Multimodal features for multi-perspective assessment of working alliance in therapist-patient psychotherapy

Rivka Vollebregt
University of Amsterdam
Amsterdam, the Netherlands
r.vollebregt@uva.nl

Lennard Bornemann
Utrecht University
Utrecht, the Netherlands

Sanne J.E. Bruijniks
Utrecht University
Utrecht, the Netherlands
s.j.e.bruijniks@uu.nl

Albert Ali Salah
Utrecht University
Utrecht, the Netherlands
a.a.salah@uu.nl

Abstract

The rapport between a therapist and a patient is called the working alliance, and monitoring it can help in preventing ruptures and drop-outs. While working alliance is typically assessed via self-reported measures, AI-based analysis tools may provide potential alternatives to assist in this task and provide insights. In this study, we use machine learning based approaches to extract multimodal working alliance features from videos of actual therapy sessions (n = 114) of patients with depression. We explore the connections between five modalities (i.e. conversational and movement features, affect from speech, text, and face) and the working alliance as reported by the patient, the therapist, and independent observers, respectively. We show that these perspectives overlap, but also show important differences. We incorporate state of the art off-the-shelf tools into a simple classification pipeline to identify several features as good predictors of working alliance, and we make our code base available for replication.

CCS Concepts

• **Applied computing** → **Psychology**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

Keywords

Working alliance, social signal processing, deep learning, multi-modality, face analysis, text analysis, conversational features

ACM Reference Format:
Rivka Vollebregt, Lennard Bornemann, Sanne J.E. Bruijniks, and Albert Ali Salah. 2026. Multimodal features for multi-perspective assessment of working alliance in therapist-patient psychotherapy. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3776574.3831165

ICMI '26, Napoli, Italy

ACM ISBN 979-8-4007-2318-6/2026/10
https://doi.org/10.1145/3776574.3831165

1 Introduction and Background

Working alliance (WA) is a widely-used factor that measures the quality of the relationship between therapist and patient in psychotherapy [14, 17], and is often defined as the agreement on treatment goals, the tasks that are worked on and the bond between the therapist and patient [22]. More specifically, *agreement on tasks* has been defined as the therapist's and patient's shared understanding of the tasks, *agreement on goals* as the therapist's and patient's shared understanding of the goals and objectives, and the *bond* as the emotional connection between the therapist and patient, consisting of mutual trust, acceptance and confidence [22].

Research on the working alliance has been repeatedly associated with treatment outcome and showed to mediate the relationship between treatment and outcome across different treatment modalities and in different patient populations [5, 17]. For example, improvement in working alliance predicted depressive symptoms in the next session [16]. In addition, recent studies distinguish between the trait and state part of the working alliance and suggest that the role of trait and state alliance may differ between different forms of psychotherapy. While *trait alliance* can be seen as the general ability of the patient and therapist to form interpersonal relationships, *state alliance* is the weakening or strengthening of alliance within and between sessions [50]. Several studies already demonstrated that trait alliance may help to predict who benefits more from treatment, while strengthening of state alliance has been related to better treatment outcome [23]. Accurately detecting the working alliance during the session may help to detect and repair weakening or poor alliance during the sessions [49].

A limitation to current research on working alliance is that most studies measure working alliance with the use of self-report questionnaires. While earlier research focused on patient's assessments, mostly at the beginning of the treatment, a general direction in the field is to measure the interplay and dynamic features between the therapist and the patient, including ruptures and repair [38]. A widely used measure of the working alliance in psychotherapy is the Working Alliance Inventory (WAI), which is a self-report questionnaire that assesses perceptions of the therapeutic alliance in three areas: goals, tasks, and bond, respectively [9, 22]. Although the original version consisted of 36 items, nowadays most often the shorter version of 12 items (WAI-S) is used [36, 37]. There are different versions of the WAI, filled in by the therapist (WAI-T), the patient (WAI-P) or an observer (WAI-O). While the WAI-P and WAI-T are completed to the perception of the therapist and patient of the working alliance, the observer ratings are conducted by an independent observer trained to detect and rate strength and weaknesses in the working alliance. Other measures, such as the Therapist Alliance Scale (TAS), assess the therapist's perceptions of the alliance [29]. However, rating the working alliance in psychotherapy sessions is done manually. This work is very labor-intensive and is sensitive to bias due to the subjectivity of the observer's judgment.

Accurately predicting WA can help the therapist to gain a better understanding of the WA itself and increase the quality of the sessions, for example by preventing ruptures (i.e. deterioration in any part of the working alliance) that might interfere with the effects of psychotherapy.

In this paper, we assess the potential of machine learning based automatic analysis of the interaction between the patient and the therapist during psychotherapy sessions. This will not only save valuable time by removing the labor-intensive scoring of the sessions, but may also provide more insight into the efficiency of specific components of the psychotherapy session. The benefit of the latter is twofold. First, it can help improve the effectiveness of the psychotherapy sessions by helping the therapist improve their methods based on the patient's perception of WA. Second, an accurate prediction of WA can make psychotherapy sessions more effective, relieving some of the pressure on the mental health system by achieving the same outcome in fewer sessions.

In our study, we use the WAI-S and WAI-O for capturing the patient's and observer's perception of the working alliance and the WAI-SRT for the therapist's perception of the working alliance. We compare self-report, patient, and observer ratings in the WAI with machine learning based approaches to extract working alliance features from videos of therapy sessions from patients with major depressive disorder (MDD).

2 Related Work in Automatic Analysis of Multimodal Cues

Multimodal automatic processing of signals during a dyadic interaction encompasses a range of verbal and non-verbal cues [45]. This area particularly benefits from recent developments in deep learning, which resulted in capabilities like automatic speech to text transcription, voice and text based affect analysis, face tracking and facial expression analysis, body tracking and action recognition. A full review of recent approaches to models developed in each modality is out of scope for this paper, but we review here a range of related works that specifically deal with dyadic therapist-patient interactions in general, and working alliance estimation in particular.

