Empirical Article

Clinical Psychological Science
1–13

DOI: 10.1177/21677026261439026
www.psychologicalscience.org/CPS

Replicability and Validity of a New Artificial-Intelligence Assessment of Posttraumatic Stress Disorder From Patient Language: A Sequential Evaluation With Model Preregistration

Oscar Kjell[1,2,3], Adithya V. Ganesan[1,3], Ryan L. Boyd[4], Joshua Oltmanns[5], Alfredo Rivero[1], Scott Feltman[6], Melissa A. Carr[7], Jorge Alves[7], Benjamin Luft[7,8], Roman Kotov[6], and H. Andrew Schwartz[1,3]

[1]Department of Computer Science, Stony Brook University; [2]Department of Psychology, Lund University; [3]College of Connected Computing, Vanderbilt University; [4]Department of Psychology, University of Texas at Dallas; [5]Department of Psychology, Southern Methodist University; [6]Department of Psychiatry and Behavioral Health, Stony Brook University; [7]World Trade Center Health and Wellness Program, State University of Stony Brook; and [8]Department of Medicine, School of Medicine, State University of Stony Brook

Abstract

Artificial intelligence (AI) shows promise in identifying psychopathology through language, but replicability in AI models remains challenging. We develop an AI-based language assessment of posttraumatic-stress-disorder (PTSD) severity and introduce the sequential evaluation with model preregistration to rigorously evaluate its validity and replicability. This design includes two phases: development with preregistration and evaluation. Data included development (N=1,437) and prospective (N=346) samples, in which participants described their lives during automated interviews. In the prospective sample, preregistered models correlated with PTSD CheckList scores (r=.38, p<.001) and converged with PTSD diagnosis (area under the curve [AUC]=.76; outperforming demographics and trauma exposures: AUC=.61, p<.01). We found that for each standard-deviation increase, mental-health-care expenditure rose by $696.50 ($p$<.001). Our preregistered PTSD model assessments are replicable in prospectively collected clinical data and showed external validity against expense criteria. With further development, such models can be used to screen for PTSD or monitor treatment response, especially in telehealth or automated interviews, in which deployment can be seamless.

Keywords

posttraumatic stress disorder, depression, disaster responders, language-based assessments, oral interviews, World Trade Center, open materials, preregistration

Received 12/28/24; Revision accepted 2/10/26

Psychological trauma is common. Seventy percent of the general population experience trauma at some point in life (Kessler et al., 2017). Only a minority of people exposed to trauma develop posttraumatic stress disorder (PTSD; Galatzer-Levy et al., 2018; Kessler et al., 2017). This subgroup needs mental-health services, but their needs are often unrecognized (Goldstein et al., 2016; Lewis et al., 2019; Wang et al., 2005). Challenges to the detection of PTSD are twofold. First, the "gold-standard" assessment is a diagnostic

Corresponding Author:
Oscar Kjell, Department of Psychology, Lund University
Email: oscar.kjell@psy.lu.se

interview, but these interviews require extensive training, and most medical clinics lack such personnel (Bruchmüller et al., 2011; Ventura et al., 1998). Second, self-report screeners are more scalable but prone to biases, such as social desirability, acquiescence, limited insight, recall biases, and others (Neal et al., 2022; Schuler et al., 2021). Moreover, the lack of objective measures of PTSD symptoms limits the ability of clinicians to monitor treatment progress and identify treatment nonresponders, which presents significant challenges to care. For instance, responders and survivors of the World Trade Center (WTC) attacks are entitled to free mental-health treatment, but 17% have clinically significant PTSD symptoms even 2 decades after the disaster, and rates of PTSD have not decreased (Lowell et al., 2018; Waszczuk et al., 2022).

Recent advances in language analyses based on artificial intelligence have demonstrated promise in addressing this gap and delivering behavioral assessments of mental health (Boyd & Schwartz, 2021; Kjell et al., 2024). However, existing language-based PTSD models have typically been limited to (a) text-based sentiment models (Sawalha et al., 2022) or unigram models (He et al., 2017) rather than large language models to detect PTSD, (b) models being trained to assess other conditions (e.g., AI models for detecting depression, anxiety, and sentiment evaluated against PTSD-symptom severity; Oltmanns et al., 2021; Sawalha et al., 2022; Son et al., 2021), or (c) models trained on nonclinical proxies for PTSD (e.g., public self-disclosures of PTSD; Coppersmith, Dredze, & Harman, 2014; Coppersmith, Harman, & Dredze, 2014; Preotiuc-Pietro et al., 2015; Todorov et al., 2020). Moreover, many of these studies used social media language (e.g., Coppersmith, Harman, & Dredze, 2014; Preotiuc-Pietro et al., 2015; Todorov et al., 2020) rather than data collected in clinical settings, and they are rarely evaluated against external clinical outcomes, such as clinician-assigned PTSD diagnoses or mental-health-care expenditure. The most significant limitation is that none of the existing models were preregistered and tested in a new, prospectively collected sample. Hence, their replicability is unknown, which is especially concerning given recent evidence that AI models can fail to perform when evaluated prospectively (Chekroud et al., 2024; Kernbach & Staartjes, 2022; Spasic & Nenadic, 2020).

Solutions to the "replication crisis" have been developed in many fields (Nosek et al., 2015), but some solutions do not directly translate to AI-based approaches. Whereas standard preregistration practices are suitable for many research goals, they fall short of meeting the needs for developing and evaluating robust models. For example, some preregistration practices require specifying exact data-processing steps, but larger and more complex data sets (e.g., open-ended language) needed for AI-model training often require unexpected bug fixes, several preprocessing steps, and hyper-parameter tuning that come to light only during model-development stages (Maharana et al., 2022; Tabassum & Patil, 2020). Furthermore, 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; Hastie et al., 2001) rather than relying on hypothesis testing—the analytic approach assumed by standard preregistration. In short, AI-model development often involves an iterative refinement process that does not fit well within the standard, a priori preregistration paradigm developed for traditional hypothesis testing.

To address these limitations, we propose a new evaluation paradigm, sequential evaluation with model preregistration (SEMP), that aims to combine good scientific practices (i.e., preregistration) 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., declaring the model code and weights). We first registered the steps for training the models and then registered the models before testing them on new prospective data (see Fig. 1). By registering the models before testing, the approach mitigates the overestimation of accuracies for clinical prediction models by (a) preventing overfitting of hyper-parameters (whether purposeful or accidental); (b) mitigating the risk of test-data leaks—when data from the test set inadvertently influence decisions in the training process, thereby yielding overestimated accuracy for new, unseen data; and (c) when using a prospective test set, more realistically testing the model application to a sample later in time, as would happen in practice.

