npj | mental health research

www.nature.com/npjmentalhealth

ARTICLE OPEN

Check for updates

A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency

Adam S. Miner[1,2,13]✉, Scott L. Fleming[3,4,13], Albert Haque[4], Jason A. Fries[2], Tim Althoff[5], Denise E. Wilfley[6], W. Stewart Agras[1], Arnold Milstein[7], Jeff Hancock[8], Steven M. Asch[9,10], Shannon Wiltsey Stirman[1,9,11], Bruce A. Arnow[1] and Nigam H. Shah[2,3,7,12]

Although individual psychotherapy is generally effective for a range of mental health conditions, little is known about the moment-to-moment language use of effective therapists. Increased access to computational power, coupled with a rise in computer-mediated communication (telehealth), makes feasible the large-scale analyses of language use during psychotherapy. Transparent methodological approaches are lacking, however. Here we present novel methods to increase the efficiency of efforts to examine language use in psychotherapy. We evaluate three important aspects of therapist language use - timing, responsiveness, and consistency - across five clinically relevant language domains: pronouns, time orientation, emotional polarity, therapist tactics, and paralinguistic style. We find therapist language is dynamic within sessions, responds to patient language, and relates to patient symptom diagnosis but not symptom severity. Our results demonstrate that analyzing therapist language at scale is feasible and may help answer longstanding questions about specific behaviors of effective therapists.

npj Mental Health Research (2022)1:19 ; https://doi.org/10.1038/s44184-022-00020-9

INTRODUCTION

Individual psychotherapy is an effective treatment for a wide range of mental health conditions[1,2]. Two problems that have emerged in research on outcomes from psychotherapy are that 1) data from meta-analyses[3,4], randomized clinical trials[5–7], naturalist settings[8], as well as qualitative reviews[9], reveal little evidence that one specific form of psychotherapy is superior to another even when hypothesized change mechanisms are significantly different; and 2) while some therapists consistently achieve better outcomes than others (i.e., therapist effects), it is unclear what individual therapists may be doing that accounts for these effects[10–12]. Indeed, a recent comprehensive review of therapist effects noted that factors accounting for therapist effectiveness are "best characterized as emergent"[13].

Studies of the psychotherapy process attempt to understand what happens during therapy sessions that may explain patient improvement[14–16]. The chief method used since the 1950s to evaluate therapist behavior in therapy sessions is to have trained humans identify clinically meaningful therapist utterances in transcripts, and draw conclusions based on observed patterns[17–20]. Although useful, relying solely on human inspection of transcripts is not likely to meet demands for improved reproducibility and scalability in psychotherapy process research[19,21–26].

Computational approaches using natural language processing offer the potential to move past human limits of attention and reproducibility[19,27–32]. Improvements in computational power, the growing ease of recording and transcribing therapy sessions, and a shift to computer-mediated communication in healthcare (i.e., telehealth) make this feasible[19,22,33,34]. Supervised machine learning has provided insight into important constructs such as empathy and therapeutic interventions but rely on time-consuming and sometimes inconsistent human evaluation, making inspectability and reproducibility a challenge[26,35,36]. Early work is promising, but does not yet translate to best practices for improved patient outcomes or provide a clear direction for therapist training[19,28,36–39]. Methodological improvements are needed to bridge divisions between theoretical schools of thought (e.g., Cognitive Behavioral, Interpersonal, Psychodynamic, Counseling) as to which therapist language patterns correlate with favorable therapy outcomes[21,23,29,32,34,40,41(pp72-73),42]. If known, the linguistic behavior of successful therapists may inform targeted clinical trials to test causality and implementability, subsequently improving clinician training.

A fundamental tenet of psychotherapy is that therapists expose patients to language that may be helpful (e.g., emotional validation) and avoid language that may be harmful (e.g., shaming). Therapist language should be well-timed and appropriate for the specific moment. Nevertheless, the specific timing, frequency, and reactivity of therapist utterances is difficult to scrutinize systematically without human inspection[21]. Difficulties are multifaceted, with key limitations being theoretical (i.e., disagreement about mechanisms of change), technical (i.e., lack of validated tools for language measurement), and practical (i.e., lack of clinically meaningful datasets). This feasibility study primarily addresses the technical limitations of language analysis

[1]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [2]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. [6]Departments of Psychiatry, Medicine, Pediatrics, and Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA. [7]Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [8]Department of Communication, Stanford University, Stanford, CA, USA. [9]VA Palo Alto Health Care System, Palo Alto, CA, USA. [10]Division of Primary Care and Population Health, Stanford University School of Medicine, Stanford, CA, USA. [11]National Center for Posttraumatic Stress Disorders, Dissemination and Training Division, VA Palo Alto Healthcare System, Menlo Park, CA, USA. [12]Technology and Digital Solutions, Stanford Healthcare, Stanford, CA, USA. [13]These authors contributed equally: Adam S. Miner, Scott L. Fleming. ✉email: aminer@stanford.edu

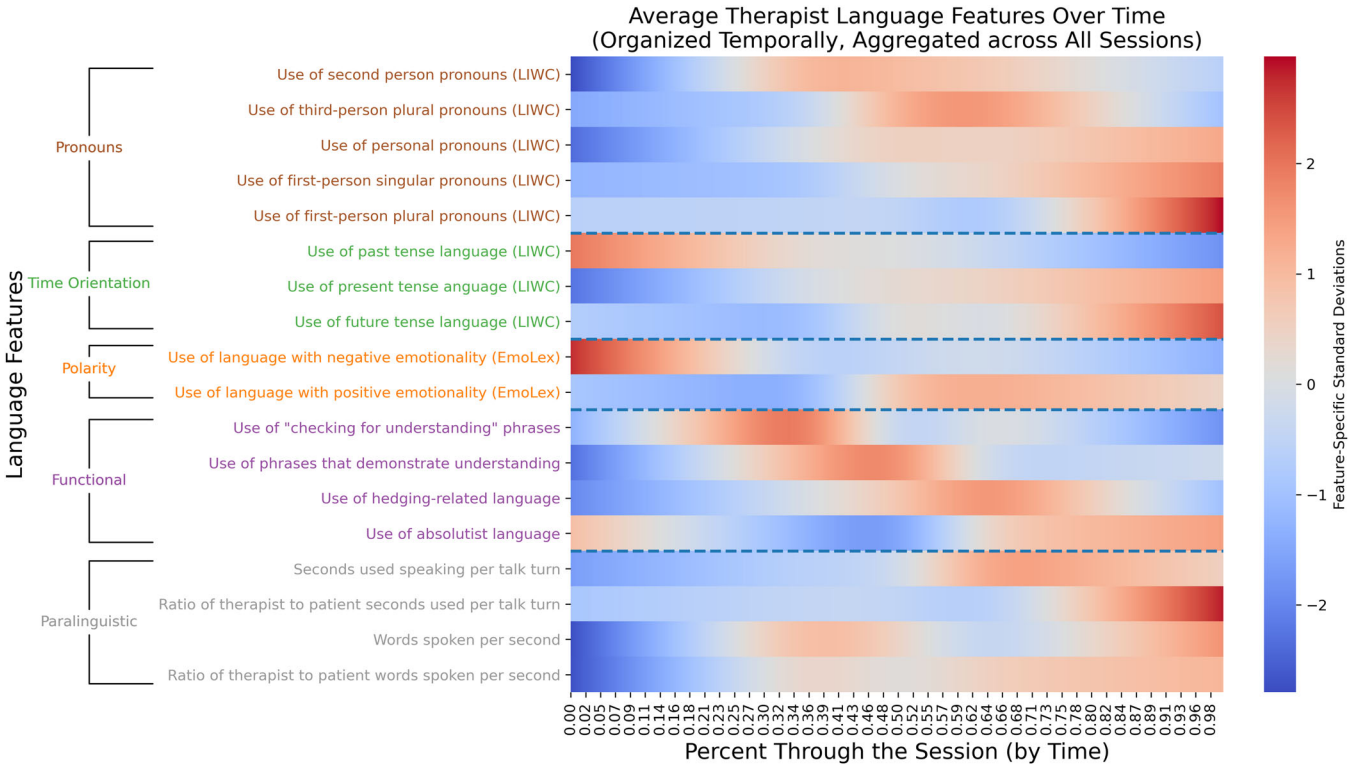

Fig. 1 Therapist speech phase-dependence. The dynamic nature of therapist speech, grouped by language feature category. It represents trends in therapist language over time after aggregating across therapists. LIWC = Linguistic Inquiry and Word Count, a dictionary-based lexicon that maps words and word stems to psychologically relevant categories. EmoLex = Word-Emotion Association Lexicon, a list of English words mapped to crowdsourced sentiment annotations. We performed smoothing/interpolation between discrete points at the level of temporal quintiles using a natural cubic spline. See Fig. 2 for per-feature examples of these trends viewed without smoothing.

in psychotherapy. Here we present a three-phase approach that measures therapist language by building on prior theoretical, methodological, and clinical insights. Phase 1 - To identify a priori language features of interest, we generate a non-exhaustive list of clinically relevant language features. Phase 2 - To observe the natural occurrence of language features identified in Phase 1, we describe the underlying structure of therapy focusing on timing, responsiveness, and consistency. Phase 3 - To demonstrate the potential for clinical utility, we evaluate the relationship between therapist language and patient symptom severity and diagnosis.