In the analysis of working alliance, the linguistic information is valuable. [47] used natural language processing (NLP) to analyze the transcripts of therapy sessions and found that certain linguistic markers, such as the use of first-person pronouns and positive emotion words, were positively associated with the working alliance. [40] used structural equation modeling to investigate the role of language entrainment, which quantifies how much the therapist and patient adapt their language to each other over the course of the sessions. They found that when the therapist adapts their language to the patient, this effects the perception of alliance of the patient.

Related to language entrainment, studies have shown that the level of empathy shown by the therapist explains some of the variance in therapeutic outcome [3]. Subsequently, it is important for the therapist to be able to recognize emotions of their patients and act accordingly. However, [31] found that trained therapists do not perform better than psychology students in recognizing emotion intensity in their patients. Using computational approaches to measure facial emotion and emotion intensity can support the therapist in interpreting the patients, as well as in assessing the affect in the interaction. The synchrony of facial expressions displayed in a dyadic interaction is also important to determine trust between two people [32].

Research has shown that eye contact can indicate a strong bond [27], whereas aversion to eye contact is indicative of someone trying to attenuate the interpersonal relationship and thus is a sign of decreased working alliance [4, 20]. A therapist making eye contact is also perceived by the patient as engaging and friendly, stimulating a strong and positive working alliance [18]. It is possible to detect eye contact using computer vision based face analysis [7].

Additional non-verbal cues come from back-channel signals. A study by [39] looked at social and bodily features concerning working alliance using six features: head nods, head shakes, speaking turn length, the waiting time (the duration of the pause between the end of one speaker's turn and the beginning of the other speaker's turn), listening nods, and listening shakes, respectively. They found that the behavior of the therapist or the patient corresponded more with their own ratings of working alliance than the rating of their conversation partner. Also, head gestures were found to be specifically predictive of the patient's rating of the working alliance. In a more recent work, [41] used structured equation modeling to combine head motion features and language use (negations, pronouns, affective features). They used the WAI-S, but combined the *task* and *goal* sub-scales due to very high correlation. Different features were highlighted for these sub-scales; we discuss their findings comparatively, later in this paper.

Using machine learning approaches to predict working alliance is explored in several works. [30] proposed a transformer-based classification model that has a psychological state encoder to infer the working alliance scores by projecting embeddings of the dialogue turns onto an embedding space of the clinical inventory for working alliance. [48] used multiple questionnaires filled in by the patient and the therapist, extracted features from these, and applied deep learning on the extracted features. This study used 325 patients and 32 psychotherapists across different settings and found that models with both therapist and patient features performed better than models with only patient features. This is a finding supported by earlier research on patient-therapist interactions during child play therapy [15].

3 Materials and Methods

In this paper, we have tested a wide range of multimodal tools to analyze a dataset of recorded psychotherapy sessions for the estimation of working alliance. In this section, we first describe the dataset, and then proceed with the analysis approach.

3.1 Data

The data used in this paper stems from a large multicenter dataset (FreqMech Study) consisting of recorded Cognitive Behavioral Therapy (CBT) and Interpersonal Therapy (IPT) sessions for patients with a diagnosis of major depressive disorder according to the DSM IV or V criteria [10]. Participants who provided consent for videotaping their therapy sessions and had available WAI-S (n=91), WAI-SRT (n=100) or WAI-O (n=88) measurements were included in the current study (total n=114). The mean age of the participants is 37.85 years (± 12.26) and 61.5% of the participants are female. While most participants were born in the Netherlands, other birth countries included Morocco, Ukraine, China, Suriname, Iran, Hong Kong, Syria, Macedonia, South Africa, Portugal, and Greece. There are 76 therapists, and 12-20 sessions taped per patient. Many sessions include WAI-S and WAI-SRT scores as rated by the patient and the therapist, respectively, and also include observational coding on the quality and psychological processes (WAI-S scores), as rated by experts. The WAI questionnaire consisted of 12 questions for the patient and observer versions (see Supplementary Material) and of 10 questions for the therapist version[1], all in 7-point Likert scale. The recording and analysis of this dataset was agreed upon by the Medical Ethical Committee of VU Medical Centre Amsterdam (registration number 2014.337) and with full knowledge and consent from the participating patients and therapists.

The contents of the recordings are very diverse, as they differ in the visibility of the patient and the therapist, the viewing angle, lighting conditions, and audio quality. Subsequently, the recording conditions are challenging and realistic for different clinical settings. There are multiple sessions recorded per patient, and these were grouped in the analysis to make sure that no subjects are simultaneously present in both training and testing folds, which might have led to over-optimistic results during automatic classification.

We worked with a subset of the original dataset, and used speaker turns as a unit of further segmenting the data. Selecting samples based on the availability of the WAI scores, we obtained a therapist dataset (37 sessions, 21700 turns), a patient dataset (43 sessions, 25000 turns), and an observer dataset (34 sessions, 20000 turns). Score distributions are reported in Figure 1. All three datasets had an average turn length of about 9,6 tokens. WAI scores are estimated at session level. The audio recordings are mono-sound and in Dutch language. We also have visuals of either the therapist or the patient (and in some cases both). The features for automatic analysis were extracted from visual, audio, and text modalities.

3.2 Feature extraction

3.2.1 Automatic Transcription. To get the verbal content of the therapist-patient interactions automatically, we conducted speech-to-text transcription using the state-of-the-art WhisperX model [6], using the "large-v1" pre-trained model with a batch size of five. Automatic transcription is much faster than manual transcription, but it may be error prone. Especially for non-English languages, automatic analysis tools may perform with lower accuracy, as they are typically trained on smaller datasets.

[1]Available at https://wai.profhorvath.com/downloads.