SEMP integrates established reproducibility conventions—preregistration in psychology (e.g., Nosek et al., 2018), prospective evaluation in clinical prediction (e.g., see Collins, Dhiman, et al., 2024), and locked-algorithms and hidden-test-set frameworks common in machine-learning competitions (shared task; e.g., see Coppersmith et al., 2015)—into a unified, practical workflow tailored to clinical research and models. Specifically, we preregistered the data-cleaning pipeline, feature construction, final model weights, and analysis scripts before accessing a later-in-time test set, thereby minimizing leakage and overfitting and mimicking real clinical settings. This design preserves the freedom of iterative model development before preregistration while enforcing a fully frozen, prospective evaluation phase. Finally, SEMP is aligned

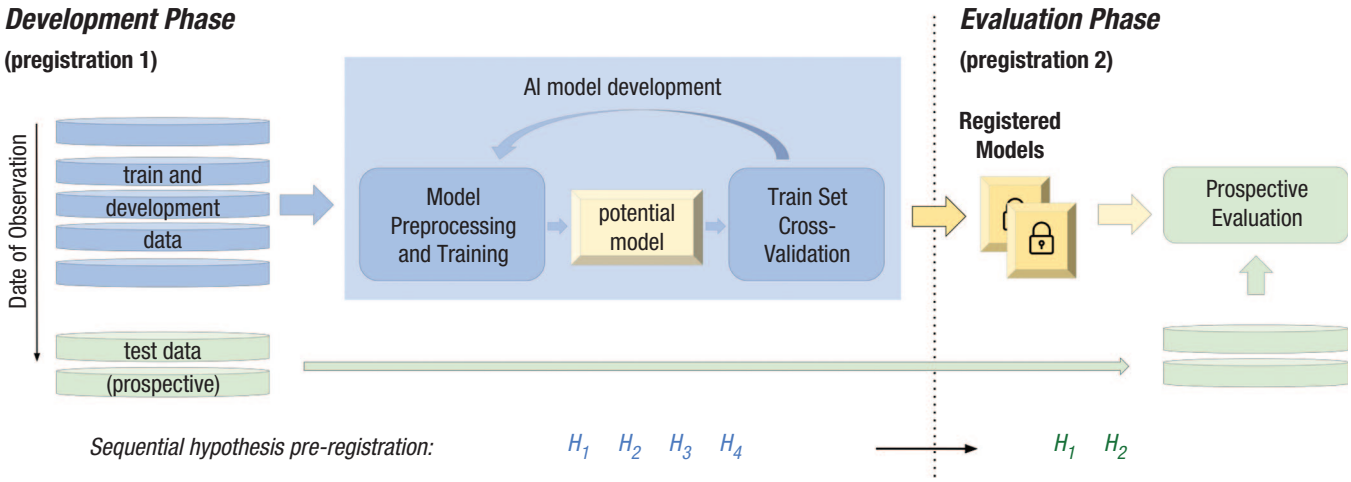

Fig. 1. Overview of the sequential evaluation with model preregistration (SEMP). The stages include data split (i.e., prospective data), preregistering the prediction-model development (e.g., ridge regression), *n*-fold cross-validation (e.g., 10-folds), hypotheses (H1–H4; in which hypotheses can change across stages, as visualized with the different colors), registered prediction models (i.e., the exact models to validate), preprocessing (e.g., removal of parts in which research assistants speak), and held-out perspective set evaluation (e.g., control variables).

with clinical-AI reporting guidance and emphasis on transparency in prediction-modeling research (Collins, Moons, et al., 2024).

In this study, we aimed to advance the application of AI in PTSD assessment through three key objectives. First, we aimed to develop AI models that assess PTSD-symptom severity, including overall severity and the four symptom clusters, from natural spoken language. These models were designed for use in clinical settings with populations exposed to documented trauma. Second, we rigorously evaluated the replicability of these models using the SEMP framework. Third, we validated these models against practitioner-relevant criteria, including PTSD diagnoses from electronic health records and the use of mental-health services. The sample comprises 1,783 WTC responders who completed a health-monitoring visit at the WTC Health Program (a setting akin to primary care) in 2021–2022.

Transparency and Openness

The two preregistration reports and the preregistered language-based assessment models are available on OSF (https://osf.io/ebgp3/?view_only=7f13f644dc594102a1e995729bd44ac3). The open materials are presented in the Supplemental Material available online. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The Institutional Review Board (IRB No. 604113) approved the study at Stony Brook University.

Materials and Method

Design: SEMP

SEMP is a two-phase procedure to preregister models before using held-out test data. The development phase comprises developing the models; the evaluation phase involves a preregistration 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.

In 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 preregistration Phase 1 (the development phase) to produce and preregister pretrained models. In the second phase (the evaluation phase), we applied the preregistered 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 the following 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 cluster dimensions (the preregistered correlations are based on training-set cross-validated $r = .41$, 99% confidence interval [CI] = [.35, .46], $N = 1{,}437$, for the combined-symptom severity and $r = .25$, 99% CI = [.18, .31] for the individual-symptom dimensions, $N = 1{,}422$); and (c) are predictive above and beyond the pretrained 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 preexisting language-based models) that also accompanied the SEMP process (depicted at the bottom of Fig. 1 as "sequential hypothesis preregistration"; for these hypotheses, see Section 3 in the Supplemental Material).

Participants

Participants were recruited from the Stony Brook WTC Health and Wellness Program, where their health has been monitored over several years. Participant data were split into two nontemporally overlapping parts: the "development" data (September 9, 2021–July 29, 2022) and the held-out prospective "evaluation" data (August 1, 2022–September 30, 2022). The development data totaled 1,437 participants (female = 7%, male = 93%; age: $M = 57.9$ years, $SD = 8.0$; 14.5% with reported PTSD diagnosis in their medical record). The prospective data include an additional 346 participants (female = 9%, male = 91%; age: $M = 58.5$ years, $SD = 7.8$; 15.6% with reported PTSD diagnosis in their medical record) enrolled after the participants in the development data set (see Section 1 in the Supplemental Material).

