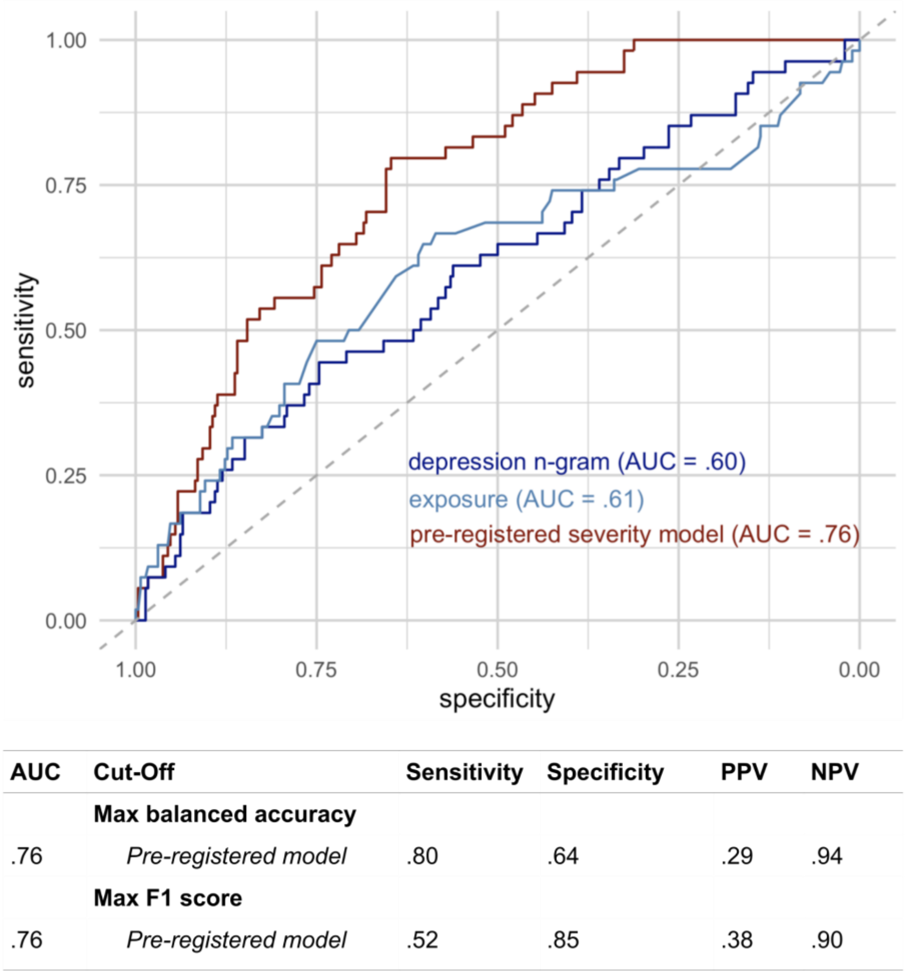
Table 1.
Development and Prospective Evaluation Pearson r of PCL Subscales Severity

Scale	*N*		*Demographics (baseline)*		*Language*		*Language + Demographics*	
	Devel.	*Prosp.*	*Devel.*	*Prosp.*	*Devel.*	*Prosp.*	*Devel.*	*Prosp.*
PCL	1437	346	.15***	.13*	.42***	.38***↑↑↑	.41***	.38***↑↑↑
PCL-subscales								
Reexperiencing	1407	341	.10***	.15**	.36***	.30***	.34***	.31***
Avoidance	1422	346	.05^{ns}	.11^{ns}	.25***	.28***↑↑	.22***	.28***↑↑↑
Emotional Numbing	1407	336	.15***	.14*	.39***	.31***	.38***	.31***↑
Hyperarousal	1402	342	.13***	.10^{ns}	.35***	.37***↑↑↑	.30***	.37***↑↑↑

Models are based on embeddings (layer 23 roberta-large) and topics;
*** = P< .001; ** = P < .01; * = P<.05; ^{ns} = P > .05;
↑ = the model accuracy is significantly higher than the corresponding demographics model (i.e., produces significantly lower error; ↑↑↑ = P < .001, ↑↑ = P < .01**,** ↑ = P < .05); this was only tested for Language and Language + Demographics in the prospective dataset.
PCL = The PTSD CheckList, devel. = development dataset; prosp. = prospective dataset
These results are without any adjustments to account for shrinkage via regularization in the machine learning models (see supplement section 3).

Clinical validation

The pre-registered model for PTSD severity was finally validated against participants' PTSD diagnosis from their medical records (not pre-registered). For the max balanced accuracy score (.72), the pre-registered model for PTSD severity yields a sensitivity of .80 and a specificity of .64. The pre-registered model yields an AUC of .76 (Figure 2); it significantly outperforms the demographics-based model (AUC=.61, P=.006), an exposure-based model (AUC=.61, P=.007) with predictors related to PTSD in previous research (26–28), and a depression n-gram model (AUC=.60, P<.001).



AUC	Cut-Off	Sensitivity	Specificity	PPV	NPV
	Max balanced accuracy				
.76	*Pre-registered model*	.80	.64	.29	.94
	Max F1 score				
.76	*Pre-registered model*	.52	.85	.38	.90

Figure 2. ROC curve (receiver operating characteristic curves) and classification accuracy metrics for PTSD diagnosis in medical records by language-based assessment using pre-registered models applied to prospective data· The pre-registered language model (green) significantly outperforms a WTC exposure model (red; DeLong's test: Z=2.70, P=.007 with predictors associated with PTSD in previous research (26–28)) and a depression model (blue; DeLong's test: Z=3.47, P<.001), which was the most accurate language-based assessment for PTSD severity in Son et al. (10)).

Lexical assessments of PTSD symptom severity

Lexical (word-based) assessments of depression (*r*=.18) and anxiety (*r*=.16) correlated significantly with overall PCL scores in the prospective language (Table 2). The difference in correlation between the development and prospective datasets ranges from .01–.04, demonstrating that the relationship is robust but small.

In addition, the three pre-registered topics were also significantly associated with PCL scores (Figure 3): topic 288 (on *stress, anxiety, and pain)*, topic 286 (on *control)*, and topic 65 (on *mental health issues)*. The difference in standardized beta between the development and prospective datasets ranged from .03–.05, showing that the topic model generalized to future data by producing reliable results.

Table2.
Association of Lexicon (word-based) assessments with PTSD symptom severity

	***r* (Controlled for demographics)**	
Model, Lexical, or Topics	***Development dataset (N = 1437)***	**Prospective dataset** (only pre-registered models) **(*N = 346)***
<u>Pre-trained Ngram Models</u>		
Anxiety (Son et al., 2021) (+)	.17***	.16**
Depression (Son et al., 2021) (+)	.14***	.18**
Neuroticism (+)	.11***	-
<u>Word Count Lexica</u>		
First-person singular pronouns (I, me, my) (+)	.05	-
First-person plural pronouns (we, our) (-)	-.04	-
Word lengths (-):	-*.02*	-
LIWC2022 Death (+)	*.02*	-
<u>Hypothesized Topics in Pre-Registration 1</u>	.06*	-
Reexperiencing the WTC attack 1 (+)		
Reexperiencing the WTC attack 2 (+)	.05	-
<u>Open Vocabulary Topics</u>	**Standardised beta**	
topic 288 (on *stress, anxiety, and pain)*	.22***	.17***
topic 286 (on *control)*	.16***	.13***
topic 65 (on *mental health issues)*	.13***	.10**

*** = P < .001; ** = P < .01; ** = P < .05; - = not pre-registered.
Reexperiencing the WTC attack: 1 = 5 topics; 2 = 7 topics.