Many forms of therapy exist, along with an abundance of theoretically and practically motivated therapist approaches. Thus, we suggest a reasonable but non-exhaustive list of domain-focused concepts that balance face-validity and technical implementability using modern linguistic and statistical approaches. We posit, based on prior research, and our clinical judgment, that five clusters of language features may be clinically important across theoretical orientations, meriting close inspection (for details, see Methods, Phase 1: Feature generation). We limit our focus to characterizations of human language most amenable to machine learning, and that may correlate with favorable patient improvement. We acknowledge that other modern sensing technologies will allow for more rich characterization of human interaction such as facial expressions, body movement, and voice tone that may also be related to therapy outcomes[43].

The five feature clusters we seek to describe are pronouns, time orientation, emotional polarity, therapist tactics, and paralinguistic style. Pronouns (e.g., I, me, you, them) reflect internal psychological attention[37,44,45]. Measuring the relative frequency of self-focused pronouns (i.e., I, me, my) and other-focused pronouns (e.g., you, your, they) has demonstrated theoretical and practical value in psychological research[44,46,47]. Time orientation is a longstanding focus of psychotherapy. Some theoretical orientations advise therapists to focus on past experiences (e.g., early childhood), while some encourage focus on the present[48–51]. Emotions are important in most clinical psychology theoretical orientations[48,49,52–54]. There is strong disagreement, however, on how to represent and measure polarity and emotionality in clinical contexts[55–57]. Therapist tactics are used to help develop a therapeutic relationship and engender patient change, including statements that demonstrate understanding[19,41]. Paralinguistics refers to the way words are said, not the words themselves, for example, rate of speech[35,58,59]. Based on prior work, these language-focused constructs are theoretically important, but poorly measured moment-to-moment in psychotherapy. Although a full review of the theoretical importance and practical application of these clusters is beyond the scope of this work, we briefly summarize each feature in our Methods (Phase 1: Feature generation).

Uncovering modifiable, therapist-focused interventions that are associated with patient improvement is a key objective of therapy process research[21,23,41,60]. Our approach presents a systematic way to generate or evaluate hypotheses about psychotherapy process at scale. This study identifies potentially modifiable features of interest in psychotherapy (Phase 1), measures feature timing, responsiveness, and consistency (Phase 2), tests clinical usefulness (Phase 3), and shares methods to encourage critical peer review and collaboration.

RESULTS

Overall, our results surface linguistic nuance in psychotherapy that previously has not been directly measured. Therapist language timing is dynamic (Fig. 1) and does not mirror patient language consistently (Fig. 2). Therapist language appears to be responsive

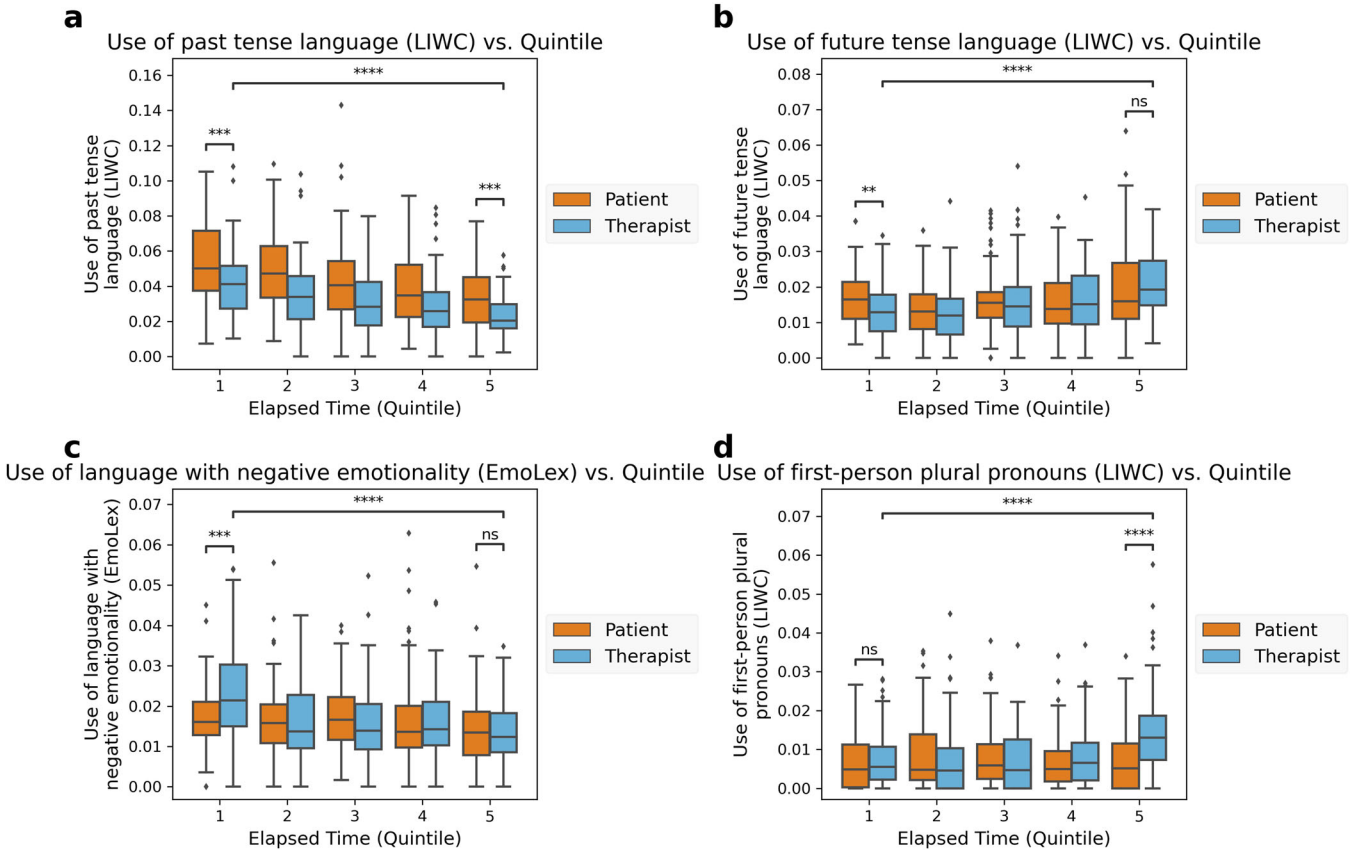

Fig. 2 Therapist and patient language within-session changes. Quantitative assessment of changes in therapist language features over time, as well as within-quintile differences between patient and therapist language. **b** and **c** show examples of patient and therapist language features that converged over time. **d** illustrates a case where patient and therapist language features diverged over time. **a** highlights a language feature that was significantly different between therapist and patient and neither converged nor diverged over the course of the session. The center line of each boxplot shows the median value for that time bin, while the lower and upper bounds of the box indicate the first quartile (25th percentile) and third quartile (75th percentile), respectively. The lower and upper "whiskers" extend to 1.5x the interquartile range (IQR) beyond the lower and upper quartile, respectively. Observations outside this range are displayed as independent points. All differences annotated with asterisks (*) are significant at level $\alpha = 0.05$ after controlling for multiple hypothesis tests via the Benjamini-Hochberg procedure. p-value annotation: Non-significant (ns): $0.01 < p \leq 1.0$; *$0.01 < p \leq 0.05$; **$0.001 < p \leq 0.01$; ***$0.0001 < p \leq 0.001$; ****$p \leq 0.0001$.

to patient language for a number of clinically relevant language features (Figs. 3, 4). For example, Figs. 3 and 4 show that therapists decreased their rate of speech, as measured by words per second, in response to increases in the patient's rate of speech, or vice versa (i.e., therapists significantly slowed their speech as patients quickened theirs). Therapist language appears consistent across sessions: on average, within-therapist language patterns were significantly more similar than between-therapist language patterns. In relation to patient-focused characteristics, therapist language appears to be related to patient diagnosis: logistic regression models trained to classify diagnosis based on therapist language patterns performed significantly better than chance.

Study population

Therapy transcripts were created per protocol as part of a secondary analysis of a previously completed randomized controlled trial, conducted in the United States across 24 college counseling clinics from April 2013 to December 2016[22,61]. See [Miner et al., 2020][22] for details on transcription and sample selection. Our primary sample had 78 sessions, each with a unique therapist and patient. A secondary sample added an additional 20 sessions, each of which represented a second session from a therapist in the primary sample but with a different patient relative to the first. Results given are with respect to the primary sample of 78 unless explicitly stated otherwise. The demographic information of these 98 therapist-patient dyads, and their clinical information (diagnosis and symptom severity), is presented in Table 1. Patients were predominantly female (87%) and in their early 20 s (median age, 21 years). Therapists were predominantly female (78%), and in their early 40 s (median age, 41 years). Patient depressive symptom severity was mostly minimal to mild.

Therapist timing is dynamic

Here we evaluate therapist language timing. Therapists appear to use distinct types of language at specific points in the session (early vs. late feature frequency). Figure 1 presents normalized frequency over time of therapist language features. Supplementary Fig. 1 shows individual therapists as examples. Figure 2 presents differences between therapist and patient language features over time for a subset of features.