Measures and materials

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 about serious events (e.g., COVID-19 and 9/11; e.g., "How does 9/11 affect you now?"; for all questions, see Section 2 in the Supplemental Material). To maximize the generalizability of content, the questions were aimed at being broad and 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 s answering each question. The questions were the same for everyone in the evaluation tests, although the questions were updated and changed over three iterations of the development phase to increase engagement and elicit more detailed answers. The recording for individuals responding with at least 150 words (the preregistered threshold) took, on average, 7.5 mins ($SD = 4.1$; range = 1.1–43.0).

The PTSD CheckList. The PTSD CheckList (PCL; Blanchard et al., 1996) comprises 17 items assessing PTSD-symptom severity based on the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1994). Respondents are asked to rate symptoms over the past month "in relation to 9/11" using a severity scale from 1 (*not at all*) to 5 (*extremely*). We computed the total mean score and the four subscales (King et al., 1998; Ruggero et al., 2013), 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's alphas were acceptable for all scales in both data sets (≥ .70; see Section 1 in the Supplemental Material).

PTSD diagnosis in medical record. Diagnoses in the medical records are certifications that the participant has been diagnosed with a WTC-related condition (i.e., a WTC-related PTSD diagnosis) at any point since September 11, 2001. The psychiatrists at the Stony Brook WTC Health and Wellness Program diagnosed based on clinical history in the medical records and the semistructured Diagnostic Interview Schedule (see Dasaro et al., 2017).

Mental-health-care-service usage. Mental-health-care-service usage was operationalized as the total cost of services received by a given patient over the past 12 months, extracted from electronic health records of the WTC program.

WTC exposure. WTC exposure was assessed using a clinical interview at the initial monitoring visit (Dasaro et al., 2017). We use 10 dichotomous (yes/no) WTC-exposure variables that were associated with increased risk of PTSD and other health outcomes in prior work (Bromet et al., 2016; Pietrzak et al., 2014; Zvolensky, Farris et al., 2015; Zvolensky, Kotov et al., 2015).

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.

Statistical analysis

The analyses were conducted using the Differential Language Analysis Toolkit (DLATK; Version 26; Schwartz, et al., 2017). Alpha was set at $p < .05$ with Benjamini-Hochberg adjustment to control for false-discovery (Type 1 error) rates. For analyses specific to preprocessing, the development of models, and Preregistration 1, see Section 1 in the Supplemental Material.

Linguistic-feature extractions. Two types of linguistic features were extracted for the mental-health assessment: (a) word embeddings (i.e., numeric representations of words that capture their meaning based on their context) from a large language model (RoBERTa-large, Layer 23; Liu et al., 2019) and (b) topics' ($N = 300$) prevalence scores based on the topic model created in the development data set using latent Dirichlet allocation (LDA; Blei et al., 2003).

Machine learning. Models were developed using cross-validation on the development set. Following the procedure outlined in Preregistration 1, we trained the models using all development data with L2-penalized (ridge) linear regression. L2 regularization is a method that shrinks regression coefficients by applying a penalty to the maximum likelihood parameter estimates, reducing overfitting. Using 10-fold cross-validation, we partitioned the development data into 10 similar-sized subsets (folds). For each fold, the model was trained on nine folds and tested on the remaining fold, ensuring that each subset served as the test set once. This process was repeated across penalties ranging from 10^1 to 10^6 to identify the penalty minimizing prediction error. The penalty that yielded the best performance was then used to develop the final models. We created models for the PCL total score and its four subscales based on three input sets: (a) language only, (b) demographic controls only, and (c) a combination of language and demographic controls.

Preregistered models, lexica, and topics

Preregistered PTSD-severity models. We preregistered models for PCL total score and the four subscales based on (a) only language, (b) only the demographic controls, and (c) language and demographic controls (for more details about how these were developed, see the Supplemental Material).

Preregistered pretrained n*-gram models and word-count lexica of theoretically grounded dimensions.* To quantify the associations between PTSD severity and related theoretically grounded dimensions, we use pretrained *n*-gram models (weighted lexica) trained on Facebook and Twitter language (Park et al., 2015; Schwartz et al., 2014) to predict their self-reported neuroticism, depression, and anxiety. We use word-count lexica, including categories from the Linguistic Inquiry Word Count (Boyd et al., 2022), including death, first-person singular and plural pronouns, and word lengths, because these were significantly related to PTSD in previous research (Son et al., 2021, p. 20). To assess respondents' reexperience of the WTC attack, we selected a lexicon combining five and seven LDA topics relating to reexperiencing the attack.

Open vocabulary topics: preregistered topics (word clouds). After controlling for demographics, we plotted topics significantly associated with PCL scores using DLATK defaults. Three topics are preregistered from analyses of the development data set (for more details, see the Supplemental Material).

Results

Participants' PCL scores did not significantly differ between the development ($M = 26.29$, $SD = 11.7$) and the prospective ($M = 26.04$, $SD = 10.5$) data sets ($t = 0.36$, $df = 1{,}781$, $p = .722$). The proportion of participants with a PCL score ≥ 44 was 10.0% in the development set and 8.4% in the prospective set. Likewise, the mean number of words in the development ($M = 838$, $SD = 629$) and prospective ($M = 776$, $SD = 475$) automated interviews ($t = -1.74$, $df = 1{,}781$, $p = .082$) did not differ significantly between the two data sets.

