**SUPPLEMENTARY INFORMATION**

**Supplementary Materials and Methods**

*Data labeling: visual and auditory markers of arousal and mood*

Facial Features of arousal and mood: Facial expression of emotion were coded based on visible facial movements. Facial features corresponding to action units (AUs) identified by the Facial Action Coding System (FACS) 8were labeled from raw MP4 video files using the OpenFace package in python 13, which has a confidence score that is more than 75% for face detection. The extracted raw features were used to compute a Facial Expressivity Score, Expressivity score for each emotion (Happy, Sad, Anger, Disgust, Surprise, Fear, Contempt), and Peak Expressivity (1, 3, 6, 9, 12, 15) sec windows. We calculated:

* Normalized EMFACS emotions (per frame):Facial landmark analysis generates a series of AUs that are coded according to the Emotional Facial Action Coding System (EMFACS) to label emotions in CDX videos. The EMFACS emotions are composite scores driven by between 2 and 7 individual AUs. Each action unit has a range of 0.0000 to 5.0000. To normalize each emotion to a value between 0.0000 and 1.0000, the composite AU scores are added and then divided by the total number of action units involved in coding the emotion. The composite equation is as follows:  
  *Normalized EMFACS emotion = (****Σ*** *AU\_corr) / AU\_Tot*
* Facial Expressivity Index (FEI)**:** A composite view of general expressivity derived from a subset of AUs. The FEI averages all of the normalized expressivity scores for EMFACS emotions and determines whether a participant is eliciting any emotion, rather than a “correct” or relevant emotion to the stimulus presented. The FEI functions independently of valence of the stimulus or question presented.
* Peak Count: Leveraging the peak count from the normalized emotion windows above, a composite variable called Expressivity Peak Count is derived from adding together all of the peaks counted across the specified time window for the seven EMFACS coded emotions. This gives a time-filtered view of composite expressivity using the following equation.

*Expressivity Peak Count (window) = Normalized EMFACS happiness (window)+  
 Normalized EMFACS sadness (window) +  
 Normalized EMFACS surprise (window)+  
 Normalized EMFACS fear (window) +  
 Normalized EMFACS anger (window) +  
 Normalized EMFACS disgust (window) +  
 Normalized EMFACS contempt (window)*

Voice prosody features of arousal: For verbal analysis, PRAAT Software python library Parsel-mouth was used. We analyzed the following parameters:

* Audio Expressivity Index: Sum of Intensity, Formant, Pitch Variability and NAQ  
   *AEI\_x = IV\_x\_zsc + PV\_x\_zsc + NAQ\_x\_zsc + FV\_x\_zsc*
* Audio Intensity(dB): Intensity(dB) of Sound in air.   
   *AI = 10 log10 { 1 / (T P02) ∫dt x2(t) }21*
* Fundamental Frequency(hz): Lowest Frequency of the periodic waveform.  
   *rx (τ) ≈ rxw (τ) / rw (τ)*
* Harmonic Noise Ratio: The degree of acoustic periodicity of audio wave, the ratio between periodic and aperiodic components.  
   *HNR = 10\*log10 { ACV(t)/ (ACV(0) - ACV(t))}*
* Glottal to Noise Excitation Ratio: The ratio between the noise generated by vocal fold and vocal tract.  
   *GNE = Noise by vocal folds / Noise by vocal tract*
* Voice Frame Score: Audio frame where participants are speaking with respect to the entire audio segment.  
   *VFS = (VF\_Speech / VF\_Tot) \* 100*
* Formant Frequency Variability (FV): Formant Frequency isclustering of acoustic energy at a particular frequency in audio stream. ***FV\_x\_zsc*** is a Z-score of *Formant Variability* of an individual from a single question. ***FV\_x*** Formant variability of one individual for a single question, ***FV\_mean*** is the mean of Formant variability from all participants across all questions, ***FV\_std*** is standard deviation from all individuals across all questions  
   *FV\_x\_zsc = (FV\_x – FV\_mean) / FV\_std*
* Intensity Variability (IV): Variability of audio intensity for an individual during the audio stream. ***IV\_x\_zsc*** is a Z-score of *IV* of an individual from a single question. ***IV\_x*** Intensity variability of one individual for a single question, ***IV\_mean*** is mean of Intensity variability from all participant across all questions, ***IV\_std*** is standard deviation from all individuals across all questions  
   *IV\_x\_zsc = (IV\_x – IV\_mean) / IV\_std*
* Pitch Variability (PV): Variation in fundamental frequency frame by frame. ***PV\_x\_zsc*** is a Z-score of *PV* of an individual from a single question. ***PV\_x*** Pitch variability of one individual for a single question, ***PV\_mean*** is mean of Pitch variability from all participant across all questions, ***PV\_std*** is standard deviation from all individuals across all questions  
   *PV\_x\_zsc = (PV\_x – PV\_mean) / PV\_std*
* Normalized Amplitude Quotient (NAQ): It is the ratio between ac flow amplitude and negative peak amplitude of ac flow and then, normalized w.r.t. to period. ***NAQ\_x\_zsc*** is a Z-score of *NAQ* of an individual from a single question. ***PV\_x*** NAQ of one individual for a single question, ***PV\_mean*** is mean of NAQ from all participant across all questions, ***PV\_std*** is standard deviation from all individuals across all questions.  
   *NAQ\_x\_zsc = (NAQ\_x – NAQ\_mean) / NAQ\_std*