Therapist speech changed significantly between the start and the end of the session. As illustrated in Fig. 1 and Supplementary Table 1, relative to the first quintile of the session, therapists in the last quintile of the session used a smaller proportion of words with negative emotionality (0.0136 vs. 0.0227, $p = 3.97 \times 10^{-7}$); a greater proportion of present-focused words (0.1697 vs. 0.1271, $p = 1.30 \times 10^{-15}$) and future-focused words (0.2084 vs. 0.01314, $p = 2.46 \times 10^{-7}$), but a smaller proportion of past-focused words

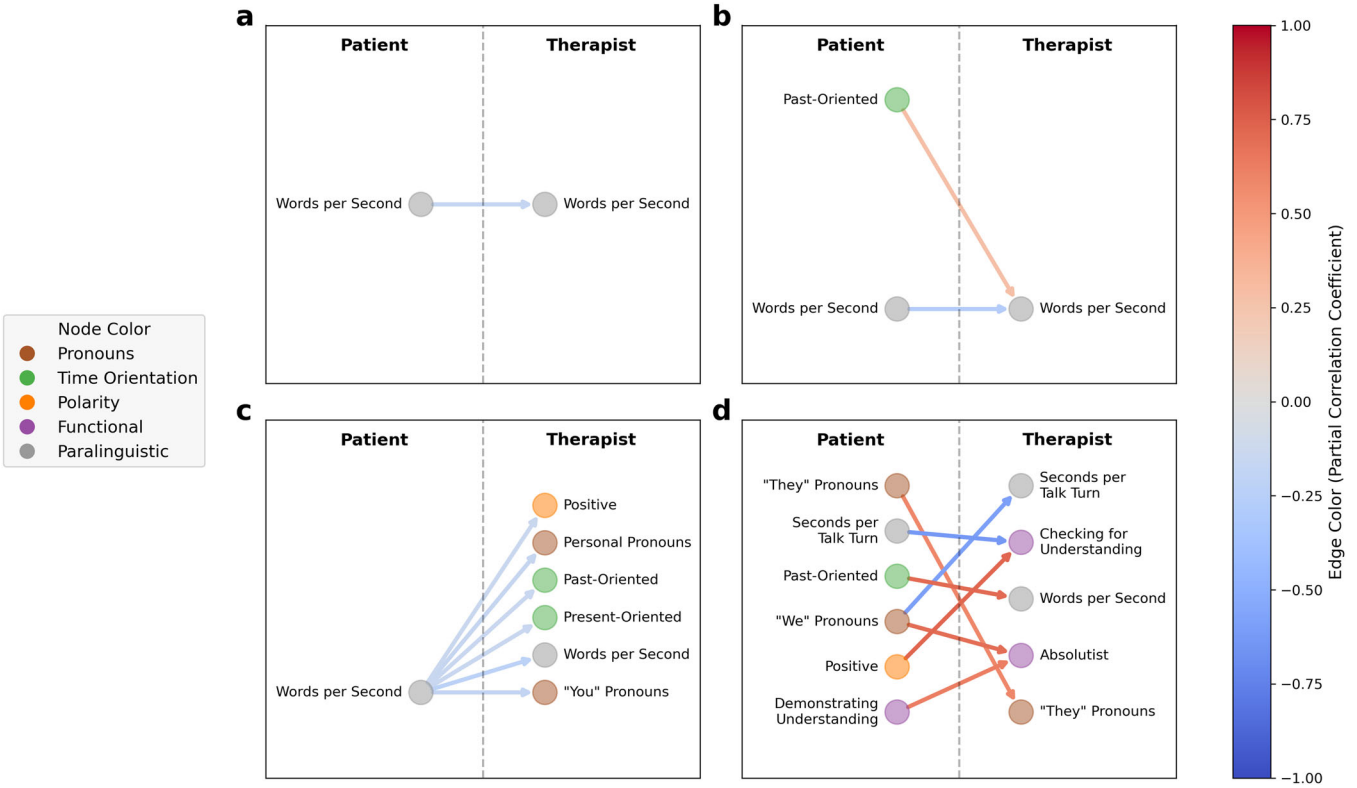

Fig. 3 Therapist responsiveness patterns at the level of individual sessions. Illustration of significant directional associations between patient language and therapist language in four sessions, each representing a unique patient-therapist dyad. Language features are colored by feature group (see Table 2). Edges are colored according to the average partial correlation coefficient. **a** illustrates an example of one patient-therapist dyad in which there was just one significant association: increases in patient rate of speech, as measured in words per second, were associated with decreases in therapist rate of speech, and vice versa. **b** shows a patient-therapist dyad in which the patient's past-oriented speech and rate of speech had opposite effects on the therapist's rate of speech. **c** demonstrates a case where decreases in the patient's rate of speech led to increases in a diverse array of therapist language features, or vice versa. **d** highlights a patient-therapist dyad with varied significant associations: increased patient use of third-person plural pronouns ('"They" Pronouns') drove increased therapist use of third-person plural pronouns ('"They" Pronouns'), increased use of positive language by the patient ("Positive") was associated with increased use of checking for understanding phrases by the therapist ("Checking for Understanding"), etc. These are four of the 73 network diagrams produced, one for each session/patient-therapist dyad.

(0.0231 vs. 0.0416, $p = 6.87^{-11}$); and a greater proportion of personal pronouns (0.1500 vs. 0.1182, $p = 3.86 \times 10^{-10}$), including first-person singular pronouns (0.0415 vs. 0.0238, $p = 1.93 \times 10^{-8}$), first-person plural pronouns (0.0150 vs. 0.0072, $p = 8.25 \times 10^{-8}$), and second-person pronouns (0.0808 vs. 0.0748, $p = 1.88 \times 10^{-2}$). Additionally, relative to the first quintile of the session, therapists in the final quintile tended to speak for longer durations, measured both in terms of raw seconds per talk turn (7.1615 seconds vs. 4.8952 seconds, $p = 7.35 \times 10^{-4}$) as well as the ratio of therapist-to-patient seconds per talk turn (1.879 vs. 0.938, $p = 4.95 \times 10^{-6}$). While therapists tended to speak longer in each talk turn near the end of the session, they also tended to speak faster relative to the patient, such that the ratio of therapist words per second to patient words per second was higher in the last quintile relative to the first (1.1715 vs. 1.040, $p = 9.62 \times 10^{-3}$). These results were all significant after controlling the False Discovery Rate at level $\alpha = 0.05$ via the Benjamini-Hochberg procedure.

The aggregate trends in therapist language highlighted above were in some cases also present in patient language, but the starting point and relative alignment (i.e., parallel, convergent, divergent) varied significantly depending on the language feature under consideration. See Fig. 2 and Supplementary Table 1 for additional details.

Although therapist language appears dynamic within sessions, patient language does not always follow the same trends. Figure 2 presents therapist-patient within-session language changes organized by quintile. Therapists' use of negative and past-oriented language decreased significantly over the course of the session (Figs. 2a and 2c), while their use of future-oriented language and first-person plural pronouns increased significantly (Figs. 2b and 2d). In some cases, patient and therapist language features converged over time (e.g., Figs. 2b and 2c: therapists used significantly less future-oriented language and significantly more negative language early in the session relative to patients, but these differences disappeared later in the session). In other cases, patient and therapist language diverged (e.g., Fig. 2d: there were no significant differences between patient and therapist use of first-person plural ("We") pronouns early in the session, but near the end of the session therapists used significantly more first-person plural pronouns than patients). In yet other cases, therapist and patient language differed significantly but seemed neither to converge nor diverge (e.g., Fig. 2a: use of past-oriented language).

Therapist speech is responsive

Here we evaluate therapist language responsiveness, specifically the degree to which changes in patient speech patterns are associated with subsequent changes in therapist speech patterns after controlling for potential confounding factors. Out of the 78 sessions we considered, two were excluded because they exhibited non-stationarity after differencing (differencing is a common technique in time series analysis for removing macro-