3.2.2 Diarization. Once the transcripts are obtained, the second task is automatic diarization, which separates the speech of the therapist and the patient, based on the audio track. We used the PyAnnote tool [35], with automatic segmentation and clustering threshold selection to determine the difference between noise, speech, and different speakers. Automatic threshold selection is preferred, as the audio quality and presence of noise could be very different between patients and between sessions.

3.2.3 Conversational features. From transcribed and diarized text, we extracted conversational features consisting of turn-taking (i.e. participation equality, turn-level freedom and percentage of overlapping turns) and turn-level features (i.e. per speaker turn duration, token count, speech rates and pause-based features for different speech parts, decomposed as between and within the speakers, as well as the total percentage of speaking time for the interlocutors during the whole conversation), following a recently introduced approach by [8].

3.2.4 Textual features. We tested out three different approaches for automatic sentiment analysis on the transcripts, using multiple Bidirectional Encoder Representations from Transformers (BERT)-based models [13]. To obtain positive and negative sentiments, we used RobBERT-v2-Dutch-Sentiment [43], which is a version of the Dutch language model RobBERT [12], fine-tuned with sentiment polarity labels. Additionally, we used the VADER sentiment analyzer to extract the polarity (positive/negative) and the intensity of the emotion present in the text (a summation of the intensity of each word in the text) [24].

For discrete emotion classification, we used the EmoRoBERTa model [26], which is able to classify 28 different types of emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, and neutral, respectively [28].

Since both EmoRoBERTa and VADER tools were trained in English, we translated the Dutch transcriptions into English using the offline Argos Translate package before sentiment and emotion classification. Due to the sensitivity of the data, cloud-based services were not usable, and only local models were used. Finally, we extracted (dimensional) emotions in the form of arousal and valence values, using the XLM-RoBERTa-large model [11]. This pre-trained model is fine-tuned on 34 publicly available datasets containing arousal and valence ratings, including one consisting of 4300 Dutch words labeled for valence, arousal, and dominance [33]. The sentiment and specific emotion features were extracted per utterance, but combined to form session-level features.

3.2.5 Audio features. We extracted arousal and valence scores from the audio of the sessions using Audeering's Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0 [46]. This model was fine-tuned with 100 hours of annotated speech data originating from podcast recordings.

3.2.6 Facial expressions and head gestures. We extracted Facial Action Units (FACs) using OpenFace [7] and used these to determine head nods and head shakes, following [39]. Face detection confidence was thresholded at 75%, and incorrectly detected faces

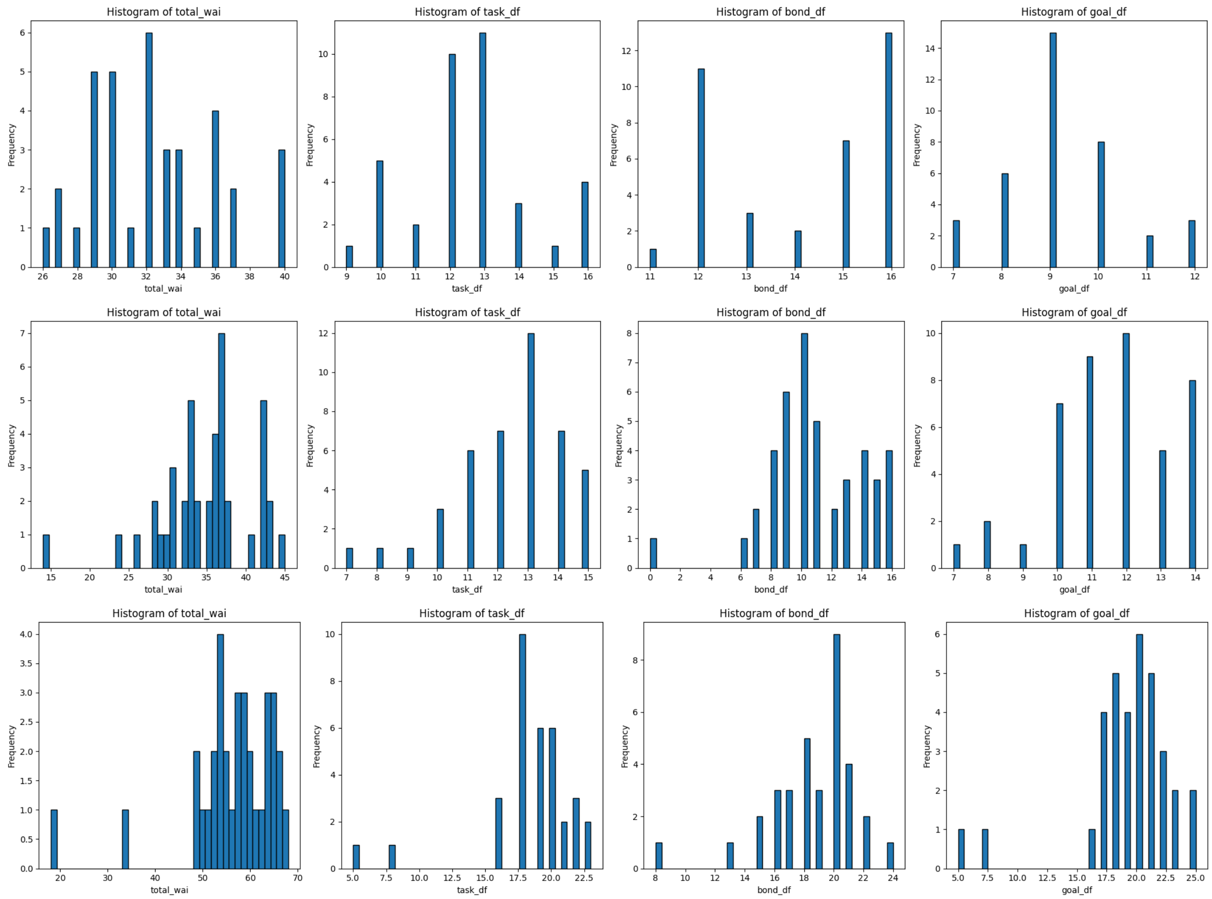

Figure 1: The score distributions for therapist (top row), patient (middle row), and observer (bottom row) scored sessions. Each row contains the WAI, task, bond, and goal scores from left to right. The shapes of the score distributions are similar across therapist, patient, and observer scores, for each category. However, the score ranges show differences, with observer ratings typically higher than either therapist or patient ratings.

were filtered out. We then calculated the distance traveled along the pitch and yaw dimensions per video frame over a rolling window of one second. Using smoothness constraints on these features, we segmented individual head gestures. Finally, the timestamps of the head gestures of each face are compared to the transcription file to split the head gestures that occurred during the speaking time of the therapist or the patient. The head movements were categorized as occurring during either speaking of listening, for both the therapist and the patient.