Language-based assessment of PTSD severity

In the prospective-language data, the preregistered pretrained language-based assessments of PTSD-symptom severity produced scores that significantly correlated with the PCL total scores ($r = .38$; Fig. 2 and Table SM11 in the Supplemental Material) and subscales ($rs = .28$–$.37$). All correlations are above the preregistered cutoffs

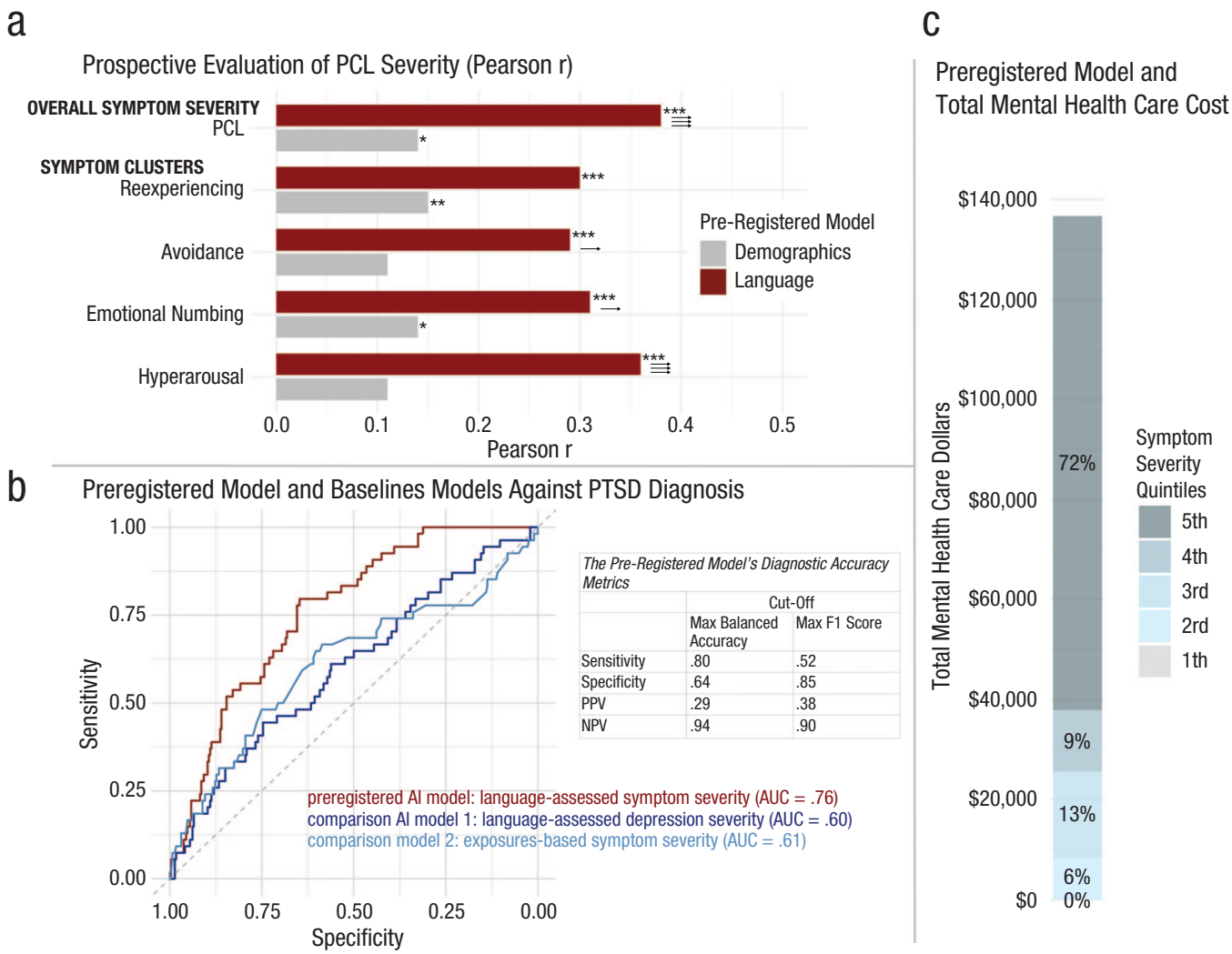

Fig. 2. (a) Models based on embeddings (Layer 23, RoBERTa-large) and topics; demographics = age at visit, gender, occupation (police or not), and race. Asterisks indicate significance, *$p < .05$, **$p < .01$, ***$p < .001$. ↑ = the model accuracy is significantly higher than the corresponding demographics model (i.e., produces a significantly lower error; ↑ = $p < .05$, ↑↑ = $p < .01$, ↑↑↑ = $p < .001$); PCL = the Posttraumatic Stress Disorder (PTSD) CheckList. These results do not include adjustments to account for shrinkage via regularization in the machine-learning models (see Section 3 in the Supplemental Material available onlie). (b) Receiver operating characteristic (ROC) curves and classification-accuracy metrics for PTSD diagnosis in medical records. The preregistered language model (red) significantly outperforms a World Trade Center (WTC) exposure model (light blue; DeLong's test: $Z = 2.70$, $p = .007$, with predictors associated with PTSD in previous research; Bromet et al., 2016; Pietrzak et al., 2014; Zvolensky, Farris et al., 2015; Zvolensky, Kotov et al., 2015) and a depression model (dark blue; DeLong's test: $Z = 3.47$, $p < .001$), which was the most accurate language-based assessment for PTSD severity in Son et al. (2021). (c) The mental-health-care expenditure stratified by language-assessed PTSD-severity quintiles.

based on the cross-validated correlations from the development set. Furthermore, all the PTSD-symptom-severity models based on language and demographics, except for the Re-Experience subscale, produced significantly less error than the preregistered baseline models using only demographics (rs = .10–.15; age, gender, occupation, and race). Overall, the preregistered models yielded correlations in the prospective data set corresponding to the cross-validated correlations in the development training data set (r difference range = .02–.08).

Clinical validation

The preregistered model for PTSD severity was finally validated against participants' PTSD diagnosis from their medical records. For the maximum balanced-accuracy score (.72), the preregistered model for PTSD severity yields a sensitivity of .80 and a specificity of .64. The preregistered model yields an area under the curve (AUC) of .76 (Fig. 2); it significantly outperforms the demographics-based model (AUC = .61, $p = .006$), an exposure-based model (AUC = .61, $p = .007$) with

Table 1. Association of Lexicon (Word-Based) Assessments With PTSD-Symptom Severity

Model, lexical, or topics	Standardized β	
	Development data set (N=1,437)	Prospective data set (only preregistered models) (N=346)
Pretrained *n*-gram models		
Anxiety (Son et al., 2021) (+)	.17***	.16**
Depression (Son et al., 2021) (+)	.14***	.18**
Neuroticism (+)	.11***	—
Word-count lexica		
First-person singular pronouns ("I," "me," "my") (+)	.05	—
First-person plural pronouns ("we," "our") (–)	–.04	—
Word lengths (–)	*–.02*	—
LIWC2022 death (+)	*.02*	—
Hypothesized topics in Preregistration 1		
Reexperiencing the WTC attack 1 (+)	.06*	—
Reexperiencing the WTC attack 2 (+)	.05	—
Open vocabulary topics		
Topic 288 (on "stress, anxiety, and "pain")	.22***	.17***
Topic 286 (on "control")	.16***	.13***
Topic 65 (on "mental health issues")	.13***	.10**

Note: β coefficients are from a standardized linear regression with demographics as covariates. (+) = positive associated registered; (–) = negative associateion registered;— = not preregistered; reexperiencing the WTC attack 1 (+) = five topics; reexperiencing the WTC attack 2 (+) = seven topics; demographics = age at visit, gender, occupation (police or not), and race; PTSD = posttraumatic stress disorder; WTC = World Trade Center.
*p<.05. **p<.01. ***p<.001.

predictors related to PTSD in previous research (Bromet et al., 2016; Pietrzak et al., 2014; Zvolensky, Farris et al., 2015; Zvolensky, Kotov et al., 2015), a model combining demographics and exposure (AUC = .61, p < .005), and a depression *n*-gram model (AUC = .60, p < .001).