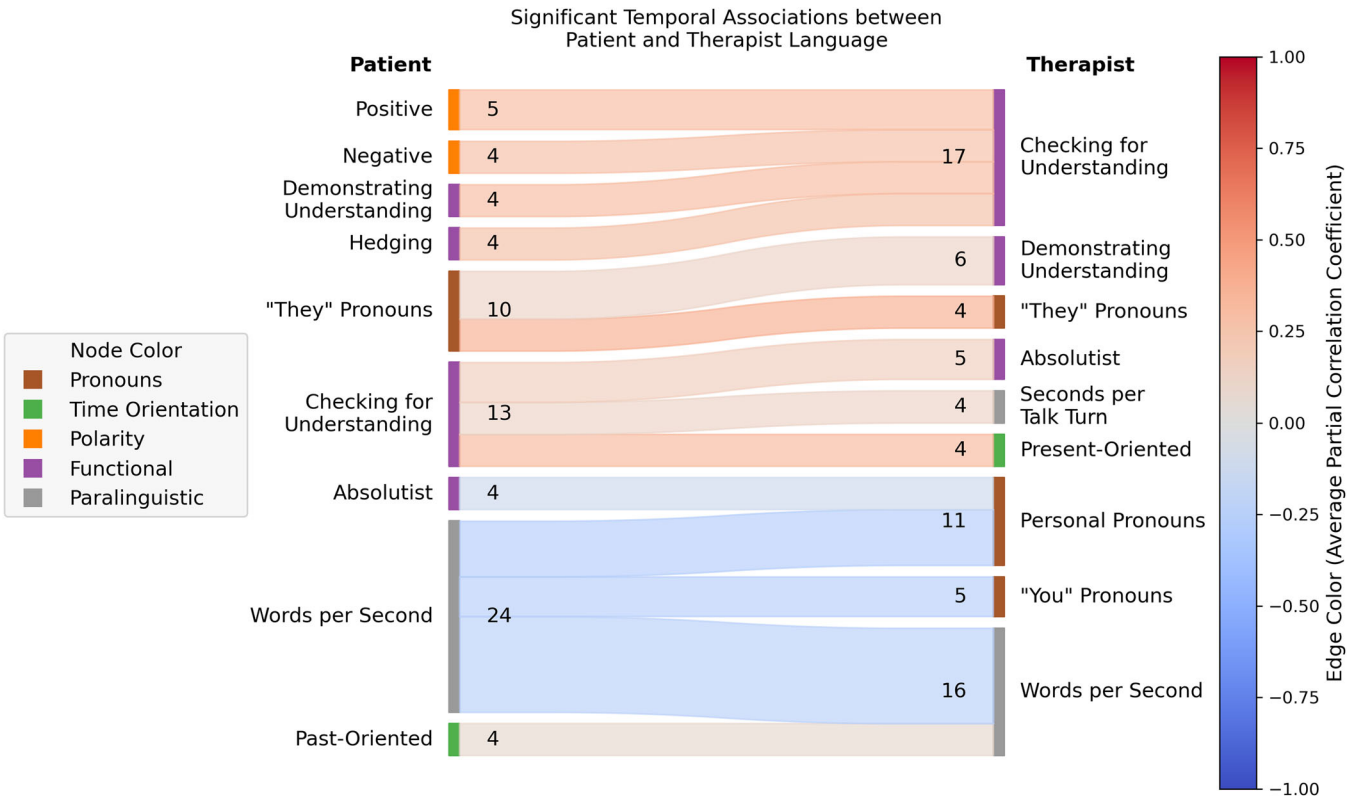

Fig. 4 Therapist responsiveness patterns aggregated over all sessions. The number of times a particular type of association between patient language features and subsequent/accommodating therapist language features was found, across all sessions. Patient language features are on the left, therapist language features on the right. For the purposes of illustration, only associations that were found in at least 4 patient-therapist dyads are displayed (see Supplementary Fig. 2 for a similar plot containing all significant associations). There were 72 such associations from 43 unique patient-therapist dyads, of which 24 involved changes in the patient's rate of speech ("Words per Second"). Language features are colored by feature group (see Table 2). Edges are colored according to the average partial correlation coefficient amongst all patient-therapist dyads in which that association was found. For example, 12 patient-therapist dyads exhibited a significant negative association between patient rate of speech and therapist rate of speech, such that increases in the patient's words per second ("Words per Second") were associated with subsequent decreases in the therapist's words per second ("Words per Second") and/or vice versa (i.e., decreases in the patient's words per second were associated with subsequent increases in the therapist's words per second).

level trends from time series whereby differences between consecutive observations are computed and treated as the primary subject of analysis; see Supplementary Methods for additional details)[62]. Another three were excluded because the patient and/or therapist had one or more language features with zero variance. Across the remaining 73 sessions analyzed, of the 18,688 possible dyad-specific associations between patient and therapist language features (16 language features each for patient and therapist, for 73 dyads) that were tested, 303 (1.6%) were significant after controlling the false discovery rate at level $\alpha = 0.05$. The mean (median) number of significant associations per therapist-patient dyad was 4.2 (3.0), with the minimum number of links in a session being 0, the maximum being 16, and the interquartile range (25th percentile, 75th percentile) being (2, 5). See Supplementary Fig. 3 for the distribution of the number of significant links per session. Figure 3 shows directed acyclic graphs illustrating the set of associations for a subset of the therapy sessions. As illustrated in Fig. 3, while the exact combinations of significant associations describing each therapist's accommodation patterns were almost all unique, some forms of accommodation (i.e., the therapist modulating their speech patterns in response to changes in patient speech patterns) were more common than others. The top three most frequent accommodation patterns were as follows: of 78 therapists in the sample, (1) 12 therapists significantly decreased their rate of speech (as measured by words per second) in response to increases in the patient's rate of speech, or vice versa (mean [SD] partial correlation: −0.24 [0.069]); (2) seven therapists significantly decreased their use of personal pronouns in response to increases in the patient's rate of speech, or vice versa (mean [SD] partial correlation: −0.27 [0.064]); (3) six therapists significantly altered the frequency with which they used phrases that demonstrate understanding in response to increases/decreases in their patients' use of third-person plural pronouns, though we note that four therapists increased their use of such phrases in response to increased patient third-person plural pronoun use (or vice versa) while two therapists' use of such phrases moved in the opposite direction (mean [SD] partial correlation: 0.10 [0.34]). Figure 4 presents the frequency with which certain associations between patient language features and subsequent/accommodating therapist language features appeared, across all sessions (for the sake of readability, only associations represented by at least three dyads are presented - see Supplementary Fig. 2 for all associations).

Therapists are consistent between sessions

Here we evaluate therapist language consistency across sessions. The average pairwise correlation of language patterns between therapists in our primary sample, averaged across 3003 (78 choose 2) distinct pairs of therapists, was −0.012 (95% CI: [−0.0218, −0.0024]), while the average pairwise correlation within therapists (comparing language patterns from two sessions with the same therapist but different patients) was 0.253 (95% CI: [0.1299,

Table 1. Clinician and patient demographic information.

Dataset			N	Min, 25%, Median, 75%, Max*
	Sites		24	
	Therapists		78	
	Patients		98	
	Session Duration in minutes			13, 39, 47, 53, 69
Patients			98	
	Gender			
		Male	13	
		Female	85	
	Age			18, 20, 21, 25, 52
	Session PHQ-9			0, 3, 7, 9, 25
		Minimal (PHQ-9 < 5)	35	
		Mild (PHQ-9 ≥ 5, PHQ-9 < 10)	32	
		Moderate (PHQ-9 ≥ 10, PHQ-9 < 15)	16	
		Moderately Severe (PHQ-9 ≥ 15, PHQ-9 < 20)	5	
		Severe (PHQ-9 ≥ 20)	2	
		Missing	10	
Therapists			78	
	Gender			
		Male	17	
		Female	61	
	Age			25, 34, 41, 51, 72
	Education			
		MA	4	
		MS	8	
		MSW	8	
		Ed. D.	2	
		Ph.D.	38	
		Psy.D.	15	
		Other	3	

*Min = Minimum value, 25% = value at 25% range, Median = Median value, 75% = value at 25% range, Max = Maximum value.

0.3794]) across 20 samples. A t test comparing the two distributions revealed that this difference was significant at level $\alpha = 0.05$ ($t = 4.39$, $p = 1.15 \times 10^{-5}$), suggesting that on average, within-therapist language patterns were significantly more similar than between-therapist language patterns.

Clinical relevance: Diagnoses and symptom severity

Here we evaluate therapist language as it relates to patient diagnosis and symptom severity. Logistic regression models trained to classify diagnosis based on therapist language patterns performed significantly better than chance in terms of accuracy on a held-out evaluation set (72.04% vs. 55.26%), with an average [95% CI] model accuracy improvement over chance (i.e., always guessing the majority class) of 16.78% [5.13%, 28.21%] ($p = 0.008$). Logistic regression models trained to classify symptom severity also performed better than chance in terms of accuracy (81.97% vs. 74.45%), though the improvement of model accuracy over chance accuracy (7.52%, 95% CI: [−2.56%, 17.95%], $p = 0.094$) was not significant at level $\alpha = 0.05$.

DISCUSSION

In this work, we provide researchers a transparent computational approach for representing, measuring, and analyzing therapist language in psychotherapy without time-consuming human inspection. We apply our approach to directly measure and analyze therapist language - both individually and in aggregate, and at multiple time scales (at the level of entire sessions, session quintiles, and utterances). We examine three clinically relevant but computationally neglected aspects of therapeutic discourse analysis: therapist language timing, responsiveness, and consistency across five clinically relevant domains: pronouns, time orientation, emotional polarity, therapist tactics, and paralinguistic style. We demonstrate the feasibility and potential clinical utility of this approach by evaluating the association between therapist language and two aspects of patient treatment: diagnosis and symptom severity. We conclude that increased use of computational language analysis of therapy will allow researchers and clinicians to transition from simply knowing what was said, to understanding what is most therapeutic[63].

Although therapists need to decide what to say and when to say it, the temporal sequencing (i.e., timing) of therapist language has been poorly measured[21,23]. Moreover, clinical features of interest are typically analyzed in isolation, leaving potential sequencing or interactions unexplored. Our approach puts multiple clinically relevant features in context across an entire session (Fig. 1), substantiating claims from discourse analysis and linguistics that words and phrases have layered and hierarchical interpretations[63(p350),64]. We find that prospectively identified language features (i.e., pronouns, therapist tactics, etc.) display a layered and temporally nuanced pattern that may be clinically relevant, meriting further inspection in observational or controlled studies.

Therapist-patient dyads actively adjust their speech based on emergent characteristics of the conversation[64]. Yet the specific language used by a therapist may be deployed in non-obvious ways in response to their conversation partner[58]. Our findings suggest that clinically relevant language features from each speaker appear to follow both similar and different trends between language features (Fig. 2). We see evidence of multiple alignments and directions of change when therapist and patient language are directly compared. Therapist and patient language are sometimes misaligned (Fig. 2a), convergent (i.e., start apart and become similar) (Figs. 2b and 2c), or divergent (i.e., start similar and diverge) (Fig. 2d). This finding is consistent with dyadic communication research in and outside of therapy, which uses related concepts such as language accommodation, entrainment, linguistic synchrony, adjustment, style matching, and affordances[30,35,63–70]. Despite a lack of harmony in concept terminology, our findings align with prior work suggesting that complex linguistic interactions are likely playing out during therapy. In prior psychotherapy research, higher empathy was observed when patients and therapists had more similar rates of speech[35]. Outside of clinical settings, in a study of romantic couples' texting patterns, couples' language converged over time towards a plateau, suggesting some normative or optimal level of linguistic alignment in romantic relationships[67]. Of note in our work, some language features converged, while others diverged, suggesting an opportunity for hypothesis generation and testing of language accommodation in psychotherapy[64,71]. For example, is emotional language convergence or divergence related to patient symptom improvement? Well-powered clinical trials or naturalistic data

Table 2. Summary of language features.