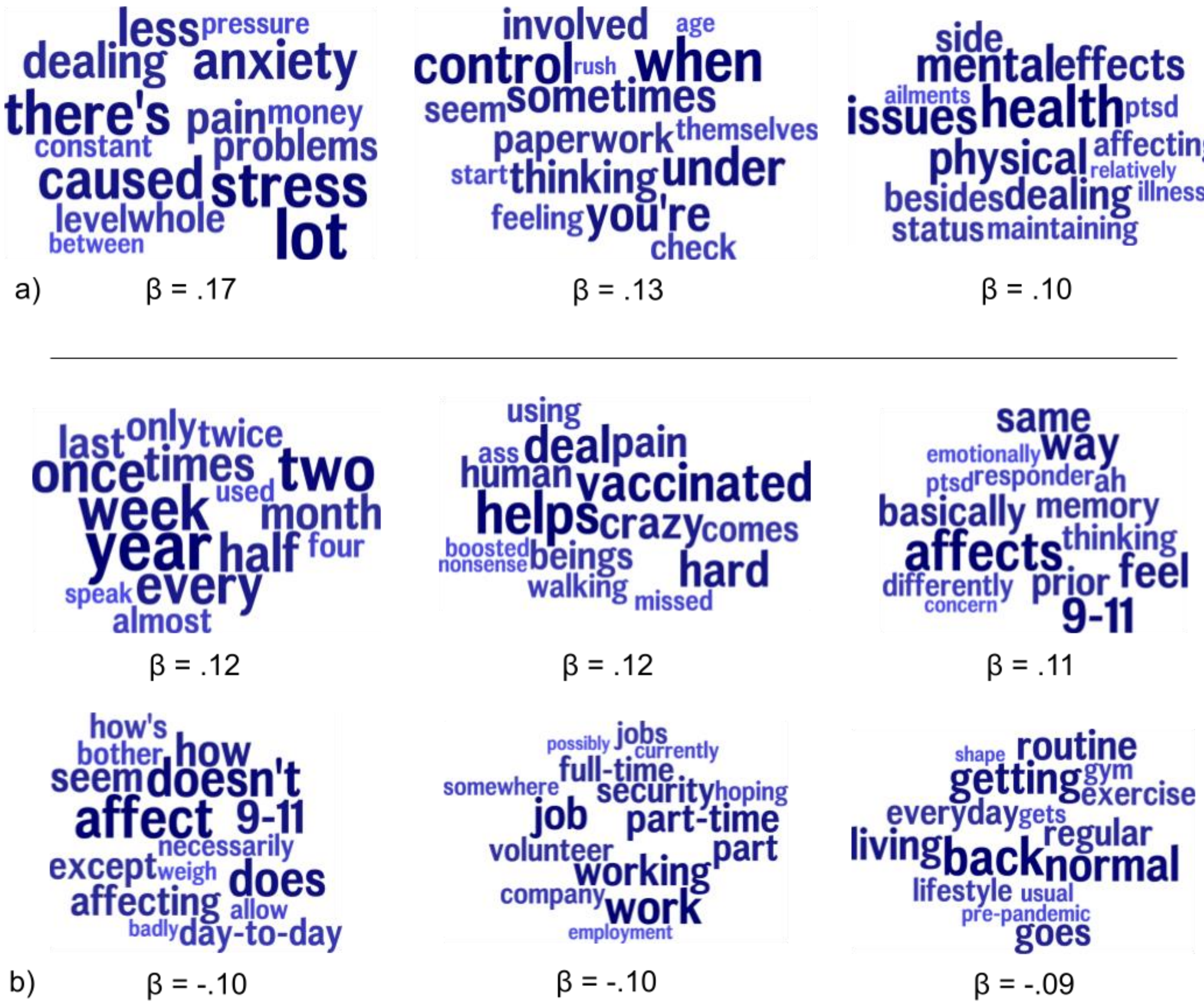


Figure 3. Topics (automatically grouped similar words) from the automated interview significantly related to PTSD severity. Scores are standardized linear coefficient (beta) controlling for demographics. Topics under (a) were pre-registered, while those in (b) were also significant in the development set but did not meet the power analysis criteria to be tested in the smaller test set.

DISCUSSION

The present study examined the extent to which language from automated clinical interviews relates to PTSD severity. We aimed to produce realistic estimates of psychiatric assessment accuracies, which was achieved by proposing the SEMP study design.

According to the first set of research questions, observable markers in the responders' natural language (via topics and word embeddings) were robustly associated with PTSD severity. The pre-registered models yielded correlations in the prospective dataset corresponding to the cross-validated correlations in the development dataset, where all correlations were above the pre-registered cut-offs. These results were also true for the PTSD severity subscales: Re-experiencing, Avoidance, Numbing, and Hyperarousal. Further, all the assessment models based on language and demographics, except for *Re-experiencing*, produced significantly less error than the baseline models based on only demographics. Notably, the SEMP procedure enables researchers to produce precise hypotheses regarding the effect size (which is uncommon in current practice).

Importantly, the pre-registered models were subjected to an external validation against PTSD diagnosis from individuals' medical records. The AUC score (.76) was significantly higher than baselines including demographics, exposures, and a previous state-of-the-art depression severity model (10); the AUC score was well above the cut-off for common clinical standards (.70, e.g., see the *COnsensus-based Standards for the selection of health Measurement INstruments*; COSMIN; (35)).

According to the second set of research questions, pre-trained n-gram models for depression and anxiety were positively correlated with observed PTSD severity. Further, the three pre-registered topics were significantly related to PTSD severity. Patients using language part of topics indicating struggles with *stress, anxiety, and pain*; *control*, and *mental health issues* are more likely to report higher PTSD severity. These findings provide insights about the type of language related to PTSD severity, and they are consistent with research showing that PTSD is comorbid with mental health issues, including depression, anxiety, and stress (10,36,37).

Implications

The AI-language models produced robust, behavior-based psychiatric assessments when evaluated prospectively over patients unseen by the model. This supports their strong potential clinical utility. The open-endedness of language responses allows patients to freely express heterogeneous symptoms and their unique experiences (1,2); the automated data collection and analyses of interview language can function as robust and scalable behavioral markers for research and clinical assessment among populations where the models have been prospectively evaluated. AI-based assessment methods have the potential to go beyond the current reliance on closed-ended rating scales, for example, they are suitable for screening (35) in telemedicine (38) or as outcome variables in Randomized Controlled Trials (RCTs).

Our new study design for rigorously developing and evaluating clinical AI-models –*SEMP*– yields results that are safeguarded against accuracy over-estimations. Accuracy estimations are crucial when implementing clinical AI-models (35). As such, the SEMP design could play the same critical role in developing and evaluating clinical AI assessment as the RCT design for evaluating interventions.

Limitations

While we suggest a procedure to assess the generalizability of AI models, the context for our analyses should still be considered. First, the AI models were developed and validated over a specific population – emergency responders to the WTC attacks. Future research should examine generalizability to other populations and traumas. Second, during the development phase, the interview questions underwent two iterations. The most substantial difference was that participants in the later stages (approximately 65% of the participants in the development and 100% in the prospective dataset) answered questions about the COVID-19 pandemic. Importantly, though, the final question set used during development was also used for evaluation within the prospective dataset, and having the exact same questions across all phases would likely have provided for greater accuracy. Additionally, motivated to prompt language generalizable to multiple outcomes, interview questions concerned participants' lives and experiences overall rather than specific clinical symptoms (e.g., see (39,40)), but further work is needed to verify if the question set actually generalizes better. Lastly, the models were trained to self-reported severity, while other options for PTSD assessment exist. Still, our validation against PTSD diagnosis from medical records mitigates this limitation.