Furthermore, the AI-based measure demonstrated significant incremental validity over PCL symptom severity with respect to external criterion of healthcare expenditure. A multiple linear regression analysis was conducted to examine the contributions of PCL ratings and language-assessed PTSD severity (both standardized) in predicting mental-health-care expenditure. The model was statistically significant overall (p < .001). For each standard-deviation increase in language-assessed PTSD severity, expenditure rose by approximately $696.50 ($p$ < .001). This relationship was stronger than the one observed with standardized PCL scores, which accounted for only a $254.10 increase and was not statistically significant (p = .183). Note that individuals in the top quintile of language-assessed PTSD-severity scores accounted for 72% of total mental-health-care expenditure.

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 1). The difference in correlation between the development and prospective data sets ranges from .01 to .04, demonstrating that the relationship is robust but small.

In addition, the three preregistered topics were also significantly associated with the PCL scores (Fig. 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 data sets ranged from 0.03 to 0.05, showing that the topic model generalized to future data by producing reliable results.

Discussion

In the present study, we developed the first replicated language-based AI models of PTSD symptoms—covering overall severity and the four symptom clusters—and introduced the SEMP framework for their prospective

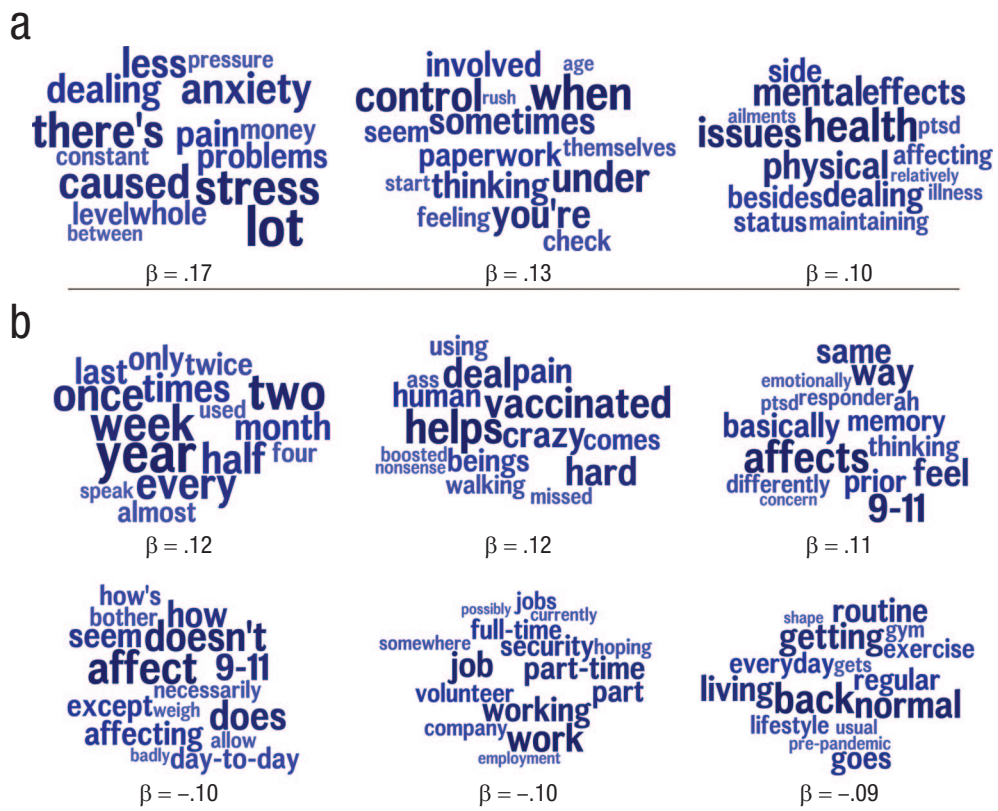

Fig. 3. Topics (automatically grouped similar words) from the automated interview significantly related to posttraumatic-stress-disorder severity. Scores are standardized linear coefficients (beta) controlling for demographics. Topics under (a) were preregistered, and those in (b) were also significant in the development set but did not meet the power-analysis criteria to be tested in the smaller test set.

evaluation. The model was trained in a sample of 1,437 WTC responders to assess PCL scores and evaluated in another sample of 346 responders. We used the SEMP design to ensure rigorous replication and found that models perform equally well in the prospective held-out sample as in the sample they were trained in. Moreover, the models showed external validity in explaining PTSD diagnosis in medical records and the use of mental-health-care services.

According to the SEMP design, the models were preregistered before the evaluation was conducted in the prospective sample so that no changes could be made to the models, addressing a common concern about overfitting in machine-learning models for clinical prediction (Chekroud et al., 2024). This test is particularly stringent compared with typical retrospective validation practices, which often do not evaluate model generalizability in a new prospective sample (Son et al., 2021). Nevertheless, we found that each model produced predictions that correlated with its target (Re-Experiencing, Avoidance, Numbing, and Hyperarousal subscales and PCL total), in line with the preregistered cutoffs. The SEMP procedure enables precise hypotheses regarding the effect size (which is uncommon in current practice). Furthermore, all the assessment models based on language and demographics, except for reexperiencing, produced significantly less error than the baseline models based on only demographics.