Feature name/description	Feature abbreviation	Examples	Feature group	Source
Second-person pronouns (LIWC)	"You" Pronouns	"you", "yours", "you'll", "y'all"	Pronouns	LIWC
Third-person plural pronouns (LIWC)	"They" Pronouns	"they", "their", "themselves", "they'll"	Pronouns	LIWC
Personal Pronouns (LIWC)	Personal Pronouns	All of the above, and third-person singular pronouns ("he", "she", "it")	Pronouns	LIWC
First-person singular pronouns (LIWC)	"I" Pronouns	"I", "I'll", "mine", "my", "myself"	Pronouns	LIWC
First-person plural pronouns (LIWC)	"We" Pronouns	"we", "us", "ours", "let's"	Pronouns	LIWC
Past-oriented language (LIWC)	Past-Oriented	"ago", "yesterday", "remember"	Time Orientation	LIWC
Present-oriented language (LIWC)	Present-Oriented	"now", "current", "is"	Time Orientation	LIWC
Future-oriented language (LIWC)	Future-Oriented	"we'll", "upcoming", "eventual"	Time Orientation	LIWC
Negative emotionality (EmoLex)	Negative	"frustrated", "scream", "hurt", "loathe"	Emotional Polarity	EmoLex
Positive emotionality (EmoLex)	Positive	"calm", "peace", "love", "enjoy", "satisfied"	Emotional Polarity	EmoLex
"Checking for understanding" phrases (active listening)	Checking for Understanding	"it sounds like", "that seems", "heard you correctly", "you sound", "let me make sure"	Therapist Tactics	Althoff et al.[30]
"Demonstrating understanding" phrases (active listening)	Demonstrating Understanding	"I hear you", "I see", "I understand"	Therapist Tactics	This study
"Hedging" phrases (active listening)	Hedging	"maybe", "from my perspective", "apparently"	Therapist Tactics	Althoff et al.[30]
"Absolutist" phrases (non-judgmental stance)	Absolutist	"absolutely", "always", "completely", "everyone", "must", "never", "nothing"	Therapist Tactics	Al-Mosaiwi, & Johnston[105]
Average seconds per talk turn	Seconds per Talk Turn	N/A	Paralinguistic Style	This study
Therapist to patient ratio of seconds per talk turn	Seconds per Talk Turn (Ratio)	N/A	Paralinguistic Style	This study
Average number of words spoken per second	Words per Second	N/A	Paralinguistic Style	This study
Therapist to patient ratio of words spoken per second	Words per Second (Ratio)	N/A	Paralinguistic Style	This study

repositories would help discern which patterns are most associated with clinical effectiveness.

Therapist responsiveness to a patient's personal experience is a crucial difference between in-person therapy and more easily accessible mental health treatments such as bibliotherapy or internet-delivered treatment[72]. Despite the importance of patient language in therapy discourse analysis, the moment-to-moment association of therapist and patient language has been difficult to operationalize. Our findings suggest a non-obvious and complex relationship between therapists' and patients' language features (Figs. 3 and 4). For example, it does not appear that therapists are following simple rules such as mirroring patient language and speaking style exactly, which would be relatively easy to observe and teach to future clinicians. We build on prior work which often focuses on patient or therapist language in isolation, specific therapeutic approaches (e.g. motivational interviewing), or language convergence (e.g. linguistic alignment)[19,35,66,67,73–75]. Our findings suggest that many-to-one and one-to-many associations are playing out between therapist and patient language features.

We do not claim originality for the idea that therapist language is responsive. In early work in discourse analysis of psychotherapy, Pittenger and colleagues (1960) wrote "the details of how [language] adjustment takes place in any given instance are worth looking for… indeed, we should venture to assert that the sequential pattern of adjustment lies at the very heart of psychotherapy process[76(p245)]." More recent work by Xiao et al (2015) found that, averaged over an entire session, therapist rate of speech is positively correlated with patient rate of speech. Our findings complement and add nuance to this finding by showing that some therapists respond to momentary increases in patient rate of speech by temporarily decelerating their own rate of speech. Thus therapists may both match the patient's rate of speech in aggregate, while converging or diverging from patient's rate of speech moment-to-moment. This may have a smoothing effect on the overall dialogue speed over time, but such claims are purely conjecture and more research is warranted. What accounts for these micro and macro processes, and whether they are related to symptom improvement is unknown. Our contribution here is a method to enable analysis of such micro-level trends across features of interest in psychotherapy. Future work is needed to establish whether specific language adjustments are helpful, inert, or harmful to patients in psychotherapy.

If best practices are to be developed to improve therapist training and create useful markers of therapy quality, comparisons are needed across clinicians, patients, treatment settings, and time[77]. Our findings suggest some degree of linguistic stability (i.e., consistency) in therapists' use of within-session language. We refer to this as a therapist's 'signature', consistent with prior work finding linguistic 'signatures' of emotion regulation in laboratory-based emotion regulation tasks[53]. Therapists appear to be both idiosyncratic and consistent in their use of language. Some language patterns are similar across sessions (i.e., therapist signature), while some language patterns adjust to patient or other situational factors. Therapist signatures may reflect their lived experience, preferences, or clinical training. Whether certain signatures are more clinically effective, and whether they are modifiable, is an important direction of future research. For example, some clinicians may regularly use more empathic language, a learnable skill, which may improve patient outcomes[36,78].

Our study has several limitations in how features were selected; these potentially may confound variables and generalizability. Phase 1 - feature selection. A small group of clinicians identified clinically relevant language features based on their training and

personal experience. Other reasonable people almost certainly would have made different selections. Also, our selected features do not address multilingualism or cultural variation in language use[79–82]. Phase 2 - language evaluation. We caution against an overly reductionist view of therapy as primarily or exclusively language based. Visual, auditory, biological, demographic, cultural, and other contextual factors may enhance, mitigate, or contradict interpretations made from language alone. We do not evaluate, nor do we claim, that therapist language always directly causes patient language or symptom improvement. It may be that patient improvement is caused by unmeasured covariates, or that therapist language is responsive to patient improvement or decompensation. Other approaches exist for feature implementation and should be evaluated, especially in the context of accuracy and appropriateness across demographic and clinical patient characteristics[55,83–87]. For example, in our study sample, both therapists and patients were mostly female, limiting generalizability. Phase 3 - clinical relevance. Clinical symptom severity measures were gathered in a college counseling setting, and thus our findings may not be generalizable to other clinicians, patients, or treatment settings. In college counseling sites, symptom severity often ranges from mild-moderate, as is true in our sample. It is unknown whether results would differ in patients with more severe symptoms. Additionally, the sample of 98 sessions is small relative to other AI and machine learning-based studies, reflecting a well-documented limitation in psychotherapy process research[34].

If successful, computational language analysis of entire psychotherapy sessions may address long-standing criticisms of methodological rigor in psychotherapy evaluation centered on reproducibility[23,28,88]. If deployed ethically and fairly, this approach would assist evaluations of treatment adherence and quality in real-world treatment settings and controlled trials[23,89–94]. To appreciate the full diversity of expression in therapy, computationally-conducted, theoretically informed evaluation may be a practical necessity[22,95]. Natural language processing of therapy transcripts is currently feasible and should seek to establish how moment-to-moment therapist language relates to the therapeutic relationship and meaningful clinical improvement. Our goal is not to reduce opportunities for clinical spontaneity and improvisation but to develop methods to learn from skilled therapists. Our results suggest that therapist language timing, responsiveness, and consistency demonstrate patterns that merit more rigorous inspection across populations and contexts.

METHODS

Study design

This is a retrospective cohort study of patient-therapist dyads that uses psychotherapy transcripts gathered from a completed clinical trial. The original study objectives, methods, and results have been published previously[61,96]. Written informed consent was obtained per protocol in the original trial from both patients and therapists. The study presented here was designed and conducted independently of the original clinical trial's primary objectives and approved by the Stanford University IRB. Our study had three phases: feature generation, feature measurement, and clinical relevance. In Phase 1 (feature generation), our team, including clinical psychologists, a psychiatrist, and a biomedical informaticist, used a modified Delphi approach to generate a list of clinically relevant language features related to therapist skill (authors ASM, BA, SA, NS)[97]. This feature list was refined based on its ability to be implemented by an expanded team of clinicians, informaticists, and computer scientists (authors ASM, SF, JF, TA, JH, AH, NS). Each feature was then implemented based on prior research and researcher judgment (authors ASM, SF). Features were selected that maximized reproducibility and transparency[98]. In Phase 2 (feature measurement), features were measured and standardized for therapists and patients in 98 professionally transcribed psychotherapy transcripts. Each transcript represents a unique patient-therapist dyad. We quantitatively assessed the structure of therapist and patient language. To evaluate timing, we measured the occurrence and frequency of the clinically relevant language features noted above (grouped into pronouns, time orientation, emotional polarity, therapist tactics, and paralinguistic style) in full therapy sessions. To evaluate responsiveness, we evaluated whether changes in therapist language were associated with immediately preceding utterance-level changes in patient language. To measure consistency, we tested whether or therapists have a consistent linguistic signature across sessions with different patients. In Phase 3 (clinical relevance), the relationship between therapists' language and patients' clinical presentation (i.e., diagnosis and symptom severity) was evaluated. Diagnosis was rated by the therapist, and depression symptom severity was assessed in the original trial using a common symptom severity measure, the Patient Health Questionnaire (PHQ-9), a patient-reported assessment of symptom frequency[99]. The study protocol was approved by the Institutional Review Board at Stanford University.