Conclusions

Our pre-registered models demonstrated reliable and accurate assessments in prospective data, showing convergent validity with self-report and external validity with PTSD diagnoses from medical records. The accuracy of our language-based PTSD severity assessments outperforms models based on demographics, exposures, and depression. By introducing new safeguards to evaluate models, we demonstrate robust results supporting the potential of language-based assessments in psychiatric research and practice. Analyses of behavioral, observable markers in automated interview language produced robust, scalable psychiatric assessments, overcoming limitations found in traditional assessments.

Acknowledgements. We express our gratitude to the rescue and recovery workers of the World Trade Center attacks for their selfless dedication following the WTC attacks and for participating in this continuous research. Our thanks also extend to the clinical staff of the World Trade Center Medical Monitoring and Treatment Programs for their unwavering commitment and to the labor and community organizations for their ongoing support.

This work was supported in part by grant U01 OH012476 from CDC/NIOSH, *Developing and Evaluating Artificial Intelligence-based Longitudinal Assessments of PTSD in 9/11 Responders a grant from the NIH-NIAAA (R01 AA028032) awarded to H. Andrew Schwartz at Stony Brook University*, and by a NARSAD Young Investigator Grant to Joshua Oltmanns from the Brain & Behavior Research Foundation.

Conflict of Interest. Oscar Kjell has co-founded and holds shares in a start-up using language-based assessments to diagnose mental health problems.

The authors report no additional biomedical financial interests or potential conflicts of interest.

References

1. Galatzer-Levy IR, Huang SH, Bonanno GA. Trajectories of resilience and dysfunction following potential trauma: A review and statistical evaluation. Clin Psychol Rev. 2018;63:41–55.
2. Galatzer-Levy IR, Bryant RA. 636,120 Ways to Have Posttraumatic Stress Disorder. Perspect Psychol Sci. 2013 Nov;8(6):651–62.
3. Kjell O, Kjell K, Schwartz HA. Beyond Rating Scales: With Targeted Evaluation, Language Models are Poised for Psychological Assessment. Psychiatry Res. 2024;115667.
4. Baumeister RF, Vohs KD, Funder DC. Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? Perspect Psychol Sci. 2007 Dec;2(4):396–403.
5. Spitzer RL. Psychiatric diagnosis: are clinicians still necessary? Compr Psychiatry. 1983;
6. Schuler K, Ruggero CJ, Mahaffey B, Gonzalez A, L. Callahan J, Boals A, et al. When Hindsight Is Not 20/20: Ecological Momentary Assessment of PTSD Symptoms Versus Retrospective Report. Assessment. 2021 Jan;28(1):238–47.
7. Takayanagi Y, Spira AP, Roth KB, Gallo JJ, Eaton WW, Mojtabai R. Accuracy of reports of lifetime mental and physical disorders: results from the Baltimore Epidemiological Catchment Area study. JAMA Psychiatry. 2014;71(3):273–80.
8. Boyd RL, Schwartz HA. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. J Lang Soc Psychol. 2021;40(1):21–41.
9. Sawalha J, Yousefnezhad M, Shah Z, Brown MR, Greenshaw AJ, Greiner R. Detecting presence of PTSD using sentiment analysis from text data. Front Psychiatry. 2022;12:811392.
10. Son Y, Clouston SA, Kotov R, Eichstaedt JC, Bromet EJ, Luft BJ, et al. World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews. Psychol Med. 2021;53(3):918–26.
11. Oltmanns JR, Schwartz HA, Ruggero C, Son Y, Miao J, Waszczuk M, et al. Artificial intelligence language predictors of two-year trauma-related outcomes. J Psychiatr Res. 2021;143:239–45.
12. Preotiuc-Pietro D, Sap M, Schwartz HA, Ungar LH. Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In 2015. p. 40–5.
13. Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in Twitter. In: Proceedings of the international AAAI conference on web and social media [Internet]. 2014 [cited 2024 Feb 8]. p. 579–82. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14574
14. Todorov G, Mayilvahanan K, Cain C, Cunha C. Context-and subgroup-specific language changes in individuals who develop PTSD after trauma. Front Psychol. 2020;11:989.
15. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality [Internet]. 2014 [cited 2024 Mar 17]. p. 51–60. Available from: https://aclanthology.org/W14-3207.pdf
16. Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A, et al. Illusory generalizability of clinical prediction models. Science. 2024 Jan 12;383(6679):164–7.
17. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform. 2020;8(3):e17984.
18. Kernbach JM, Staartjes VE. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting. In: Staartjes VE, Regli L, Serra C, editors. Machine Learning in Clinical Neuroscience [Internet]. Cham: Springer International Publishing; 2022 [cited 2024 Feb 22]. p. 15–21. (Acta Neurochirurgica Supplement; vol. 134). Available from: https://link.springer.com/10.1007/978-3-030-85292-4_3
19. Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. Glob Transit Proc. 2022;3(1):91–9.
20. Tabassum A, Patil RR. A survey on text pre-processing & feature extraction techniques in natural language processing. Int Res J Eng Technol IRJET. 2020;7(06):4864–7.
21. Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning [Internet]. New York, NY:

Springer New York; 2001 [cited 2024 Mar 17]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-21606-5

22. Blanchard EB, Jones-Alexander J, Buckley TC, Forneris CA. Psychometric properties of the PTSD Checklist (PCL). Behav Res Ther. 1996;34(8):669–73.
23. King DW, Leskin GA, King LA, Weathers FW. Confirmatory factor analysis of the Clinician-Administered PTSD Scale: evidence for the dimensionality of posttraumatic stress disorder. Psychol Assess. 1998;10(2):90.
24. Ruggero CJ, Kotov R, Callahan JL, Kilmer JN, Luft BJ, Bromet EJ. PTSD symptom dimensions and their relationship to functioning in World Trade Center responders. Psychiatry Res. 2013;210(3):1049–55.
25. Dasaro CR, Holden WL, Berman KD, Crane MA, Kaplan JR, Lucchini RG, et al. Cohort profile: world trade center health program general responder cohort. Int J Epidemiol. 2017;46(2):e9–e9.
26. Bromet EJ, Hobbs MJ, Clouston SA, Gonzalez A, Kotov R, Luft BJ. 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. 2016;46(4):771–83.
27. Zvolensky MJ, Farris SG, Kotov R, Schechter CB, Bromet E, Gonzalez A, et al. World Trade Center disaster and sensitization to subsequent life stress: A longitudinal study of disaster responders. Prev Med. 2015;75:70–4.
28. Pietrzak RH, Feder A, Singh R, Schechter CB, Bromet EJ, Katz CL, et al. Trajectories of PTSD risk and resilience in World Trade Center responders: an 8-year prospective cohort study. Psychol Med. 2014;44(1):205–19.
29. Schwartz HA, Giorgi S, Sap M, Crutchley P, Ungar L, Eichstaedt J. Dlatk: Differential language analysis toolkit. In 2017. p. 55–60.
30. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. ArXiv Prepr ArXiv190711692. 2019;
31. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
32. Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, Stillwell DJ, et al. Automatic personality assessment through social media language. J Pers Soc Psychol. 2015;108(6):934.
33. Schwartz HA, Eichstaedt J, Kern ML, Park G, Sap M, Stillwell D, et al. Towards assessing changes in degree of depression through facebook. In 2014. p. 118–25.
34. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. Austin TX Univ Tex Austin. 2022;
35. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018 May;27(5):1147–57.
36. Brady KT, Killeen TK, Brewerton T, Lucerini S. Comorbidity of psychiatric disorders and posttraumatic stress disorder. J Clin Psychiatry. 2000;61:22–32.
37. Caramanica K, Brackbill RM, Liao T, Stellman SD. Comorbidity of 9/11-Related PTSD and Depression in the World Trade Center Health Registry 10–11 Years Postdisaster. J Trauma Stress. 2014 Dec;27(6):680–8.
38. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: Capabilities, features, barriers, and applications. Sens Int. 2021;2:100117.
39. Kjell O, Kjell K, Garcia D, Sikström S. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. Psychol Methods. 2019;24(1):92.
40. Kjell O, Sikström S, Kjell K, Schwartz HA. Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. Sci Rep. 2022 Mar 10;12(1):3918.