Importantly, the preregistered models were subjected to external validation against clinically salient criteria derived from medical records and thus methodologically independent from the collected language. Language-assessed PTSD severity discriminated patients with PTSD diagnosis with an AUC of .76, which is significantly higher than for other types of predictors, including demographics, exposures, and a previous state-of-the-art depression-severity model (Son et al., 2021). This AUC score was well above the cutoff for common clinical standards (.70; e.g., see the consensus-based standards for the selection of health measurement

instruments, Prinsen et al., 2018). Furthermore, language-assessed PTSD severity demonstrated incremental validity over self-reported PCL ratings. When both were included as predictors of mental-health-service usage, only language-assessed PTSD severity was statistically significant, and the top quintile accounted for more than 70% of total mental-health-care expenditures.

Note that natural-language questions were not designed to maximize convergence with PTSD-symptom measures, in which patients would be asked to describe their symptoms. Rather, patients were asked more natural questions to speak freely about their past and present and their aspirations in life. Thus, the associations observed reflect psychological signals embedded in more spontaneous self-expression than explicit symptom reporting. For such contexts of comparing behavior with a psychological score, the modal effect size is between 0.1 and 0.4 (Meyer et al., 2001; Roberts et al., 2007). This puts the model's convergent validity at the upper end ($r = .38$, AUC = .76), representing a substantial, meaningful effect, comparable with correlations typically observed between naturalistic behaviors and psychological constructs (Baumeister et al., 2007).

Pretrained *n*-gram models for depression and anxiety were positively correlated with observed PTSD severity. Furthermore, the three preregistered topics were significantly related to PTSD severity. Greater symptoms were associated with the use 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 into the specific language related to PTSD severity. They are consistent with research showing that a stressful life contributes to the exacerbation of these symptoms (Gold et al., 2005; Zvolensky, Farris et al., 2015; Zvolensky, Kotov et al., 2015) and with evidence of high comorbidity between PTSD, depression, anxiety, and pain (Brady et al., 2000; Caramanica et al., 2014; Sharp & Harvey, 2001).

Implications

With further development, language-based AI models can have multiple potential applications. First, AI can facilitate clinical research by offering systematic, fine-grained, and reliable measures of PTSD. For example, these behavioral markers may serve as complementary outcome metrics in randomized clinical trials (e.g., to address placebo effects) or as targets for studies investigating the neural underpinnings of PTSD. In addition, techniques to understand the performance of AI in predicting PTSD from language can help to elucidate psychological mechanisms underpinning this disorder (Boyd & Schwartz, 2021).

Second, language-based AI can be deployed as a screener. Existing screeners, such as the PCL, are short and easy to administer. Nevertheless, they pose a certain burden on patients and clinicians and as a result, are not routinely administered in many medical settings (Greene et al., 2016). In contrast, AI-based language assessments can be seamlessly integrated into telemedicine (Haleem et al., 2021), unobtrusively estimating the likelihood of PTSD after a session. High language scores may trigger the administration of a traditional screener or the referral to a mental-health provider for an evaluation. Evidence is emerging that AI-based screening may improve patient outcomes (Rollwage et al., 2024).

Third, once PTSD treatment is initiated, language-based AI can be used to monitor treatment response objectively, estimating symptom severity after each session. These AI assessments would avoid biases inherent in self-report measures (Neal et al., 2022; Schuler et al., 2021) and offer clinical utility over these scales, as we found in the present incremental-validity analyses. Implementing AI is not limited to telehealth and can be performed with any audio-recorded visits. Certainly, the use of AI in clinical settings should be combined with best practices for ensuring patient privacy, informed consent, transparency, and data security to form a responsible and ethical implementation of AI in mental-health care (for a more elaborate discussion on responsible AI and ethics, see Peters et al., 2020).

Fourth, by introducing the SEMP study design, we provide a rigorous framework for developing and evaluating clinical AI models, helping to safeguard results against accuracy overestimations and poor generalization. Accuracy estimations are crucial when implementing clinical AI models (Prinsen et al., 2018). The SEMP design could play a critical role in clinical AI assessments similar to the randomized-controlled-trial design used to evaluate interventions. These reliable accuracy estimates would be combined with best practices for ensuring patient privacy, informed consent, transparency, and data security to form a responsible and ethical implementation of AI in mental-health care (for a more elaborate discussion on responsible AI and ethics, see Peters et al., 2020).

Limitations

Although 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 in a relatively homogeneous population and linguistic context—responders to the WTC attacks comprising mainly males, sharing similar occupational and trauma histories, and all assessed

through a standardized interview. Such sample homogeneity may reduce linguistic variability and could inflate model performance when tested in other populations or settings. Furthermore, the context in which language is elicited (e.g., intake vs. follow-up sessions, in person vs. telehealth, structured vs. open interviews) can influence both how patients express symptoms and how language-based AI models perform. Occupational role was the only available proxy for socioeconomic status (SES); direct measures such as income, education, or composite SES indices were not collected, which limits our ability to characterize socioeconomic diversity in the sample and assess the generalizability of the findings across SES subgroups. Therefore, future research should evaluate and validate model performance across more diverse populations, trauma types, and language contexts.

Second, motivated to prompt language generalizable to multiple outcomes, interview questions concerned participants' lives and experiences overall rather than specific clinical symptoms (see e.g., Kjell et al., 2019, 2022), but further work is needed to verify if this language actually generalizes better. Third, the analyses are based on cross-sectional evaluations and do not examine the sensitivity to change in treatment effects. Finally, because the models were trained to self-reported PTSD severity, they may partly reflect biases inherent in self-report measures (e.g., social desirability, recall errors, limited insight). Although other options for PTSD assessment exist (e.g., the Clinician-Administered PTSD Scale; Weathers et al., 2018), our validation against PTSD diagnosis from medical records and service utilization mitigates this limitation.

Conclusions

This is the first preregistered and prospectively validated language-based model of PTSD severity—including both total- and cluster-level-severity assessments—developed from clinically collected, automated interview data in the unified SEMP workflow. Our preregistered PTSD-severity models demonstrated reliable and accurate assessments in prospective data, showing convergent validity with self-report and external validity with PTSD diagnoses from medical records and mental-health-service usage. The accuracy of our language-based 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.

Transparency

Action Editor: Jennifer Lau
Editor: Jennifer L. Tackett

Author Contributions

Oscar Kjell: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

Adithya V. Ganesan: Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing.

Ryan L. Boyd: Methodology; Writing – original draft; Writing – review & editing.

Joshua Oltmanns: Conceptualization; Methodology; Writing - review & editing.