Dataset

Audio recordings of psychotherapy were collected per protocol during a randomized controlled trial[96]. The sessions took place between April 2013 and December 2016 at 24 college counseling sites across the United States. Non-directed counseling was offered to participants presenting with symptoms of depression or eating disorders. Transcripts were created using professional human transcriptionists; details are provided in prior work[22]. For the current study, a convenience sub-sample of unique therapist-patient dyads was selected, yielding 78 session transcripts. For therapists with more than one patient or session in our sample, a single session was randomly selected. Thus, our primary sample had 78 sessions, across 78 unique therapists and 78 unique patients. We generated a secondary sample with an additional 20 sessions, each representing a second session from a therapist in the primary sample but with a unique patient. Unless otherwise explicitly stated, any analyses are with respect to the primary sample of 78 unique therapist/unique patient sessions.

Diagnosis was made by the treating clinician during the original clinical trial using the DSM-IV diagnostic criteria. Depression symptom severity was measured at the start of each session using the Patient Health Questionnaire-9 (PHQ-9), a common and validated measure of depression severity[99–101].

Phase 1: Feature generation

Due to a lack of validated clinical ontologies for psychotherapy, we first identified clinically relevant features using a modified Delphi approach[97]. Features reflect either clinically important constructs (e.g., emotions) or paralinguistics (e.g., rate of speech). Features were manually clustered into five domains based on prior research and clinical judgment: pronouns, time orientation, emotional polarity, specific tactics, and paralinguistics. Examples of features considered but not selected for final analysis were 'conveying warmth', 'tracks and remarks on therapeutic alliance ruptures', n-grams from the process measure The Multitheoretical List of Therapeutic Interventions - 30 items (MULTI-30) (Supplementary Table 2.).

Pronouns. The Linguistic Inquiry and Word Count (LIWC) program is a validated lexicon containing psychologically meaningful categories of words and word stems, including categories for various kinds of personal pronouns[44]. Our "Pronouns" features represent the number of matches between spoken words and terms in the relevant pronoun-specific LIWC category.

Time orientation. Time orientation of the patient and therapist language is a key focus of research in mental health[50,51]. Each "Time Orientation" feature represents the number of times a word/word stem from a relevant time orientation lexicon in LIWC appears in speech[44].

Emotional polarity. Emotions are important in most clinical psychology theoretical orientations[46,48,49,52]. Nevertheless, there is strong disagreement on how to measure emotionality[55,102,103]. We chose to use the NRC Emotion Lexicon (EmoLex) to measure whether a word conveyed positive or negative sentiment because of its expansive coverage (14,182 unigrams/words) and inspectable approach, rooted in a crowdsourced layman's understanding of each word. The "Positive emotionality" feature represents the number of words considered to have positive polarity, and similarly for the "Negative emotionality" feature.

Therapist tactics. We used small, non-exhaustive lexicons to detect two clinically important but rarely measured therapist tactics: active listening and non-judgmental stance, adapted from prior work[30]. Active listening entails speech acts that seek to validate the patient, clarify meaning, or direct the patient towards useful experiences[104]. A non-judgmental stance is created and maintained in many ways, but one approach is to avoid absolutist language (e.g., "always", "never")[105]. See Supplementary Methods for additional details.

Paralinguistic style. The meanings of words are influenced by how the words are said[106,107]. We focus on paralinguistic aspects of speech that can be measured using only transcripts. We measured the seconds taken by each therapist per talk turn, with talk turn boundaries delineated by a change in speaker in the transcript. We additionally measured therapists' rate of speech by dividing the number of therapist-spoken words by the amount of time that the therapist spoke, as indicated by the time stamps in the transcripts. We also measured the therapist-to-patient ratio of both seconds taken per talk turn and words spoken per second. Including these ratios provides insight into whether the therapist was speaking faster or slower than the patient, as well as taking more time in each talk turn compared to the patient.

Phase 2: Feature implementation

Temporal aggregation and granularity. In addition to analyzing therapist language at the level of talk turns/utterances, we aggregated features (1) at the level of session quintiles (e.g., the first 20% of the session, by time), and (2) at the entire session-level. We indicate which level of aggregation was used in each subsection of the methods. There is no standard approach, and prior work has used both quintiles and deciles to segment discourse analysis[30,45]. We analyzed sessions at the level of quintiles to reduce the variance of aggregate language feature statistics within each time window while nevertheless providing sufficient temporal granularity so as to make meaningful deductions about changes in language use over time.

Therapist speech changes. To represent therapist language, we calculated the average value of each language feature within each quintile of therapist speech. For count-based lexicon-matching features, we calculated the proportion of total words that matched a term appearing in the associated lexicon for each quintile.

To qualitatively analyze the dynamic nature of therapist language over time, we fit a natural cubic spline to the data represented by ordinally indexed session quintiles (independent variable) and quintile-aggregated language features, averaged across therapists (dependent variable)[108]. This procedure was also performed for individual therapist language features to additionally highlight heterogeneity in the way therapist language changes over time. See Fig. 1 and Supplementary Fig. 1.

We also quantitatively assessed patterns in therapist language features over time. For each language feature, we compared the distribution of that therapist language feature in the first and last quintile of therapy sessions, using a nonparametric Mann–Whitney *U* test to test for significant differences in distribution between the two quintiles[109]. Within the first and last quintiles, we also analyzed differences between patient and therapist language features using the Mann–Whitney *U* test. We used the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR) at level $\alpha = 0.05$. See Fig. 2.

Evaluating therapist speech responsiveness. Here we describe how our therapist language representations were used to analyze individual therapist's accommodation patterns at the level of utterances. To better answer the question of how therapists adapt their language to patient language, we leveraged recent methodological advances in time series causal discovery for dynamical systems to identify temporal dependencies between patient and therapist language features[110]. The algorithm we employed, PCMCI, applies momentary conditional independence (MCI) tests to identify temporal links between variables, accounting for potential observed confounding. PCMCI has been shown to identify such links in observational data with good statistical power and low Type I error. For each therapist, we used PCMCI with partial correlation to identify significant links between patient language and therapist language. Patient-to-therapist associations were recorded as significant if the associated MCI test was significant at level $\alpha = 0.05$, after controlling the FDR with the Benjamini-Hochberg procedure. We additionally calculated and reported the frequency of each type of association across all sessions.

Phase 3: Measuring clinical relevance

We next describe our approach to differentiating between therapy sessions. By aggregating therapists' language features over the entire time course of the session, we obtain a 16-dimensional vector for each therapist (i.e., the therapist's linguistic "signature"). We sought to examine: (1) whether a therapist's "signature" is consistent across patients, over and above chance; and (2) whether these "signatures" are associated with clinically relevant patient variables, namely symptom severity and psychiatric diagnosis.

To answer (1), we calculated the cross-therapist "signature" correlations between all pairs of therapists in our primary sample, then compared that distribution to the distribution of "signature" correlations within therapists but across different patients. We used a *t*-test to test whether there were any differences in the distribution of correlations between the two groups.

To answer (2), we performed two predictive analyses via logistic regression, treating the therapist "signatures" as independent variables and the patient symptom severity classification (admitting PHQ-9 < 10 vs. PHQ-9 ≥ 10) and admitting diagnosis (depression vs. eating disorder) as the dependent variables, respectively. We randomly divided our dataset into two equally sized halves, trained a logistic regression model on one half, and evaluated the model's accuracy on the second half. The test accuracy on the second half was compared to chance, which in this case we defined as always predicting the majority label of the dependent variable in the subsampled evaluation dataset. The difference between our model's accuracy and chance accuracy was recorded, and this process was repeated 1000 times, using random splits of the data each time. We used the resulting distribution of accuracy differences to estimate the probability that our logistic regression model would perform no better than chance, defining a significant result as $p < 0.05$.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The dataset is not publicly available due to patient privacy restrictions but may be available from the corresponding author on reasonable request. Additional requirements may be required, such as a data use agreement or limitations to non-commercial purposes.

CODE AVAILABILITY

The code used in this study can be found at: https://github.com/som-shahlab/psych-nlp.