Supplemental

Demonstrating High Validity of a New AI-Language Assessment of PTSD: Sequential Evaluations with Model Pre-registration

Section 1: Hypotheses, methods, analyses, and results for pre-registration 1

In this section, we describe the hypotheses, methods, analyses, and results for pre-registration 1.

Hypotheses for pre-registration 1

Pre-registration 1 registered four sets of hypotheses:

H_1 Concurrent PTSD symptom severity is predictable from automated clinical interview language.

H_2 Concurrent PTSD symptom severity is predictable from automated clinical interview language above and beyond three levels of baselines:

A. demographics (age, gender, occupation, and race),
B. prior PTSD Symptom Severity and demographics
C. concurrent MCS and PCS prior PTSD Symptom Severity and demographics.

H_3 Language-based assessments of each of the following from the automated clinical interview will be associated with concurrent PTSD symptom severity (positively or negatively as indicated by + or - respectively).

Replicating Son et al., 2021:

D. Anxiety (+)
E. Depression (+)
F. Neuroticism (+)
G. First-person singular pronouns (I, me, my) (+)
H. First-person plural pronouns (we, our) (-)
I. Word lengths (-)

Additional Features:

J. Discussion of death (+)
K. Words related to re-experiencing the WTC attacks (+) (c.f. Hellawell & Brewin, 2004; Papini et al., 2014).

H_4 Open-vocabulary linguistic features of automated clinical interviews will be associated with concurrent PTSD Symptom Severity.

Methods for pre-registration 1

Participants

Table Supplement 1 shows the demographics across the datasets.

Table Supplement 1.
Demographics across data split

Dataset	N	Age (SD)	Female	Male	Asian	Black	Multi	White	Unk.	Police
Development *(before August 1, 2022)*	1437	57.9 *(8.0)*	107 (7%)	1330 (93%)	8 (<1%)	29 (2%)	42 (3%)	779 (54%)	579 (40%)	963 (67%)
Prospective *(after August 1, 2022)*	346	58.5 (7.8)	30 (9%)	316 (91%)	1 (<1%)	3 (<1%)	16 (5%)	171 (49%)	155 (45%)	240 (69%)

Note. Unk. = Unknown (participant has not self-reported any race).

Reliability of the PCL

We computed the total mean score (Cronbach alpha [**α** development set/prospective set] = .95/.93; Hierarchical omega (**ω** developing set/prospective set]) = .82/.81) and the four subscales (4,28), including Reexperiencing (**α** = .90/.86; **ω** = .87/.78), Avoidance (**α** = .84/.76; **ω** = –), Numbing (**α** = .95/.82; **ω** =.82/.82), and Hyperarousal (**α** = .87/.85; **ω** = .84/.70).

Description of the analyses in the development set

In pre-registration 1 we did not specify the diarization and removal of instances where respondents read the questions out loud (they were specified in pre-registration 2). The pre-processing, including transcriptions and diarization, and removing parts where respondents read the question out loud were developed in the development set as specified in the main paper and below.

Preprocessing

Audio from the recordings was transcribed using *Whisper Large* (v2; (30)), a state-of-the-art AI transcription model. The average word error rate was 2.18% (over 28 randomly selected transcripts); calculated as the ratio between the number of word changes made to the Whisper transcriptions to match human transcriptions. Subsequently, the transcriptions were stripped of segments corresponding to the interviewer's speech using the Pyannote speaker-diarization AI model (31). After verifying that the interviewee always spoke the most by an order of magnitude, we limited the text to that identified as the most frequent speaker.

Removing parts where respondents read the question out loud

Within the training data, we noticed a strong variation in how participants read the questions on the screen aloud. To limit analyses to the participants' original language, the reading of questions was removed by automatically finding sequences of tokens that matched the subsequences of the questions. The approach to remove questions used a 2-gram lexicon, topics, and contextualized word embeddings (that capture longer sequences) from a large language model (*RoBERTa-large, layer 23*; (32)) as features in a ridge regression model ((33)). To evaluate its accuracy, one of the authors labeled 55 transcriptions (see Supplement for more information). The question detection model yielded an accuracy of 97% (*F1* = .78, *AUC* = .94; using 10-fold cross-validation).

Linguistic feature extraction

As specified in pre-registration 1, three types of language features were extracted on the development set, including 1) word embeddings from a large language model (*RoBERTa-large, layer 23*; Liu et al., 2019) which have been found suitable for human-level predictions from word embeddings (Ganesan et al., 2021), 2) words and phrases encoded as relative frequencies and as binary indicators, and 3) topics (N = 300) prevalence scores based on Latent Dirichlet Allocation (LDA; Blei et al., 2003).

Model training and selection

Models were examined using cross-validation. Following the procedure described in pre-registration 1 the models were developed using all development data using ridge regression with a penalty ranging from 10^1 to 10^6. We created models for PCL total and the four subscales based on i) only language, ii) only the demographic controls, and iii) language and demographic controls.

Significance testing accuracy from different assessment models

To significance test the accuracy of two language-based assessment models, the accuracy of model A (ŷ1) and model B (ŷ2) to predict (i.e., estimate) the observed scores (y) were compared using a paired t-test between the absolute error from ŷ1-y and ŷ2-y.

Lexicon

To quantify the associations between PTSD severity and related theoretically grounded dimensions, we used a lexicon trained on Facebook and Twitter language (Park et al., 2015; Schwartz et al., 2014; Guntuku et al., 2019) to predict their self-reported neuroticism, depression, and anxiety. These lexicons were applied to the language data to derive a language-based estimate of the expression of these constructs in the participants' language.

We had registered only to include the 1400 most frequently used words, which resulted in substantially lower significance, although here we present both results).

Domain Adaptation. The language-based models for depression, anxiety, and neuroticism were trained on 10s of 1000s of Facebook users, which have been shown to be effective in previous works (Eichstaedt et al., 2018; Guntuku et al., 2019). However, the depression, anxiety, and neuroticism scores estimated for these interviews using models trained on language that has different lexical usage patterns would result in large errors (Rieman et al., 2017). Thus, as part of exploratory analyses (i.e., not part of pre-registration 1 and not used in pre-registration 2), these models were adapted to account for the change in domain, which acts as a forcing function to the change in lexical usage. In this process, we removed the words that had dissimilar distributions in terms of either frequency or sparsity from the source (FB lexica) and re-trained the model from scratch.

Open vocabulary. To assess respondents' re-experience of the World Trade Center attack, we selected five and seven LDA topics relating to re-experiencing the attack (as specified in the pre-registration 1).