The facial expressions were extracted for every second of video, according to a rule-based method based on FACs as described by [42]. The facial expressions were similarly categorized as occurring during speaking or listening times.

3.3 Models

Since the amount of data we have is limited for end-to-end learning, we used a support vector machine (SVM) in the classification task. In our preliminary analysis, we have tested a number of regression models, including k-Nearest Neighbor (kNN), Support Vector Regressor (SVR), Elastic Net, Random Forest, Multilinear Regression, and Extreme Gradient Boosting (XGB) Regressor. However, a classification approach produced more stable results, and higher accuracy overall. We used 5-fold cross validation, and all models were optimized using a grid search on the training folds to find the best hyperparameter settings. Only training folds were used for model selection. The features were z-normalized before being fed to the models to a mean of zero and a standard deviation of one, and normalization parameters are learned from the training data. For feature selection, we used the Maximum Relevance Minimum Redundancy algorithm (MRMR) [34].

The individual feature analysis was done on multiple datasets with a larger size, which were merged for the combined feature analysis using the machine learning models. Table 1 shows the number of data points (sessions) for each dataset. Conversational level features had the largest dataset, and the visual datasets were the smallest, as the faces were not often visible in the recordings.

Table 1: Number of sessions in the individual datasets for each type of feature.

Feature	Patient scores	Therapist scores	Observer scores
Audio affect	43	40	35
Text emotions	52	44	39
Text affect	75	79	71
Conversational level	91	100	88
Facial expressions	28	38	34
Head gestures	24	33	29

4 Experimental Results

4.1 Automatic transcription

We evaluated the performance of the transcription tool using the Word Error Rate (WER), Match Error Rate (MER), and Word Insertion Likelihood (WIL) measures. WER quantifies the accuracy of the transcription based substitutions (S, incorrectly transcribed words), insertions (I, additional words in the transcription compared to the ground truth) and deletions (D, missing words in the transcription compared to the ground truth), given a text of N words:

$$WER = \frac{S + D + I}{N} \tag{1}$$

MER represents the probability of a match between the transcribed and the ground truth texts being incorrect. It is calculated by dividing the total number of errors (S+D+I) by the total number of words (N) and the insertions (I):

$$MER = \frac{S + D + I}{N + I} \tag{2}$$

Lastly, WIL represents the probability of insertions in the transcription:

$$WIL = \frac{I}{N} \tag{3}$$

We tested automatic transcription first on a video with perfect audio quality, and then on two sample sessions of the actual dataset, which were manually transcribed. The WhisperX transcription performed well on the test set with good audio quality. Table 2 shows that the transcription accuracy on a real session is less accurate than on the high-quality test video.

Table 2: Evaluation using the Word Error Rate (WER), Match Error Rate (MER), and Word Insertion Likelihood (WIL) of the automatic transcriptions on a high-quality test video and two samples from the actual dataset.

Dataset	WER	MER	WIL
Test Video	0.0085	0.0085	0.0141
1025 session 6	0.1791	0.1782	0.2339
6013 session 2	0.0993	0.0986	0.1479

To illustrate how the Whisper model we have used fares against other transcription models in the literature, we performed a comparison of five models on the test video. Table 3 shows that our selected method is clearly preferable to the other approaches.

Table 3: WER, MER, and WIL scores for different systems on the test video

Model	WER	MER	WIL
whisper large-v2	0.01	0.01	0.01
whisper large-v1	0.06	0.06	0.09
whisper medium	0.07	0.07	0.12
word online	0.23	0.22	0.36
google	0.39	0.38	0.55

4.2 Diarization

A similar approach was used to test diarization, where a test video with crisp audio quality and two samples from the actual dataset were manually referenced and used to evaluate the performance of the PyAnnote diarization pipeline. The evaluation measures were the Diarization Error Rate (DER) and Jaccard Error Rate (JER) (see Table 4). DER measures the error rate for the entire text, while in JER, the error rates are first computed per speaker and then averaged.

$$DER = \frac{FA + Missed + Incorrect}{T} \tag{4}$$

$$JER = \frac{1}{N} \sum_{i=1}^{N} \frac{FA_i + Missed_i}{TOTAL_i} \tag{5}$$

Table 4: Evaluation using the DER and JER of the PyAnnote-produced diarizations on a test video and two samples from the actual dataset.

Dataset	Diarization Error Rate (%)	Jaccard Error Rate (%)
Test Video	1.7	0.02
1025 session 6	9.1	0.13
6013 session 2	6.3	0.12

We compared the diarization of PyAnnote system to several alternatives. Figure 2 shows that it has the best performance on diarization with a DER of 1.7% compared to 14.1% for the closest contender.

4.3 Classification

We formulate the problem of automatic WA assessment as a binary classification. Given the data from a single session, the classifier produces a score to indicate high or low working alliance. This is a cautious approach, as our dataset is limited in size. We used an equal frequency binning strategy for the binary split. As a result, Observer data is split as 19/15 high/low, Patient data as 22/21, and Therapist data 22/15. The test set is randomly drawn after binning, and was not stratified. have 4/3, 4/5, 5/3 samples (for high/low) for Observer, Patient, and Therapist sets, and a random baseline accuracy would be around 50%. Table 5 summarizes the results, in terms of accuracy, precision, recall, and F1 scores.