Received: 12 May 2022; Accepted: 18 October 2022;
Published online: 02 December 2022

REFERENCES

1. Holmes, E. A. et al. The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *Lancet Psychiatry.* **5**, 237–286 (2018).
2. Lambert, M. J. Outcome in psychotherapy: the past and important advances. *Psychotherapy.* **50**, 42–51 (2013).
3. Cuijpers, P., van Straten, A., Andersson, G. & van Oppen, P. Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *J. Consult. Clin. Psychol.* **76**, 909–922 (2008).
4. Barth, J. et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS Med.* **10**, e1001454 (2013).
5. Arch, J. J. et al. Randomized clinical trial of cognitive behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for mixed anxiety disorders. *J. Consult. Clin. Psychol.* **80**, 750–765 (2012).
6. Bögels, S. M., Wijts, P., Oort, F. J. & Sallaerts, S. J. M. Psychodynamic psychotherapy versus cognitive behavior therapy for social anxiety disorder: an efficacy and partial effectiveness trial. *Depress Anxiety.* **31**, 363–373 (2014).
7. Markowitz, J. C. et al. Is Exposure Necessary? A Randomized Clinical Trial of Interpersonal Psychotherapy for PTSD. *Am J. Psychiatry.* **172**, 430–440 (2015).
8. Wampold, B. E. & Brown, G. S. J. Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. *J. Consult. Clin. Psychol.* **73**, 914–923 (2005).
9. Lambert, M. J. ed. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. John Wiley & Sons; 2013.
10. Barkham, M., Lutz, W., Lambert, M. J. & Saxon, D. Therapist effects, effective therapists, and the law of variability. In Castonguay L. G., Hill C. E., eds. *How and Why Are Some Therapists Better than Others?: Understanding Therapist Effects*. Vol 356. American Psychological Association, xv; 2017:13–36.
11. Owen, J. et al. Are high-performing therapists both effective and consistent? A test of therapist expertise. *J. Consult. Clin. Psychol.* **87**, 1149–1156 (2019).
12. Erekson, D. M., Clayson, R., Park, S. Y. & Tass, S. Therapist effects on early change in psychotherapy in a naturalistic setting. *Psychother Res.* **30**, 68–78 (2020).
13. Wampold, B. E. & Owen, J. Therapist effects: History, methods, magnitude. In: Barkham M., Lutz W., Castonguary L. G., eds. *Bergin and Garfield's Handbook of Psychotherapy*. John Wiley & Sons, Inc.; 2021:297–326.
14. Castonguay, L. G., Boswell, J. F., Constantino, M. J., Goldfried, M. R. & Hill, C. E. Training implications of harmful effects of psychological treatments. *Am Psychol.* **65**, 34–49 (2010).
15. Elliott, R. Psychotherapy change process research: realizing the promise. *Psychother Res.* **20**, 123–135 (2010).
16. Goldfried, M. R. & Wolfe, B. E. Toward a more clinically valid approach to therapy research. *J. Consult. Clin. Psychol.* **66**, 143–150 (1998).
17. Stiles, W. B. Verbal response modes and psychotherapeutic technique. *Psychiatry.* **42**, 49–62 (1979).
18. Benjamin, L. S. Structural analysis of social behavior. *Psychol. Rev.* **81**, 392–425 (1974).
19. Flemotomos, N. et al. Automated evaluation of psychotherapy skills using speech and language technologies. *Behav Res Methods*. Published online August 3, 2021. https://doi.org/10.3758/s13428-021-01623-4.
20. Rogers, C. R. The use of electrically recorded interviews in improving psychotherapeutic techniques. *Am. J. Orthopsychiatry.* **12**, 429–434 (1942).
21. Hofmann, S. G. & Hayes, S. C. The future of intervention science: process-based therapy. *Clin. Psychol. Sci.* **7**, 37–50 (2019).
22. Miner, A. S. et al. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digit. Med.* **3**, 82 (2020).
23. Goldfried, M. R. Obtaining consensus in psychotherapy: What holds us back? *Am. Psychol.* **74**, 484–496 (2019).
24. Kazdin, A. E. Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behav. Res. Ther.* **88**, 7–18 (2017).
25. Hill, C. E. & Lent, R. W. A narrative and meta-analytic review of helping skills training: Time to revive a dormant area of inquiry. Gaines AN, Goldfried MR, Constantino MJ. Revived call for consensus in the future of psychotherapy. *Evid. Based. Ment. Health.* **24**, 2–4 (2021).
26. Mehta, M., et al. Psychotherapy is Not One Thing: Simultaneous Modeling of Different Therapeutic Approaches. In *Proc. the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics; 2022:47–58.
27. Lee, M. & Martin, J. L. Coding, counting and cultural cartography. *Am. J. Cultural Sociology.* **3**, 1–33 (2015).
28. Goldberg, S. B. et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *J. Couns. Psychol.* **67**, 438–448 (2020).
29. Gaines, A. N., Goldfried, M. R. & Constantino, M. J. Revived call for consensus in the future of psychotherapy. *Evid Based Ment. Health.* **24**, 2–4 (2021).
30. Althoff, T., Clark, K., & Leskovec. J. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. In *Transactions of the Association for Computational Linguistics, Volume 4*.; 2016:463–476.
31. Eichstaedt, J. C. et al. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychol. Methods.* **26**, 398–427 (2021).
32. Imel, Z. E., Steyvers, M. & Atkins, D. C. Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions. *Psychotherapy.* **52**, 19–30 (2015).
33. Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M. S., & Chakraborty, T. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proc. the Fifteenth ACM International Conference on Web Search and Data Mining*.; 2021:735–745.
34. Doorn, K. A. van, Kamsteeg, C., Bate, J. & Aafjes. M. A scoping review of machine learning in psychotherapy research. *Psychother Res*. Published online August 29, 2020. https://doi.org/10.1080/10503307.2020.1808729.
35. Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P. G., Narayanan, S. S. Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In *Proc. Interspeech*; 2015:2489–2493.
36. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff T. Towards facilitating empathic conversations in online mental health support: a reinforcement learning approach. In *Proc. of the Web Conference 2021*. WWW '21. (Association for Computing Machinery, 2021:194–205).
37. Nook, E. C., Vidal Bustamante, C. M., Cho, H. Y. & Somerville, L. H. Use of linguistic distancing and cognitive reappraisal strategies during emotion regulation in children, adolescents, and young adults. *Emotion.* **20**, 525–540 (2020).
38. Lee, F. T., Hull, D., Levine, J., Ray, B. & McKeown, K. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proc. of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. (Association for Computational Linguistics, 2019:12–23).
39. Ewbank, M. P. et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry.* **77**, 35–43 (2020).
40. Roth, A. & Fonagy P. *What Works for Whom?: A Critical Review of Psychotherapy Research*. (Guilford Press, 2006).
41. Castonguay, L. G. & Hill, C. E., (eds) *How and Why Are Some Therapists Better than Others?: Understanding Therapist Effects*. Vol 356. (American Psychological Association, 2017).
42. Lattie, E. G., Stiles-Shields, C. & Graham, A. K. An overview of and recommendations for more accessible digital mental health services. *Nat. Rev. Psychol.* **26**, 1–14 (2022).
43. Haque, A., Milstein, A. & Fei-Fei, L. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature* **585**, 193–202 (2020).
44. Tausczik, Y. R. & Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang Soc. Psychol.* **29**, 24–54 (2010).
45. Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H. & Wolf, M. First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clin. Psychol. Psychother.* **24**, 384–391 (2017).
46. Vine, V., Boyd, R. L. & Pennebaker, J. W. Natural emotion vocabularies as windows on distress and well-being. *Nat. Commun.* **11**, 4525 (2020).
47. Wadden, D., August, T., Li, Q. & Althoff, T. The effect of moderation on online mental health conversations. In *Proc. of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)*.; 2021:751–763.