Closed vocabulary. To assess *Death*, *First-person singular pronouns* (I, me, my) and *First-person plural pronouns* (we, our), Linguistic Inquiry, and Word Count (LIWC-22; Boyd et al., 2022) dictionaries were used (as specified in the pre-registration 1). The LIWC dictionaries are comprised of psychologically meaningful categories of words that have been developed and extensively validated by psychologists. The assessment is computed as the relative frequency of each LIWC dictionary by summing the within-participant word frequencies within each of the three LIWC dictionaries.

Word clouds

Topics that are significantly associated with PCL scores after controlling for demographics and multiple corrections using Bonferroni correction were plotted using DLATK defaults (as specified in the pre-registration 1).

Missing data, exclusions, and transformation

Participants with responses of fewer than 150 words were excluded from the analyses. We had registered 100 words as a response to the life narrative as a minimum for including a participant in the analyses in pre-registration 1. However, we found that a minimum of 150 words yielded more accurate results, which we thus registered for pre-registration 2. We are presenting the 150 threshold for consistency even for the pre-registration 1 hypothesis below. We also present analyses showing accuracy across different word response cut-offs. Missing data were imputed for PCL total scores by computing the mean, where participants had answered at least 14 of the 17 items (see Schafer & Graham, 2002).

Measures Unique for Pre-registration 1

The Short Form 12 (SF-12) Health Survey

Physical and mental health was assessed using the SF-12 (Ware et al., 1995) from the Short Form 36 Health Survey (Brazier et al., 1992), which focuses on functional status, well-being, and overall evaluation of health. It includes 12 closed-ended items assessing eight broad domains: Physical functioning, Social functioning, Role limitations (physical problems), Role limitations (emotional problems), Pain, Mental Health, Vitality, and General health perception. From the SF-12, we computed the Physical Component Summary (PCS) and Mental Component Summary (MCS) using standardized procedures based on regression weights (Ware et al., 1995).

Labeling scheme for when participants read questions out loud

The labeling scheme developed during analyses of the development set comprised three levels indicating different degrees of the extent to which the respondent was reading the question out loud (see below). The first author labeled fifty-eight transcripts; twenty were transcribed by a research assistant to examine the inter-rater reliability. Overlap was calculated as average(number of spans matched / total number of spans), i.e., average(Jaccard). A span is considered to be matched if, between two annotators, there is at least a 50% overlap. The inter-rater correlation was .60. In the final analyses, the models were trained to predict levels 2 - 4 as indicative of a part where participants read the question out loud.

Labeling scheme

Levels	**Description**
(no label)	No reference to the question (will never be removed)
<< **2** >>	Reminding of language in question
<<< **3** >>>	Referencing the question connected to the response
<<<< **4** >>>>	Disconnected reading of the questions; the part is not needed to understand the response.

Results for pre-registration 1

Table Supplement 2 shows descriptive statistics of rating scales. Table Supplement 3 presents descriptive data on the number of words in the narrative in the development and prospective data sets. Table Supplement 4 shows correlations among rating scales and language-based assessed PCL scores.

Table Supplement 2.
Descriptive Word Statistics of the Transcribed Video-Recordings

	N	**Mean**	***SD***	**Min**	**Max**
Development data					
Before removing questions, read aloud	1437	921	665	140	6068
After removing questions, read aloud	1437	838	629	123	5949
Prospective data					
Before removing questions, read aloud	346	831	495	138	2368
After removing questions, read aloud	346	776	475	133	2290

Table Supplement 3.
Descriptive Statistics of Rating Scales

Measure	*N*	Mean	*SD*	Min-Max	Skew	Kurtosis
PCL development dataset	1437	26.29	11.7	17.0 – 75.0	1.63	2.21
PCL prospective dataset	346	26.04	10.5	17 .0 – 63.0	1.34	1.10
PCS development dataset	1308	46.07	10.64	15.4 – 64.0	-0.72	-0.59
MCS development dataset	1308	52.49	8.86	16.7 – 67.9	-1.25	1.22

Note. PCL = The PTSD CheckList; PCS = Physical Component Summary; MCS = Mental Component Summary.

Table Supplement 4.
Correlations among variables in the development dataset

Measures	1	2	3
PCL			
Predicted PCL	.42		
PCS	-.39	-.28	
MCS	-.71	-.36	.16

Notes. All significant at < .001; *N* ≈ 1437 for PCL and predicted PCL, and 1308 for correlations with MCS and PCS. PCL = The PTSD CheckList.

H_1: Concurrent PTSD symptom severity is predictable from automated clinical interview visit language

The results about the accuracy of the machine learning-based assessments are presented in Table Supplement 5. These analyses were carried out in accordance with H1 and the pre-registered secondary analysis, including an ablation analysis of the different types of features, including 1) Embeddings; 2) Embeddings + topics; 3) Ngrams rel/bin + topics (was SOTA for personality); 4) Embeddings + topics + grams. The results show that the most accurate model is based on Embeddings + Topics.

The results presented in Table Supplement 6 and Supplement 7 were not registered in pre-registration 1. Table Supplement 6 shows that the minimum number of 150 words in a response yields the most accurate predictions. This is why we changed this threshold from 100 to 150 words in pre-registration 2.

Table Supplement 7 shows that the models for PCL subscales perform with significant assessment accuracy beyond demographic information, leading us to include these in pre-registration 2.

Table Supplement 5.
Prediction Accuracy Pearson r of Concurrent PTSD Symptom Severity

Features	$\boldsymbol{r}_{\text{(PCL)}}$	$\boldsymbol{r}_{\text{(PCL Anscombe)}}$
N-grams	.32	.33
N-gram top 1400	.29	.31
Topics	.31	.32
Embeddings	.39	.39
N-grams + Topics	.33	.35
N-gram_1400 + Topics	.33	.35
Embeddings + N-grams	.35	.37
Embeddings + Topics	**.42**	**.42**
Embeddings + Topics + N-grams	.36	.37
Embeddings + Topics + N-grams_1400	.39	.40

Notes. Embeddings = Layer 23 Roberta-large; $p < .001$, $N = 1437$; N-gram top 1400 was pre-registrated in registration 1, but not in 2. PCL = The PTSD CheckList

Table Supplement 6.
Prediction Accuracy Pearson r of Concurrent PTSD Symptom Severity

Group word threshold		**Pearson's *r***	
		PCL	
	N	**10-folds**	**5-folds**
1	1535	.38	.37
50	1526	.39	.39
100	1508	.40	.40
150	**1475**	**.41**	**.41**
200	1416	.40	.39
250	1358	.40	.40
300	1299	.41	.41
400	1166	.41	.38
500	1031	.39	.39
1000	510	.35	.33

Notes. Features included Embeddings + Topics. PCL = The PTSD CheckList

Table Supplement 7.
Prediction Accuracy Pearson r of Concurrent PCL Subscales Severity

Scale	***r***			
	Language	***Demographics***	***Language + demographics***	***N***
PCL	.42	.15	.41	**1437**
PCL-subscales				
Reexperiencing	.36	.10	.34	1407
Avoidance	.25	.05 (ns)	.22	1422
Emotional Numbing	.39	.15	.38	1407
Hyperarousal	.35	.13	.30	1402

Notes. Prediction based on the best model from Table 5, including embeddings (layer 23 roberta-large) and topics; all significant at p < .001, except for ns.
PCL = The PTSD CheckList

H_2: Concurrent PTSD symptom severity is predictable from automated clinical interview language above and beyond three levels of baselines

The analyses show that the language-based models are more accurate than demographics alone (by H_2A; Table Supplement 7). However, language-based models are not more accurate than prior PCL scores and demographics (H_2B; Table Supplement 8) or concurrent MCS and PCS scores and prior PCL scores and demographics (H_2C; Table Supplement 9). Therefore, pre-registration 2 focuses on the language-based assessments' ability to predict beyond demographics, as we note that these variables do not have shared method variance with the three rating scales PCL, MCS, and PCS.