While the SVM classifier shows good prediction results, we caution the reader that the database is relatively small, although comparable to similar studies in the literature. We have 76 therapists,

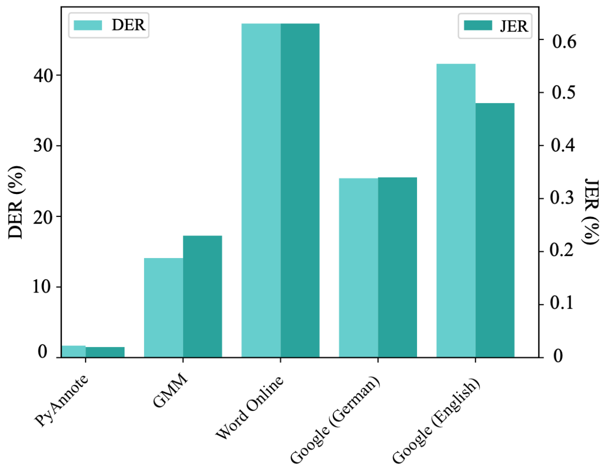

Figure 2: Performance of different diarization models on a test video.

114 patients, and 12-20 sessions per patient. [40] reports 66 therapy sessions between 39 unique clients and 11 unique therapists, and [2] reports 252 in-person and teletherapy sessions from 47 patients treated by 10 clinicians. [48] trained models with data from 325 patients and 32 therapists, and tested them on a clinical sample sample of 85 patients and 8 therapists.

By using off-the-shelf tools and pre-trained models, we have sought to minimize the risk of overlearning. As the accuracy of the classifier is relatively high across all three datasets (ranging from 0.81 to 0.93, with 1.0 representing perfect classification), we conclude that supervised classification is promising for assessing the therapy sessions broadly into low or high working alliance classes. The comparatively low recall for the low working alliance class shows that the model has some bias towards producing a high WA rating. We next look at individual features to gain more insight about which feature is informative for this problem.

Table 5: Performance results of the SVM Classifier on patient, therapist, and observer WAI score prediction. The precision, recall, and the F1 scores are reported per class (High or Low WA)

Dataset	Accuracy	Precision (H-L)	Recall (H-L)	F1-Score (H-L)
Patient	0.81	0.75-**0.80**	0.75-**0.80**	0.75-**0.80**
Therapist	0.93	0.71-**1.00**	**1.00**-0.33	**0.83**-0.50
Observer	0.93	0.80-**1.00**	**1.00**-0.67	**0.89**-0.80

4.4 Correlation analysis

We evaluated the correlation of each feature used in our models with WAI scores. Table 6 shows an overview of the features per modality for each rater (patient, therapist, and observer) with a significant Spearman correlation result ($p < 0.05$), with Bonferroni correction. The features displayed in black color have a positive correlation with the WAI score, whereas the features in red color have a negative correlation.

A quick look at the table shows that some conversational features (such as the turn length of the patient and the therapist), as well as some speech emotion related features, consistently correlate with the WAI scores. However, for the text, face and movement modalities, correlations with WAI scores depend on the perspective of the scorer. The therapist, the patient, and the external observer presumably use different cues to assess WA. We have already observed some of these differences in the ground truth score distribution, but looking at the correlation structure is informative. We provide some observations in the next section.

5 Discussion

5.1 Findings

While it was clear from the onset that the scoring of the WAI by patients and therapists differed, our automatic feature analysis reveals patterns that can be considered in more detail. The differences were most prominent in conversational features and head movements.

While the head movement of therapists during listening turns is weakly correlated with a higher working alliance according to the therapist scores (not significant, hence not shown on Table 6), it is correlated with a lower working alliance as rated by the patient. When patients are speaking, excessive head movement by the therapist may be perceived by the patient as a sign of distraction, causing a lower perception of working alliance. The predictability of a session's conversational structure correlated positively in the therapist ratings, whereas patient ratings pointed out to a negative correlation. Furthermore, the results on emotion display suggest that patients appreciate an open conversation with space to show their emotions, and especially negative emotions.

For reading the patient's response during a session, the therapist can pay attention to facial expressions. Displayed anger is a sign of a low working alliance, and elevated amounts of head nods during listening times is a sign of a high working alliance perception.

The classification analysis showed a similar level of predictability between the patient, therapist, and observer WAI scores, but differences in which specific features contributed to the prediction. There was a difference in the importance of which modality was predictive for which perception, as conversational features were more correlated with the patient WAI scores, and the audio affect features were more correlated with the therapist WAI scores. Furthermore, the WAI scores of patients and therapists were more informed by the rater's behavior than that of the other speaker. The observer results do not necessarily resemble the patient or therapist results, indicating that the three different raters each have different underlying features for perceiving the working alliance.

Due to the size limitations of the dataset, feature selection did not result in a significant improvement in the predictive performance of the classification models. According to our analysis, conversational features were the most important indicators for predictive models.

This corroborates the main findings of [8]. They report that participation equality was correlated with the task bond and goal components with a correlation ranging from 0.42 to 0.53, which we did not find within the observer dataset. However, there was

Table 6: Overview of all significant features by Spearman correlation test sorted per modality and WAI category. Both positive (black) and negative (red) correlations with the WAI are shown.