Speech content features of arousal and mood: Speech content was extracted with natural language processing (NLP) using Receptiviti which uses the LIWC 2015 dictionary30. Extracting features include for instance, Summary Language variables, Linguistic Dimensions, Psychological, Social, Cognitive, Perceptual and Biological Processes. Furthermore, we extracted content using DeepSpeech, which is an open source pre-trained neural network model to extract Text from Speech. This identifies features like rate of speech, Intent Expressivity, Emotion label, Word Repetition out of it.

* Intent Expressivity: The emotional expressivity (positive, negative, or neutral) is measured using NLP whereby word use is identified using speech detection and the emotional value and strength of the word is compared to a standardized library (Sentiment Analysis API- Text Processing) and then labeled appropriately for each word. Variables are then calculated for as an average strength of words of a labeled valence (i.e. negative, neutral, or positive).
* Emotion Label: The emotion label is the specific emotion associated with each word identified using speech detection. The emotional value of the word Is compared to a standardized library and then labeled appropriately for each word. Variables are then calculated for as a count of words of a labeled emotion (e.g. fear, anger, happiness, etc.).
* Rate of Speech: Rate of speech was calculated as the number of words identified with NLP given a time segment.
* Word Repetition**:** Word repetition is the count of words that are repeated in a given segment of time.

Movement features of arousal: Movement variables was extracted from raw MP4 video files using the OpenFace package in python 13

* Head Movement: Overall body motion within the scene is calculated by observing the 3D position of the patient’s head relative to the camera as a surrogate. The working hypothesis for this item is if a patient’s head position moves from left to right or toward and away from the camera, the patient’s body is doing the same. Expressive gestures are quantified by analyzing the slope of 3 sources of head motion within the frame. The change in that position over each frame can be measured as frame by frame movement of head.  
   *Head\_Velocity = Sum(distance.euclidean(translation(x1,y1,z1),translation(x2,y2,z2)))*
* Attentiveness : Attention of participant towards camera within the video frame is measured by analyzing 3D position of the patient’s eye gaze direction. Attentiveness is estimated by examining the eye gaze direction coordinates averaged for both eyes. The working hypothesis for this item is if patient’s eye moves left-right side gaze angle for x-coordinate(gaze\_angle\_x) is from positive to negative , if eye moves up-down gaze angle for y-coordinate(gaze\_angle\_y) is from negative to positive, coordinates for both angles will be closer to 0 if patient is looking straight towards camera.
* Pupil dilation rate: Change in pupil size within the video frames is measured by analyzing 2D location of patient eye region landmarks. Pupil dilation rate is estimated by analyzing the change in pupil area over the frame segments using pupil landmarks.  
   *Pupil\_diameter = distance.euclidean(pupil(x1,y1),pupil(x2,y2))*  
   *Pupil\_area = pi \* ((Pupil\_diameter/2)\*\*2)*

Supplementary Table 1. Classification task and observed positive and negative events for each outcome.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification task** | | **Training set** | **Test set** | **Total** |
| **PCL-5 cut-off score** | Provisional PTSD | n = 19 | n = 4 | N=23 |
| No PTSD | n = 37 | n = 15 | N=52 |
| **CES-D cut-off** | Provisional MDD | n = 19 | n = 9 | N=28 |
| **sore** | No MDD | n = 31 | n = 8 | N=39 |

a)



b)



Supplementary Figure 1. Predicted and observed values for predicting symptoms severity for (a) PCL-5 and (b) CES-D.