Table Supplement 8
Prediction Accuracy of Concurrent PTSD Symptom Severity

Features	$r_{(PCL)}$		***N***
	First collected PCL	**Last previous PCL**	
Language	.43	.42	1st 1194
Prior PCL	.65	.84	last: 1135
Language + Prior PCL	.67	.84	
Language	.42	.43	1st: 1182
Prior PCL + Dem.	.64	.84	
Language + Prior PCL + Demographics	.66	.84	last: 1123

Demographics = age at visit, gender, occupation (police or not) and race. PCL = The PTSD CheckList.

First collected PCL score: Time difference of 3804 (SD = 2135; range = 339 – 7273) days, with a correlation of .64.

Last previous PCL score: Time difference of 631 (SD = 478; range = 335 - 6902) days, with a correlation of .84.

Table Supplement 9
Prediction Accuracy of Concurrent PTSD Symptom Severity

Features	***r*****(PCL)**	***N***
MCS and PCS	*.76*	1226
Language + MCS and PCS	*.76*	
MCS and PCS + PCL + Demographics	*.81*	*1118*
Language + MCS and PCS + PCL + Demographics	*.80*	

Demographics = age at visit, gender, occupation (police or not) and race; PCL = The PTSD CheckList.

H_3: Language-based assessments of each of the following from the automated clinical interview language are associated with concurrent PTSD symptom severity
The results showed that the lexical assessments of A) anxiety, B) depression, and C) neuroticism were positively related to PCL scores, which is in accordance with the pre-registered hypotheses. In contrast to the hypotheses, lexical assessments that were not significantly correlated with PCL scores included D) first-person singular pronouns (I, me, my), E) first-person plural pronouns (we, our), and F) word lengths. Nor were PCL scores significantly correlated with G) discussion of death or words related to re-experiencing the WTC attacks.

In addition, analyses we found that domain-adapted lexical models did not produce stronger correlations.

Table Supplement 10
Language-based assessments associations with concurrent PTSD symptom severity

Model/Lexica	NOT Adapted (controlled)		Domain adaptation (controlled)	
	r	***p***	***r***	***p***
Anxiety (+)	.17	<.001	.14	<.001
Depression (+)	.14	<.001	.12	<.001
Neuroticism (+)	.11	<.001	.13	<.001
First-person singular pronouns (I, me, my) (+)	.05	.124	*NA*	*NA*
First-person plural pronouns (we, our) (-)	-.04	.124	*NA*	*NA*
Word lengths (-):	-*.02*	*.580*	*NA*	*NA*
LIWC2022 Death (+)	*.02*	*.696*	*NA*	*NA*
Reexperiencing the WTC attack 1 (+)	.06	.042	*NA*	*NA*
Reexperiencing the WTC attack 2 (+)	.05	.059	*NA*	*NA*

Note. N = 1437;
Reexperiencing the WTC attack 1 (+) = 5 topics (reexp_1);
Reexperiencing the WTC attack 2 (+) = 7 topics (reexp_2);

H_4: Open-vocabulary linguistic features of automated clinical interviews will be associated with concurrent PTSD Symptom Severity

Figure Supplement 1 shows the six most positively associated topics with PCL scores (p < .001). Figure Supplement 2 shows the six most strongly negatively associated topics with PCL scores (p < .001). The three topics at the top of Figure Supplement 1 were pre-registered in pre-registration 2 (*N* [prospective dataset] = 346, alpha = .05, and a power = 80%, giving a correlational strength of .13).

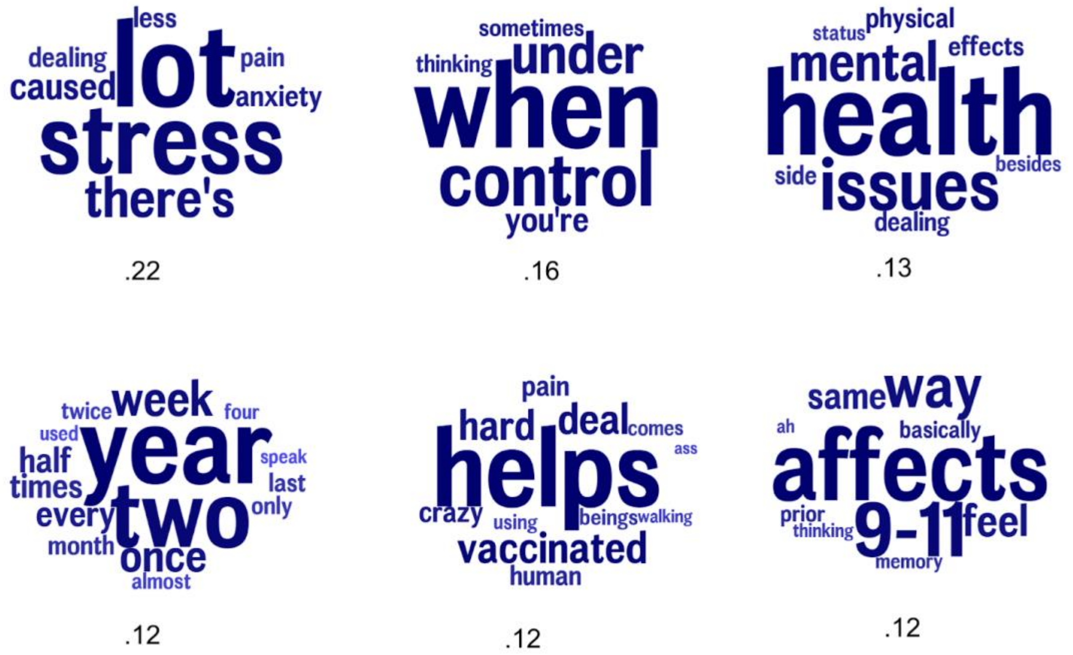


Figure Supplement 1 | Topics positively associated with PCL controlling for demographics (with betas)