	Conversational	Speech	Text	Facial	Movement
P.	Speech duration patient (0.39) Turn length patient (0.39) Participation equality (0.31) Turn length therapist (-0.31) Speech duration therapist (-0.31) Speech rate patient (-0.30) Turn-level-freedom (-0.30)		Amusement (0.40) Fear (0.39) Min arousal patient (-0.30)	Happiness (0.62) Anger (-0.63)	Listening nods therapist (-0.77) Listening shakes therapist (-0.77) Listening nods patient (-0.54)
T.	Overlapping segments (0.35) Number of turns (0.34) Speech rate patient (0.32) Turn-level freedom (0.30) Duration % therapist (0.24) Turn length therapist (-0.30) Turn length patient (-0.25)	Max valence (0.63) Min valence (0.57) Max arousal (0.47)	Disgust (0.38) Positive sentiment (-0.38)	Neutral (0.48) Sadness (-0.41) Anger (-0.30)	
O.	Number of turns (0.33) Turn-level freedom (0.28) Overlapping segments (0.25) Turn length therapist (-0.32) Turn length patient (-0.27)	Max valence (0.56) Min valence (0.53) Max arousal (0.41)	Remorse (0.46) Approval (0.44) Disgust (0.39) Confusion (-0.47) Disapproval (-0.44) Excitement (-0.40)	Anger (-0.50)	Listening nods patient (-0.57) Listening shakes patient (-0.57)

a positive correlation within the patient dataset between the participation equality with the goal and task components (0.30 and 0.31, respectively). Similarly, our research shows a positive correlation between the turn-level-freedom and the bond component (0.28). For the therapist dataset, we found a positive correlation with the task component (0.30). To summarize the differences in our settings: 1- they have used only observer-based scores in their analysis, whereas we give a more detailed decomposition; 2- they use a dataset of around 70.1k words for 23 recorded conversations, compared to 200k words in our dataset for 34 sessions; 3- their analysis language is Italian, whereas ours is Dutch; 4- they divided sessions into three segments (15, 70, 15 per cent, respectively), whereas our recordings did not necessarily start at the beginning of a session, and subsequently, we did not use such a segmentation; 5- they have used manual text diarization, whereas we used automatic transcription and diarization. It is encouraging to obtain similar results in face of these differences.

The research by [39] is also closely related to ours. Their analysis focused primarily on head gestures and turn-taking behaviors as features predictive of the WAI. We found that the features that proved important for their analysis were also important in our models. Specifically, the head gestures during listening periods were significant predictors for patient working alliance ratings in both studies. Also, our study confirms their finding that therapists rely more on conversational features, like turn length and wait time, compared to head movements, and both studies showed that head gestures were more reflective of the task component of the working alliance, while conversational behaviors tended to be more reflective of the bond component. A difference between our results and [39] is that we found that therapist ratings do not correlate much with their own head movements, but they do with the patient's head movements during speaking periods.

5.2 Limitations and challenges

In this study, we did not separate state from trait alliance. The former is more related to the therapeutical process [50], and requires a longitudinal analysis approach. Our focus has been on the automatic extraction of low-level, multimodal cues, which can be used for more fine-grained analysis.

A design choice we made in our study was to treat each session individually, instead of modeling alliance throughout the therapy as a temporal construct. [40] introduced a structured equation model to incorporate such temporal aspects. In our case, there were a higher number of therapists involved in the sessions, and it was quite common to have different therapists attending a single patient at different times. Subsequently, such modeling was not preferred.

Some studies have found that working alliance may not be an appropriate measure for all cultural groups, as cultural differences can influence perceptions of the therapeutic relationship [44]. As an example, [25] found that the engagement of East Asian Americans in psychotherapy is much lower than that of white and minority groups. This is one important argument about the applicability of the automatic tools.

A related limitation is that our data came from a relatively limited cultural background (although we have patients from 12 countries in three continents), and the automatic assessment tools we use are not always trained to work with subjects from a wide cultural and ethnic background. While the AI community is collectively striving to improve the automatic assessment tools to be more inclusive, the full potential of these approaches is not reached. This means that any practical tools to be implemented in a clinical setting would need thorough testing within the practical cultural context, and its results carefully interpreted (or ignored) if the patient or therapist demographics are less represented in the training conditions.

Another related issue is the difficulty of working with a non-English language. The majority of the existing text analysis tools are specialized on the English language, whereas we work with Dutch transcripts in this paper. We have tested a range of tools, some trained directly on Dutch texts, and some based on English, but adapted to the problem via automated translation. In some cases, the high quality of translation tools and the performance difference between models trained in different languages justifies (empirically) the use of automatic translation [21]. There is rapid progress in this area, and each year we see significant improvements.

The challenges listed above are mostly technical issues, but there are also practical issues related with the use of AI-based assessment during psychotherapy. While it is shown that monitoring WA regularly can support the therapist in improving the therapeutical process, a recent survey done with 954 patients and 248 clinicians about the usefulness of self-report and AI-based assessment provides some additional findings about the practical aspects [1]. Self-report is found to be burdensome, and AI-based assessment can alleviate this. Clinicians are more reluctant than the patients to allow recording of sessions, and patients are more positive about AI-based feedback for clinical outcomes.

A limitation that may not easily benefit from progress in technology is the problem with facial visibility, due to privacy concerns of the patients and clinicians. In our data collection, the patients (or the therapists) can easily opt out of facial recording. We should note that although facial visibility is limited, our dataset comes from an effectiveness study and likely to generalize to Dutch routine clinical practice. Nonetheless, related work in the literature reports similar issues; in [2], only 80 sessions out of 252 had video recordings. Actually, it is technologically possible (and easy) to record a patient and process facial features on the fly to store a privacy-aware representation (such as a series of facial action units). In such a case, the facial image is not stored, and biometric identification is not possible. However, this is more of a trust issue. As the recent European AI Act[2] illustrated, affective technology and automatic behavior analysis are associated with high risk levels. Strong legal mandates and requirements will enable applications that are safe and trusted, helping to overcome this issue. Related to this, we did not have all modalities available in all sessions, and data limitations prevented extensive multimodal fusion analyses, which can be considered after further data collection.