Alfredo Rivero: Data curation; Formal analysis; Writing – review & editing.

Scott Feltman: Data curation; Investigation; Methodology; Writing – review & editing.

Melissa A. Carr: Data curation; Investigation; Writing – review & editing.

Jorge Alves: Data curation; Methodology; Writing – review & editing.

Benjamin Luft: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Writing – review & editing.

Roman Kotov: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing.

H. Andrew Schwartz: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

O. Kjell cofounded and holds shares in a start-up that uses language-based assessments to diagnose mental-health problems. The authors report no additional biomedical financial interests or potential conflicts of interest.

Funding

This work was supported in part by Grant U01 OH012476 from Centers for Disease Control and Prevention (CDC/NIOSH), Developing and Evaluating Artificial Intelligence-Based Longitudinal Assessments of PTSD in 9/11 Responders; a grant from the National Institute for Alcohol Abuse & Alcoholism (NIH-NIAAA) (R01 AA028032) awarded to H. A. Schwartz at Stony Brook University; and a National Alliance for Research on Schizophrenia and Depression (NARSAD) Young Investigator Grant to J. Oltmanns from the Brain & Behavior Research Foundation.

Open Practices

This article has received the badge for Open Materials and Preregistration. More information about the Open Practices

badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Oscar Kjell https://orcid.org/0000-0002-2728-6278
Adithya V. Ganesan https://orcid.org/0000-0001-6179-624X
Ryan L. Boyd https://orcid.org/0000-0002-1876-6050
Joshua Oltmanns https://orcid.org/0000-0001-6670-6995
Melissa A. Carr https://orcid.org/0000-0002-0870-6628
Jorge Alves https://orcid.org/0009-0005-7941-2204
Benjamin Luft https://orcid.org/0000-0001-9008-7004
Roman Kotov https://orcid.org/0000-0001-9569-8381

Acknowledgments

We express our gratitude to the rescue and recovery workers of the World Trade Center (WTC) 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.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/21677026261439026

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403.

Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., & Forneris, C. A. (1996). Psychometric properties of the PTSD Checklist (PCL). *Behaviour Research and Therapy*, *34*(8), 669–673.

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

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

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41.

Brady, K. T., Killeen, T. K., Brewerton, T., & Lucerini, S. (2000). Comorbidity of psychiatric disorders and posttraumatic stress disorder. *Journal of Clinical Psychiatry*, *61*, 22–32.

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

Bruchmüller, K., Margraf, J., Suppiger, A., & Schneider, S. (2011). Popular or unpopular? Therapists' use of structured interviews and their estimation of patient acceptance. *Behavior Therapy*, *42*, 634–643.

Caramanica, K., Brackbill, R. M., Liao, T., & Stellman, S. D. (2014). Comorbidity of 9/11-related PTSD and depression in the World Trade Center Health Registry 10–11 years postdisaster. *Journal of Traumatic Stress*, *27*(6), 680–688. https://doi.org/10.1002/jts.21972

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, *383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Jr., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (Part 1): From development to external validation. *The BMJ*, *384*, Article e074819. https://doi.org/10.1136/bmj-2023-074819

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *The BMJ*, *385*, Article q902. https://doi.org/10.1136/bmj.q902.

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). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-32

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 31–39). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-12

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post-traumatic stress disorder in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 579–582. https://doi.org/10.1609/icwsm.v8i1.14574

Dasaro, C. R., Holden, W. L., Berman, K. D., Crane, M. A., Kaplan, J. R., Lucchini, R. G., Luft, B. J., Moline, J. M., Teitelbaum, S. L., & Tirunagari, U. S. (2017). Cohort profile: World Trade Center health program general responder cohort. *International Journal of Epidemiology*, *46*(2), Article e9. https://doi.org/10.1093/ije/dyv099

Galatzer-Levy, I. R., Huang, S. H., & Bonanno, G. A. (2018). Trajectories of resilience and dysfunction following

potential trauma: A review and statistical evaluation. *Clinical Psychology Review*, *63*, 41–55.

Gold, S. D., Marx, B. P., Soler-Baillo, J. M., & Sloan, D. M. (2005). Is life stress more traumatic than traumatic stress? *Journal of Anxiety Disorders*, *19*(6), 687–698.

Goldstein, R. B., Smith, S. M., Chou, S. P., Saha, T. D., Jung, J., Zhang, H., Pickering, R. P., Ruan, W. J., Huang, B., & Grant, B. F. (2016). The epidemiology of DSM-5 posttraumatic stress disorder in the United States: Results from the National Epidemiologic Survey on Alcohol and Related Conditions-III. *Social Psychiatry and Psychiatric Epidemiology*, *51*, 1137–1148.

Greene, T., Neria, Y., & Gross, R. (2016). Prevalence, detection and correlates of PTSD in the primary care setting: A systematic review. *Journal of Clinical Psychology in Medical Settings*, *23*, 160–180.

Haleem, A., Javaid, M., Singh, R. P., & Suman, R. (2021). Telemedicine for healthcare: Capabilities, features, barriers, and applications. *Sensors International*, *2*, Article 100117. https://doi.org/10.1016/j.sintl.2021.100117

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-21606-5

He, Q., Veldkamp, B. P., Glas, C. A., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*, *24*(2), 157–172.

Kernbach, J. M., & Staartjes, V. E. (2022). Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and overfitting. In V. E. Staartjes, L. Regli, & C. Serra (Eds.), *Machine learning in clinical neuroscience* (Vol. 134, pp. 15–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-85292-4_3

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., Degenhardt, L., de Girolamo, G., Dinolova, R. V., Ferry, F., Florescu, S., Gureje, O., Haro, J. M., Huang, Y., Karam, E. G., Kawakami, N., Lee, S., Lepine, J. P., Levinson, D., & Koenen, K. C. (2017). Trauma and PTSD in the WHO world mental health surveys. *European Journal of Psychotraumatology*, *8*(Suppl. 5), Article 1353383. https://doi.org/10.1080/20008198.2017.1353383

King, D. W., Leskin, G. A., King, L. A., & Weathers, F. W. (1998). Confirmatory factor analysis of the Clinician-Administered PTSD Scale: Evidence for the dimensionality of posttraumatic stress disorder. *Psychological Assessment*, *10*(2), 90–96. https://doi.org/10.1037/1040-3590.10.2.90

Kjell, O., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

Kjell, O., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Research*, *333*, Article 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kjell, O., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-07520-w

Lewis, S. J., Arseneault, L., Caspi, A., Fisher, H. L., Matthews, T., Moffitt, T. E., Odgers, C. L., Stahl, D., Teng, J. Y., & Danese, A. (2019). The epidemiology of trauma and post-traumatic stress disorder in a representative cohort of young people in England and Wales. *The Lancet Psychiatry*, *6*(3), 247–256.