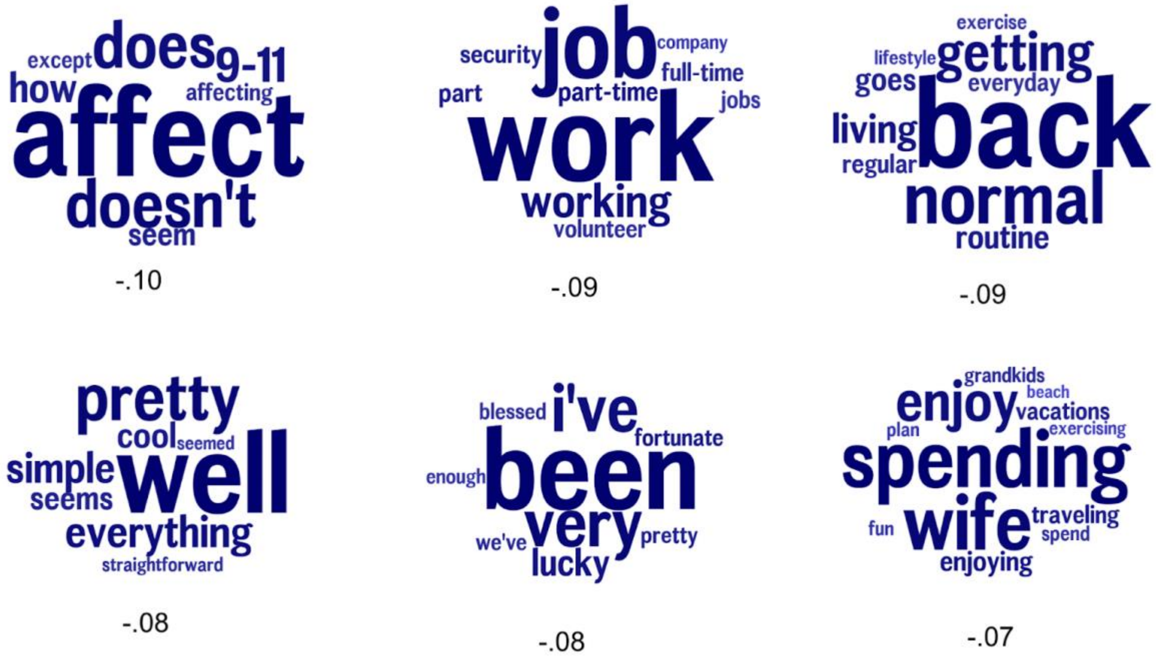


Figure Supplement 2 | Topics negatively associated with PCL controlling for demographics (with betas; top 6 of 6)

Section 2: The three different question sets across participants

The first iteration of questions (*N* = 144)

Question - code	Question
OEL_1	**How are you?**
OEL_1E	**Can you elaborate?**
OEL_2	**How's the family?**
OEL_2A	**Can you elaborate?**
OEL_3	**What's new?**
OEL_3A	**Can you elaborate?**
OEL_4	**What else?**
OEL_5	Over the past 5 years what are the three nicest things that happened to you and your family?
OEL_5A	Can you elaborate?
OEL_6	Over the past 5 years what are the three worst things that happened to you and your family?
OEL_6A	Can you elaborate?
OEL_7	Imagine you are 5 years older. Please tell us about the life you are leading, your interests, your home life, and your work.
OEL_7A	Can you elaborate?

Notes. **Bold** = questions that are removed in the forthcoming iterations; **Q-code = question code**

The second iteration of questions (*N* = 631)

Question - code	Question
Q #1	**What are the three things in your life that you look forward to the most right now?**
Q #1A	**Can you elaborate?**
Q #2	**What are the three biggest challenges that you are managing in your life right now?**
Q #2A	**Can you elaborate?**
Q #3	**Where are three places that you turn to find support right now and why?**
Q#3A	**Can you elaborate?**
Q #4	Over the past 5 years, what are the three nicest things that happened to you and your family?
Q#4A	Can you elaborate?
Q#5	Over the past 5 years, what are the worst three things that have happened to you and your family?
Q#5A	Can you elaborate?
Q #6	Imagine you are 5 years older. Please tell us about the life you are leading, your interests, your home life, and your work.

The second iteration of questions (*N* = 631)

Question - code	Question
Q #1	**What are the three things in your life that you look forward to the most right now?**
Q #1A	**Can you elaborate?**
Q #2	**What are the three biggest challenges that you are managing in your life right now?**
Q #2A	**Can you elaborate?**
Qn #6A	Can you elaborate?
Q #7	**What are three reflections that you have on this research experience? Were any parts particularly challenging or interesting? Do you have any recommendations for us continuing this research?**

Notes. **Bold** = questions that are removed in the forthcoming iterations.

The third iteration of questions (*N* = 1306)

Question - code	Questions
1	What are the three things in your life that you look forward to the most right now? Can you elaborate?
2	What are the three biggest challenges that you are managing in your life right now? Can you elaborate?
3	Where are three places that you turn to find support right now and why? Can you elaborate?
4	Over the past 5 years, what are the three nicest things that happened to you and your family? Can you elaborate?
5	Over the past 5 years, what are the worst three things that have happened to you and your family? Can you elaborate?
6	Imagine you are 5 years older. Please tell us about the life you are leading, your interests, your home life, and your work. Can you elaborate?
7	**How has the COVID-19 pandemic changed your life? Can you elaborate?**
8	**What was most difficult for you about the COVID-19 pandemic? Can you elaborate?**
9	**What do you most look forward to doing after the COVID-19 pandemic is over? Can you elaborate?**
10	**Looking back at the last two decades, what effect did 9/11 have on your life? Can you elaborate?**
11	**How does 9/11 affect you now? Can you elaborate?**
12	**What would you like future generations to know about 9/11? Can you elaborate?**
13	What are three reflections that you have on this research experience?
14	Were any parts particularly challenging or interesting?
15	Do you have any recommendations for us continuing this research?

Notes. **Bold** = questions that are removed in the forthcoming iterations.

Section 3: The Prospective Dataset

In this section, we describe supplementary details regarding hypotheses, methods, and results for the prospective datasets.

Hypotheses for pre-registration 2

Language-based assessments of specific psychologically related dimensions from stage 1 will be correlated with the overall PTSD symptom severity in the stage 2 prospective sample. These include:
a) language-based anxiety – positive association,
b) language-based depression – positive association,
c) topic 288 (on *stress, anxiety, and pain)*;
d) topic 286 (on *control)*, and
e) topic 65 (on *mental health issues)*.

Method for pre-registration 2

Details about measures

***WTC exposure*.** The questions regarding WTC exposure concerned 9/11, including 1) being caught in the dust cloud, 2) death of a colleague, family member, or friend, 3) knowing someone who was injured, 4) involved in search and rescue efforts, 5) working primarily at or adjacent to the towers collapse site, 6) exposure to human remains of victims, 7) early arrival (i.e., on 9/11 or 9/12 2001), 8) long work on-site (i.e., total hours worked in the top quartile of this sample), 9) slept on-site any nights during September or October 2021, 10) worked on-site every day from 9/11 to 9/30.

Results for pre-registration 2

Table 11.
Prediction Accuracy Pearson r of Concurrent PTSD Symptom Severity of Exposures, Demographics and Language

Scale	*N*	*Exposures (baseline)*	*Exposures + Demographics (baseline)*	*Language*	*Language + Exposure*	*Language + Demographics + Exposure*
PCL	1014	.11***	.23***	.34***	.35***	.36***
PCL-subscales						
Reexperiencing	1002	.09**	.23***	.22***	.34***	.31***
Avoidance	1012	.08**	.13***	.18***	.21***	.22***

Emotional Numbing	1004	.08*	.19***	.33***	.34***	.35***
Hyperarousal	996	.10**	.21***	.28***	.30***	.27***

Notes. Models are based on embeddings (layer 23 roberta-large) and topics;
*** = P < .001; ** = P < .01; * = P <.05;
PCL = The PTSD CheckList, devel. = development dataset; prosp. = prospective dataset

Redistributing prediction to better comply with clinical standards

AI models have a tendency to "shrink" the variance as part of their regularization, which we did not adjust for in our primary results. Applying the PCL cut-off of 44, commonly used in the literature, the model predictions yield a sensitivity of 0 and a specificity of 1. This is because the distribution of the prediction models has changed. One potential concern with many machine learning-based approaches for regression is that they are not typically constrained to have a similar variance and distribution as the training outcome. In particular, penalization, by definition, shrinks the variance to avoid overfit and often outputs a more normal distribution than the training outcome. These issues have a negligible effect on typical accuracy metrics such as correlation, but when put into practice, they can make predicted outcomes harder to interpret for practitioners used to scales having particular cut-offs, ranges, and distributional shapes. Next, we show how it is possible to update the model to yield a more similar distribution to the training outcome.