6 Conclusions

In this paper, we have tested a broad range of AI based automatic assessment tools to extract conversational, speech, text, facial, and movement related indicators from psychotherapy sessions, and performed predictive and descriptive analyses on these indicators. Our results indicate to what extent WA assessment can be automated with the existing tools. While our SVM-based classifier obtained 81% accuracy for predicting the patient-based WAI scores, we approach these results carefully. Clearly, much more extensive training sets are needed for robust prediction. However, even the feature extraction approaches without subsequent classification could be potentially useful for gaining insights into WA. Further studies can focus on prediction of potential alliance ruptures or increases, and in optimizing an automatic algorithm that addresses prediction of individual symptoms to have improved prediction accuracy combined with explainability.

As a second contribution, we looked at the correlations between individual features and WA scores given by patients, therapists, and external observers. Our findings highlight where the therapist's perception differs from the patient's perception, and raises new venues of exploration for the use of social signals during psychotherapy. Our analysis suggests that conversational features were the most important contributors to the WAI. Further research is required to systematically explore these initial findings.

Finally, we make our code base available to the readers[3] . While the patient and therapist data used during this research is confidential, the computational analysis tools can be easily shared and used across other settings. The shared code base includes more technical details about the methods used, and allows reproducibility.

7 Safe and Responsible Innovation Statement

We followed [19] to check our work on ethical risks of applying affective computing technologies in the mental healthcare domain. The most important issues are consent, biases, and privacy risks. The data were used under informed consent from subjects, and there are no risks involving the developed systems affecting the care of the people involved.

The data processed for this paper comes from a limited demographic, and we note this as a shortcoming. Furthermore, some of the approaches we used may have internal biases (such as facial analysis being more effective on certain population segments). On the other hand, due to the privacy-sensitivity of therapy data, there are not many openly distributed datasets, and limited demographic coverage is a common problem. We are not able to share any patient data used in this work, but we make our code available to enable the replication of the study, and the measurement of any such biases on any given dataset.

During the processing of sensitive personal data, GDPR was closely observed. No data processing happened in the cloud, or on remote services. Data was stored on protected university servers at all times, accessed via secure protocols, and locally (and anonymously) processed with algorithms that were either developed by the authors, or were open source.

Acknowledgments

This study was supported by a grant to S.J.E. Bruijniks and A.A. Salah from UU Applied Data Science. We acknowledge the contribution of the participating patients. Without their cooperation in the trial, this study would have been impossible to conduct.

[2]https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[3]https://github.com/VR-13/Paper_UU_code.git

References

[1] Katie Aafjes-van Doorn. 2025. Feasibility of artificial intelligence-based measurement in psychotherapy practice: Patients' and clinicians' perspectives. *Counselling and Psychotherapy Research* 25, 1 (2025), e12800.

[2] Katie Aafjes-Van Doorn, Marcelo Cicconet, Jeffrey F Cohn, and Marc Aafjes. 2025. Predicting working alliance in psychotherapy: A multi-modal machine learning approach. *Psychotherapy Research* 35, 2 (2025), 256–270.

[3] Maayan Abargil and Orya Tishby. 2022. How therapists' emotion recognition relates to therapy process and outcome. *Clinical Psychology & Psychotherapy* 29, 3 (2022), 1001–1019.

[4] Nalini Ambady, Jasook Koo, Robert Rosenthal, and Carol H Winograd. 2002. Physical therapists' nonverbal communication predicts geriatric patients' health outcomes. *Psychology and aging* 17, 3 (2002), 443.

[5] Allison L Baier, Alexander C Kline, and Norah C Feeny. 2020. Therapeutic alliance as a mediator of change: A systematic review and evaluation of research. *Clinical psychology review* 82 (2020), 101921.

[6] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH* (2023).

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[8] Sebastian P Bayerl, Gabriel Roccabruna, Shammur Absar Chowdhury, Tommaso Ciulli, Morena Danieli, Korbinian Riedhammer, and Giuseppe Riccardi. 2022. What can Speech and Language Tell us About the Working Alliance in Psychotherapy. *arXiv preprint arXiv:2206.08835* (2022).

[9] Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice* 16, 3 (1979), 252.

[10] Sanne JE Bruijniks, Lotte HJM Lemmens, Steven D Hollon, Frenk PML Peeters, Pim Cuijpers, Arnoud Arntz, Pieter Dingemanse, Linda Willems, Patricia Van Oppen, Jos WR Twisk, et al. 2020. The effects of once-versus twice-weekly sessions on psychotherapy outcomes in depressed patients. *The British Journal of Psychiatry* 216, 4 (2020), 222–230.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR* abs/1911.02116 (2019). arXiv:1911.02116 http://arxiv.org/abs/1911.02116

[12] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3255–3265. doi:10.18653/v1/2020.findings-emnlp.292

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Marc J Diener, Mark J Hilsenroth, and Joel Weinberger. 2007. Therapist affect focus and patient outcomes in psychodynamic psychotherapy: A meta-analysis. *American Journal of Psychiatry* 164, 6 (2007), 936–941.

[15] Metehan Doyran, Batıkan Türkmen, Eda Aydın Oktay, Sibel Halfon, and Albert Ali Salah. 2019. Video and text-based affect analysis of children in play therapy. In *ACM International Conference on Multimodal Interaction*. 26–34.

[16] Fredrik Falkenström, Annika Ekeblad, and Rolf Holmqvist. 2016. Improvement of the working alliance in one treatment session predicts improvement of depressive symptoms by the next session. *Journal of Consulting and Clinical Psychology* 84, 8 (2016), 738–751.

[17] C Flückiger, AC Del Re, BE Wampold, and AO Horvath. 2018. The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy* 55, 4 (2018), 316–340.

[18] Jairo N Fuertes. 2019. *Working alliance skills for mental health professionals*. Oxford University Press, USA.

[19] Jeffrey M Girard, Dasha A Yermol, Albert Ali Salah, and Jeffrey F Cohn. 2025. Computational Analysis of Expressive Behavior in Clinical Assessment. *Annual review of clinical psychology* 22 (2025).

[20] Gerald W Grumet. 1983. Eye contact: The core of interpersonal relatedness. *Psychiatry* 46, 2 (1983), 172–180.