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. https://doi.org/10.48550/arXiv.1907.11692

Lowell, A., Suarez-Jimenez, B., Helpman, L., Zhu, X., Durosky, A., Hilburn, A., Schneier, F., Gross, R., & Neria, Y. (2018). 9/11-related PTSD among highly exposed populations: A systematic review 15 years after the attack. *Psychological Medicine*, *48*(4), 537–553.

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128–165.

Neal, T., Lienert, P., Denne, E., & Singh, J. P. (2022). A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and Human Behavior*, *46*(2), 99–120. https://doi.org/10.1037/lhb0000482

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Oltmanns, J. R., Schwartz, H. A., Ruggero, C., Son, Y., Miao, J., Waszczuk, M., Clouston, S. A., Bromet, E. J., Luft, B. J., & Kotov, R. (2021). Artificial intelligence language predictors of two-year trauma-related outcomes. *Journal of Psychiatric Research*, *143*, 239–245.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020.

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, *1*(1), 34–47.

Pietrzak, R. H., Feder, A., Singh, R., Schechter, C. B., Bromet, E. J., Katz, C. L., Reissman, D. B., Ozbay, F., Sharma,

V., & Crane, M. (2014). Trajectories of PTSD risk and resilience in World Trade Center responders: An 8-year prospective cohort study. *Psychological Medicine*, *44*(1), 205–219.

Preotiuc-Pietro, D., Sap, M., Schwartz, H. A., & Ungar, L. H. (2015). Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 40–45). Association for Computational Linguistics.

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345.

Rollwage, M., Habicht, J., Juchems, K., Carrington, B., Hauser, T. U., & Harper, R. (2024). Conversational AI facilitates mental health assessments and is associated with improved recovery rates. *BMJ Innovations*, *10*, 4–12.

Ruggero, C. J., Kotov, R., Callahan, J. L., Kilmer, J. N., Luft, B. J., & Bromet, E. J. (2013). PTSD symptom dimensions and their relationship to functioning in World Trade Center responders. *Psychiatry Research*, *210*(3), 1049–1055.

Sawalha, J., Yousefnezhad, M., Shah, Z., Brown, M. R., Greenshaw, A. J., & Greiner, R. (2022). Detecting presence of PTSD using sentiment analysis from text data. *Frontiers in Psychiatry*, *12*, Article 811392. https://doi.org/10.3389/fpsyt.2021.811392

Schuler, K., Ruggero, C. J., Mahaffey, B., Gonzalez, A. L., Callahan, J., Boals, A., Waszczuk, M. A., Luft, B. J., & Kotov, R. (2021). When hindsight is not 20/20: Ecological momentary assessment of PTSD symptoms versus retrospective report. *Assessment*, *28*(1), 238–247. https://doi.org/10.1177/1073191119869826

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014). Towards assessing changes in degree of depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 118–125). Association for Computational Linguistics.

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017). DLATK: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 55–60). Association for Computational Linguistics.

Sharp, T. J., & Harvey, A. G. (2001). Chronic pain and posttraumatic stress disorder: Mutual maintenance? *Clinical Psychology Review*, *21*(6), 857–877.

Son, Y., Clouston, S. A., Kotov, R., Eichstaedt, J. C., Bromet, E. J., Luft, B. J., & Schwartz, H. A. (2021). World Trade Center responders in their own words: Predicting PTSD symptom trajectories with AI-based language analyses of interviews. *Psychological Medicine*, *53*(3), 918–926.

Spasic, I., & Nenadic, G. (2020). Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics*, *8*(3), Article e17984. https://doi.org/10.2196/17984

Tabassum, A., & Patil, R. R. (2020). A survey on text preprocessing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology*, *7*(6), 4864–4867.

Todorov, G., Mayilvahanan, K., Cain, C., & Cunha, C. (2020). Context- and subgroup-specific language changes in individuals who develop PTSD after trauma. *Frontiers in Psychology*, *11*, Article 989. https://doi.org/10.3389/fpsyg.2020.00989

Ventura, J., Liberman, R. P., Green, M. F., Shaner, A., & Mintz, J. (1998). Training and quality assurance with the Structured Clinical Interview for DSM-IV (SCID-I/P). *Psychiatry Research*, *79*(2), 163–173.

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Twelve-month use of mental health services in the United States: Results from the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 629–640.

Waszczuk, M. A., Docherty, A. R., Shabalin, A. A., Miao, J., Yang, X., Kuan, P.-F., Bromet, E., Kotov, R., & Luft, B. J. (2022). Polygenic prediction of PTSD trajectories in 9/11 responders. *Psychological Medicine*, *52*(10), 1981–1989.

Weathers, F. W., Bovin, M. J., Lee, D. J., Sloan, D. M., Schnurr, P. P., Kaloupek, D. G., Keane, T. M., & Marx, B. P. (2018). The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): Development and initial psychometric evaluation in military veterans. *Psychological Assessment*, *30*(3), 383–395. https://doi.org/10.1037/pas0000486

Zvolensky, M. J., Farris, S. G., Kotov, R., Schechter, C. B., Bromet, E., Gonzalez, A., Vujanovic, A., Pietrzak, R. H., Crane, M., & Kaplan, J. (2015). World Trade Center disaster and sensitization to subsequent life stress: A longitudinal study of disaster responders. *Preventive Medicine*, *75*, 70–74.

Zvolensky, M. J., Kotov, R., Schechter, C. B., Gonzalez, A., Vujanovic, A., Pietrzak, R. H., Crane, M., Kaplan, J., Moline, J., Southwick, S. M., Feder, A., Udasin, I., Reissman, D. B., & Luft, B. J. (2015). Post-disaster stressful life events and WTC-related posttraumatic stress, depressive symptoms, and overall functioning among responders to the World Trade Center disaster. *Journal of Psychiatric Research*, *61*, 97–105.