Redistributing the predictions was solved through two general steps: (1) *Anscombe-loss:* adding a transformation to account for shape during training and (2) adding a variance transformation (called “predictive redistribution”; Giorgi et al., 2022) on the output of the regression predictions. Hence, we executed the following steps:

1. Transformed PCL scores using Anscombe (e.g., see Anscombe, 1973)

$$2y + 38$$

2. Trained the language models to the Anscombe transformed PCL scores.

3. Redistributed the predicted Anscombe transformed PCL scores according to observed PCL score.

$$(ypred - \mu ypred\ \sigma ypred)\ \sigma y + \mu y$$

where:

ypred = the predicted value on the standardized scale,

$\mu ypred$ = the mean of the predicted values on the standardized scale,
$\sigma ypred$ = the standard deviation of the predicted values
σy = the standard deviation of the original target variable y.
μy = the mean of the original target variable

4. Un-Anscombe the redistributed Anscombe predictions.

$$(y2)^2 - 38$$

5. Finally, we constrained the range according to the observed PCL scores.

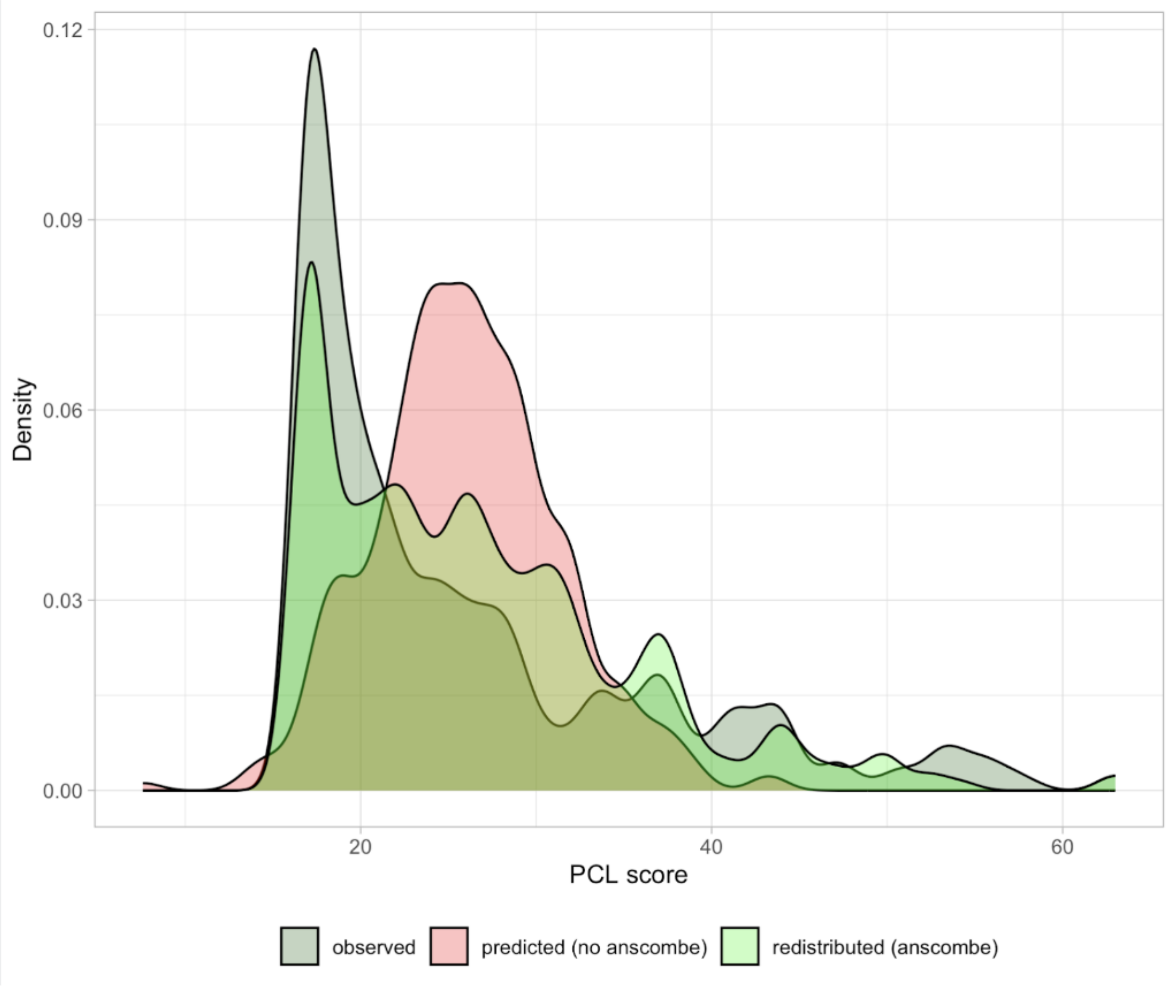


Figure Supplement 3 | Redistribution transformation procedures can make predictions Density plot of distributions for observed, predicted, and redistributed predictions of PCL scores.

Applying the PCL cutoff of 44 on the redistributed model predictions now provides a sensitivity of .96 and a sensitivity of .11. Applying the PCL cut-off of 27 (which is the cutoff with the highest balanced accuracy for observed PCL scores in the development set), the model predictions yield a sensitivity of .70, and a specificity of .70. This is because the distribution of the prediction model has changed, where the redistributed model yields a sensitivity of .78 and a specificity of .66.

Table Supplement 11. Classification accuracy metrics for the full model on PTSD diagnosis in medical record

Model	Threshold	AUC	F1	Sens	Spec	PPV	NPV
Max balanced accuracy							
Pre-registered model	26.7	.76	.43	.80	.64	.29	.94
Max F1							
Pre-registered model	29.7- 29.9	.76	.44	.52	.85	.38	.90
Threshold 44							
Pre-registered model	44	.76	NA	0	.1	NaN	.84
Redistributed model	44	.76	.17	.11	.96	.35	.85
Threshold 27 (highest balanced accuracy for observed scores in the training set)							
Pre-registered model	27	.76	.40	.70	.66	.28	.92
Redistributed model	27	.76	.43	.78	.66	.30	.94

Reference

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, *27*(1), 17–21. https://doi.org/10.1080/00031305.1973.10478966

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

Brazier, J. E., Harper, R., Jones, N. M., O'cathain, A., Thomas, K. J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: New outcome measure for primary care. *British Medical Journal*, *305*(6846), 160–164.

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. *Proceedings of the National Academy of Sciences*, *115*(44), 11203–11208.

Ganesan, A. V., Matero, M., Ravula, A. R., Vu, H., & Schwartz, H. A. (2021). Empirical evaluation of pre-trained transformers for human-level nlp: The role of sample size and dimensionality. *Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting*, *2021*, 4515.

Giorgi, S., Lynn, V. E., Gupta, K., Ahmed, F., Matz, S., Ungar, L. H., & Schwartz, H. A. (2022). Correcting Sociodemographic Selection Biases for Population Prediction from Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 228–240.

Guntuku, S. C., Preotiuc-Pietro, D., Eichstaedt, J. C., & Ungar, L. H. (2019). What twitter profile and posted images reveal about depression and anxiety. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 236–246. https://ojs.aaai.org/index.php/ICWSM/article/view/3225

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv Preprint arXiv:1907.11692*.

Rieman, D., Jaidka, K., Schwartz, H. A., & Ungar, L. (2017). Domain adaptation from user-level facebook models to county-level twitter predictions. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 764–773. https://aclanthology.org/I17-1077/

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147.

Ware, J. E., Keller, S. D., & Kosinski, M. (1995). *SF-12: How to score the SF-12 physical and mental health summary scales*. Health Institute, New England Medical Center.