[21] Sibel Halfon, Eda Aydın Oktay, and Albert Ali Salah. 2016. Assessing affective dimensions of play in psychodynamic child psychotherapy via text analysis. In *Human Behavior Understanding: 7th International Workshop, HBU 2016, Amsterdam, The Netherlands, October 16, 2016, Proceedings 7*. Springer, 15–34.

[22] AO Horvath and LS Greenberg. 1989. Development and validation of the Working Alliance Inventory. *Journal of counseling psychology* 36, 2 (1989), 223.

[23] AO Horvath and BD Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology* 38, 2 (1991), 139.

[24] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[25] Kevin C Hynes. 2019. Cultural values matter: The therapeutic alliance with East Asian Americans. *Contemporary Family Therapy* 41, 4 (2019), 392–400.

[26] Rohan Kamath, Arpan Ghoshal, Sivaraman Eswaran, and Prasad Honnavalli. 2022. An Enhanced Context-based Emotion Detection Model using RoBERTa. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. 1–6. doi:10.1109/CONECCT55679.2022.9865796

[27] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63.

[28] Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with Roberta. *arXiv preprint arXiv:2108.12009* (2021).

[29] Michael J Lambert and Dean E Barley. 2001. Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, research, practice, training* 38, 4 (2001), 357–361.

[30] Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2022. Working Alliance Transformer for Psychotherapy Dialogue Classification. *arXiv preprint arXiv:2210.15603* (2022).

[31] Paulo PP Machado, Larry E Beutler, and Leslie S Greenberg. 1999. Emotion recognition in psychotherapy: Impact of therapist level of experience and emotional awareness. *Journal of Clinical Psychology* 55, 1 (1999), 39–57.

[32] Adrien Meynard, Gayan Seneviratna, Elliot Doyle, Joyanne Becker, Hau-Tieng Wu, and Jana Schaich Borg. 2021. Predicting Trust Using Automated Assessment of Multivariate Interactional Synchrony. In *16th IEEE International Conference on Automatic Face and Gesture Recognition*.

[33] Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior research methods* 45 (2013), 169–177.

[34] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1226–1238.

[35] Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 1021–1028.

[36] Nele Stinckens, Annick Ulburghs, and Laurence Claes. 2009. De werkalliantie als sleutelelement in het therapiegebeuren: Meting met behulp van de WAV-12, de Nederlandstalige verkorte versie van de Working Alliance Inventory. *Tijdschrift Klinische Psychologie* 39 (2009), 44–60.

[37] Terence J Tracey and Anna M Kokotovic. 1989. Factor structure of the working alliance inventory. *Psychological Assessment: A journal of consulting and clinical psychology* 1, 3 (1989), 207–210.

[38] Volker Tschuschke, Margit Koemeda-Lutz, Agnes von Wyl, Aureliano Crameri, and Peter Schulthess. 2022. The impact of clients' and therapists' characteristics on therapeutic alliance and outcome. *Journal of Contemporary Psychotherapy* 52, 2 (2022), 145–154.

[39] Alexandria K. Vail, Jeffrey Girard, Lauren Bylsma, Jeffrey Cohn, Jay Fournier, Holly Swartz, and Louis-Philippe Morency. 2021. Goals, Tasks, and Bonds: Toward the Computational Assessment of Therapist Versus Client Perception of Working Alliance. In *16th IEEE International Conference on Automatic Face and Gesture Recognition*. 1–8. doi:10.1109/FG52635.2021.9667021

[40] Alexandria K Vail, Jeffrey Girard, Lauren Bylsma, Jeffrey Cohn, Jay Fournier, Holly Swartz, and Louis-Philippe Morency. 2022. Toward Causal Understanding of Therapist-Client Relationships: A Study of Language Modality and Social Entrainment. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*. 487–494.

[41] Alexandria K Vail, Jeffrey M Girard, Lauren M Bylsma, Jay Fournier, Holly A Swartz, Jeffrey F Cohn, and Louis-Philippe Morency. 2023. Representation Learning for Interpersonal and Multimodal Behavior Dynamics: A Multiview Extension of Latent Change Score Models. In *Proc. ACM International Conference on Multimodal Interaction*.

[42] Michel François Valstar and Maja Pantic. 2006. Biologically vs. logic inspired encoding of facial actions and emotions in video. In *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 325–328.

[43] Benjamin van der Burgh and Suzan Verberne. 2019. The merits of Universal Language Model Fine-tuning for Small Datasets - a case with Dutch book reviews. *CoRR* abs/1910.00896 (2019). arXiv:1910.00896 http://arxiv.org/abs/1910.00896

[44] Melba JT Vasquez. 2007. Cultural difference and the therapeutic alliance: an evidence-based analysis. *American Psychologist* 62, 8 (2007), 878.

[45] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[46] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[47] Bruce E Wampold, Gregory W Mondin, Marcia Moody, Frederick Stich, Kurt Benson, and Hyun-nie Ahn. 1997. A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially," all must have prizes.". *Psychological*

bulletin 122, 3 (1997), 203.
[48] Ying Zhou, Xiao-yu Chen, Ding Liu, Yu-lin Pan, Yan-fei Hou, Ting-ting Gao, Fei Peng, Xiao-cong Wang, and Xiao-yuan Zhang. 2022. Predicting first session working alliances using deep learning algorithms: A proof-of-concept study for personalized psychotherapy. *Psychotherapy Research* (2022), 1–10.
[49] Sigal Zilcha-Mano and Paula Errázuriz. 2015. One size does not fit all: Examining heterogeneity and identifying moderators of the alliance–outcome association. *Journal of counseling psychology* 62, 4 (2015), 579–591.
[50] Sigal Zilcha-Mano and Hadar Fisher. 2022. Distinct roles of state-like and trait-like patient–therapist alliance in psychotherapy. *Nature Reviews Psychology* 1, 4 (2022), 194–210.