48. Beck, J. S. *Cognitive Behavior Therapy, Second Edition: Basics and Beyond*. (Guilford Press, 2011).
49. Weissman, M. M., Markowitz, J. C., Klerman, G. L. *Clinician's Quick Guide to Interpersonal Psychotherapy*. Vol 165; 2008:140–141.
50. Baird, H. M., Webb, T. L., Sirois, F. M. & Gibson-Miller, J. Understanding the effects of time perspective: A meta-analysis testing a self-regulatory framework. *Psychol. Bull.* **147**, 233–267 (2021).
51. Park, G. et al. Living in the Past, Present, and Future: Measuring Temporal Orientation With Language. *J .Pers.* **85**, 270–280 (2017).
52. Keefe, J. R. et al. In-session emotional expression predicts symptomatic and panic-specific reflective functioning improvements in panic-focused psychodynamic psychotherapy. Gross JJ, Jazaieri H. Emotion, Emotion Regulation, and Psychopathology: An Affective Science Perspective. *Clin. Psychol. Sci.* **2**, 387–401 (2014).
53. Nook, E. C., Schleider, J. L. & Somerville, L. H. A linguistic signature of psychological distancing in emotion regulation. *J. Exp. Psychol. Gen.* **146**, 337–346 (2017).
54. Greenberg, L. S. Emotions, the great captains of our lives: their role in the process of change in psychotherapy. *Am. Psychol.* **67**, 697–707 (2012).
55. Fiske, A. P. The lexical fallacy in emotion research: Mistaking vernacular words for psychological entities. *Psychol. Rev.* **127**, 95–113 (2020).
56. Gross, J. J. & Jazaieri, H. Emotion, Emotion Regulation, and Psychopathology: An Affective Science Perspective. *Clin. Psychol. Sci.* **2**, 387–401 (2014).
57. Greenberg, L. S. & Safran, J. D. Emotion in psychotherapy. *Am. Psychol.* **44**, 19–29 (1989).
58. Rocco, D. et al. Beyond Verbal Behavior: An Empirical Analysis of Speech Rates in Psychotherapy Sessions. *Front Psychol.* **9**, 978 (2018).
59. Tonti, M. & Gelo, O. C. G. Rate of speech and emotional-cognitive regulation in the psychotherapeutic process: a pilot study. *Res. Psychother.: Psychopathol. Process and Outcome.* **19**, 92–102 (2016). https://doi.org/10.4081/ripppo.2016.232.
60. Barkham, M., Lutz, W. & Castonguay, L. G. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. (John Wiley & Sons, 2021).
61. Wilfley, D. E. et al. Training models for implementing evidence-based psychological treatment: A cluster-randomized trial in college counseling centers. *JAMA Psychiatry.* **77**, 139–147 (2020).
62. Hyndman, R. J., Athanasopoulos G. *Forecasting: Principles and Practice*. (OTexts, 2018).
63. Labov, W. & Fanshel D. *Therapeutic Discourse: Psychotherapy as Conversation*. (Academic Press, 1977).
64. Clark, H. H. *Using Language*. (Cambridge University Press, 1996).
65. Borelli, J. L. et al. Therapist-client language matching: initial promise as a measure of therapist-client relationship quality. *Psychoanal. Psychol.* **36**, 9–18 (2019).
66. Ireland, M. E. et al. Language style matching predicts relationship initiation and stability. *Psychol. Sci.* **22**, 39–44 (2011).
67. Brinberg, M. & Ram, N. Do New Romantic Couples Use More Similar Language Over Time? Evidence from Intensive Longitudinal Text Messages. *J. Commun.* **71**, 454–477 (2021).
68. Koole, S. L. & Tschacher, W. Synchrony in Psychotherapy: A Review and an Integrative Framework for the Therapeutic Alliance. *Front Psychol*. 2016;7. https://doi.org/10.3389/fpsyg.2016.00862.
69. Doré, B. P. & Morris, R. R. Linguistic Synchrony Predicts the Immediate and Lasting Impact of Text-Based Emotional Support. *Psychol. Sci.* **29**, 1716–1723 (2018).
70. Shapira, N., Atzil-Slonim, D., Tuval Mashiach, R. & Shapira O. Measuring Linguistic Synchrony in Psychotherapy. In: *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. (Association for Computational Linguistics, 2022:158–176).
71. Giles, H., Mulac, A., Bradac, J. J. & Johnson, P. Speech accommodation theory: the first decade and beyond. *Annals Intern. Commun. Association.* **10**, 13–48 (1987).
72. Hatcher, R. L. Interpersonal competencies: Responsiveness, technique, and training in psychotherapy. *Am. Psychol.* **70**, 747–757 (2015).
73. Duran, N. D., Paxton, A. & Fusaroli, R. ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychol. Methods.* **24**, 419–438 (2019).
74. Burkhardt, H. A. et al. Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. *J. Med. Internet Res.* **23**, e28244 (2021).
75. Nook, E. C., Hull, T. D., Nock M. & Somerville L. Linguistic measures of psychological distance track symptom levels and treatment outcomes in a large set of psychotherapy transcripts. https://doi.org/10.31234/osf.io/hqxaz.
76. Pittenger, R. E., Hockett C. F. & Danehy J. J. *The First Five Minutes: A Sample of Microscopic Interview Analysis*. (Paul Martineau, 1960).
77. Horn, R. L. & Weisz, J. R. Can artificial intelligence improve psychotherapy research and practice? *Adm. Policy Ment. Health.* **47**, 852–855 (2020).
78. Sharma, A., Miner, A., Atkins, D. & Althoff T. A computational approach to understanding empathy expressed in text-based mental health support. In *The 2020 Conference on Empirical Methods in Natural Language Processing*.; 2020:5263–5276.
79. Costa, B. & Dewaele, J. M. Psychotherapy across languages: beliefs, attitudes and practices of monolingual and multilingual therapists with their multilingual patients. *Language Psychoanalysis.* **1**, 18–40 (2012).
80. Benish, S. G., Quintana, S. & Wampold, B. E. Culturally adapted psychotherapy and the legitimacy of myth: a direct-comparison meta-analysis. *J. Couns. Psychol.* **58**, 279–289 (2011).
81. Whaley, A. L. & Davis, K. E. Cultural competence and evidence-based practice in mental health services: a complementary perspective. *Am. Psychol.* **62**, 563–574 (2007).
82. Hook, J. N. et al. Cultural humility and racial microaggressions in counseling. *J. Couns. Psychol.* **63**, 269–277 (2016).
83. Koenecke, A. et al. Racial disparities in automated speech recognition. *Proc. Natl Acad. Sci.* **117**, 7684–7689 (2020).
84. Corcoran, C. M., Benavides, C. & Cecchi, G. Natural language processing: opportunities and challenges for patients, providers, and hospital systems. *Psychiatric Annals.* **49**, 202–208 (2019).
85. Mitchell, M. et al. Diversity and Inclusion Metrics in Subset Selection. In *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. Association for Computing Machinery; 2020:117-123.
86. Rubin, E. Striving for Diversity in Research Studies. *N. Engl .J. Med*. Published online September 13, 2021. https://doi.org/10.1056/NEJMe2114651.
87. Brown, S., Davidovic, J. & Hasan, A. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*. **8**, 2053951720983865 (2021).
88. Krause, K. R., Chung, S., Sousa Fialho, M. L., Szatmari, P. & Wolpert, M. The challenge of ensuring affordability, sustainability, consistency, and adaptability in the common metrics agenda. *Lancet Psychiatry*. **8**, 1094–1102 (2021).
89. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. A. & Shah, N. H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Inform. Assoc.* **27**, 2011–2015 (2020).
90. Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C. & Narayanan, S. S. "Rate My Therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS One.* **10**, e0143055 (2015).
91. Bone, C. et al. Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health.* **3**, e231–e240 (2021).
92. Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med.* **2**, 88 (2019).
93. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Machine Intelligence.* **1**, 389–399 (2019).
94. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data.* **3**, 160018 (2016).
95. Krieger, N. *Epidemiology and the People's Health: Theory and Context*. (Oxford University Press, 2011).
96. Wilfley, D. E. et al. Training models for implementing evidence-based psychological treatment for college mental health: A cluster randomized trial study protocol. *Contemp Clin Trials.* **72**, 117–125 (2018).
97. Linstone H. A. & Turoff M. *The Delphi Method: Techniques and Applications*. First Edition. (Addison-Wesley Educational Publishers Inc, 1975).
98. Wallach, J. D., Boyack, K. W. & Ioannidis, J. P. A. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLoS Biol.* **16**, e2006930 (2018).
99. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
100. Stochl, J. et al. On Dimensionality, Measurement Invariance, and Suitability of Sum Scores for the PHQ-9 and the GAD-7. *Assessment*. 29:355–366. Published online December 3, 2020:1073191120976863.
101. Zimmerman, M. Using the 9-Item Patient Health Questionnaire to Screen for and Monitor Depression. *JAMA*. Published online October 18, 2019. https://doi.org/10.1001/jama.2019.15883.
102. Aafjes-van Doorn, K. & Barber, J. P. Systematic review of in-session affect experience in cognitive behavioral therapy for depression. *Cognit. Ther. Res.* **41**, 807–828 (2017).
103. Burkhardt, H., Pullmann, M., Hull, T., Aren, P., Cohen, T. Comparing emotion feature extraction approaches for predicting depression and anxiety. In *Proc. of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. (Association for Computational Linguistics, 2022:105–115).
104. Fitzgerald, P. & Leudar, I. On active listening in person-centred, solution-focused psychotherapy. *J. Pragmat.* **42**, 3188–3198 (2010).

105. Al-Mosaiwi, M. & Johnstone, T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clin. Psychol. Sci.* **6**, 529–542 (2018).
106. Mehrabian, A. & Ferris, S. R. Inference of attitudes from nonverbal communication in two channels. *J. Consult. Psychol.* **31**, 248–252 (1967).
107. Mehrabian, A. & Wiener, M. Decoding of inconsistent communications. *J. Personality Social. Psychol.* **6**, 109–114 (1967).
108. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer Science & Business Media, 2009).
109. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
110. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**, eaau4996 (2019).

ACKNOWLEDGEMENTS

A.S.M. was supported by grants from the National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award (KL2TR001083 and UL1TR001085), the Stanford Department of Psychiatry Innovator Grant Program, and the Stanford Human-Centered AI Institute. S.L.F. was supported by a Big Data to Knowledge (BD2K) grant from the National Institutes of Health (T32 LM012409) and a National Defense Science and Engineering Graduate Fellowship from the Department of Defense. N.H.S acknowledges support from the Mark and Debra Lesli Endowment for the Program for AI in Healthcare at Stanford Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Fei-Fei Li and G. Terence Wilson for project guidance. We thank members of 2018 Spring AI Bootcamp, Pranav Rajpurkar, Andrew Ng, Suvadip Paul, Ben Cohen-Wang, Matthew Sun for collaboration. We thank Jon-Michael Knapp for assistance editing the manuscript. We thank all of the participating counseling centers, directors, therapists, and student patients.

AUTHOR CONTRIBUTIONS

A.S.M. and S.L.F. contributed equally as co-first authors. A.S.M., S.L.F, A.H., B.A.A., W.S.A., N.H.S., J.A.F. conceptualized and designed the study. A.S.M., S.L.F., A.H., J.A.F., B.A.A., J.H., and N.H.S. acquired, analyzed, or interpreted the data. A.S.M., S.L.F., A.H., J.A.F., B.A.A., and N.H.S. drafted the manuscript. All authors performed critical revision of the manuscript for important intellectual content. S.L.F., A.H., and J.A.F. performed statistical analysis. B.A.A. and N.H.S. provided administrative, technical, and material support. B.A.A., W.S.A., and N.H.S. supervised the study. A.S.M, S.L.F, and N.H.S. had full access to all the data. A.S.M. and S.L.F. take responsibility for the integrity of the data and the accuracy of the data analysis.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44184-022-00020-9.

Correspondence and requests for materials should be addressed to Adam S. Miner.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022