JAMA Psychiatry | Original Investigation

Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning

Michael P. Ewbank, PhD; Ronan Cummins, PhD; Valentin Tablan, PhD; Sarah Bateup, MA; Ana Catarino, PhD; Alan J. Martin, BSc; Andrew D. Blackwell, PhD

IMPORTANCE Compared with the treatment of physical conditions, the quality of care of mental health disorders remains poor and the rate of improvement in treatment is slow, a primary reason being the lack of objective and systematic methods for measuring the delivery of psychotherapy.

OBJECTIVE To use a deep learning model applied to a large-scale clinical data set of cognitive behavioral therapy (CBT) session transcripts to generate a quantifiable measure of treatment delivered and to determine the association between the quantity of each aspect of therapy delivered and clinical outcomes.

DESIGN, SETTING, AND PARTICIPANTS All data were obtained from patients receiving internet-enabled CBT for the treatment of a mental health disorder between June 2012 and March 2018 in England. Cognitive behavioral therapy was delivered in a secure online therapy room via instant synchronous messaging. The initial sample comprised a total of 17 572 patients (90 934 therapy session transcripts). Patients self-referred or were referred by a primary health care worker directly to the service.

EXPOSURES All patients received National Institute for Heath and Care Excellence–approved disorder-specific CBT treatment protocols delivered by a qualified CBT therapist.

MAIN OUTCOMES AND MEASURES Clinical outcomes were measured in terms of reliable improvement in patient symptoms and treatment engagement. Reliable improvement was calculated based on 2 severity measures: Patient Health Questionnaire (PHQ-9) and Generalized Anxiety Disorder 7-item scale (GAD-7), corresponding to depressive and anxiety symptoms respectively, completed by the patient at initial assessment and before every therapy session.

RESULTS Treatment sessions from a total of 14 899 patients (10 882 women) aged between 18 and 94 years (median age, 34.8 years) were included in the final analysis. We trained a deep learning model to automatically categorize therapist utterances into 1 or more of 24 feature categories. The trained model was applied to our data set to obtain quantifiable measures of each feature of treatment delivered. A logistic regression revealed that increased quantities of a number of session features, including change methods (cognitive and behavioral techniques used in CBT), were associated with greater odds of reliable improvement in patient symptoms (odds ratio, 1.11; 95% CI, 1.06-1.17) and patient engagement (odds ratio, 1.20, 95% CI, 1.12-1.27). The quantity of nontherapy-related content was associated with reduced odds of symptom improvement (odds ratio, 0.89; 95% CI, 0.85-0.92) and patient engagement (odds ratio, 0.88, 95% CI, 0.84-0.92).

CONCLUSIONS AND RELEVANCE This work demonstrates an association between clinical outcomes in psychotherapy and the content of therapist utterances. These findings support the principle that CBT change methods help produce improvements in patients' presenting symptoms. The application of deep learning to large clinical data sets can provide valuable insights into psychotherapy, informing the development of new treatments and helping standardize clinical practice.

JAMA Psychiatry. 2020;77(1):35-43. doi:10.1001/jamapsychiatry.2019.2664
Published online August 22, 2019.

Supplemental content

Author Affiliations: Clinical Science Laboratory, Ieso Digital Health, Cambridge, England.

Corresponding Author: Michael P. Ewbank, PhD, Clinical Science Laboratory, Ieso Digital Health, Cowley Road, The Jeffreys Building, Milton, Cambridge CB4 0DS, England (m.ewbank@iesohealth.com).

Compared with treatment of physical conditions, the quality of care of mental health disorders remains poor, and the rate of improvement in treatment is slow.[1] Outcomes for many mental disorders have stagnated or even declined since the original treatments were developed.[2,3] A primary reason for the gap in quality of care is the lack of systematic methods for measuring the delivery of psychotherapy.[1] As with any evidence-based intervention, to be effective, treatment needs to be delivered as intended (also known as treatment integrity),[4,5] which requires accurate measurement of treatment delivered.[6] However, while it is relatively simple to monitor the delivery of most medical treatments (eg, the dosage of a prescribed drug), psychotherapeutic treatments are a series of private discussions between the patient and clinician. As such, monitoring the delivery of this type of treatment to the same extent as physical medicine would require infrastructure and resources beyond the scope of most health care systems.

The National Institute for Heath and Care Excellence and the American Psychological Association recommend cognitive behavioral therapy (CBT) as a treatment for most common mental health problems such as depression and anxiety-related disorders. Cognitive behavioral therapy refers to a class of psychotherapeutic interventions informed by the principle that mental disorders are maintained by cognitive and behavioral phenomena and that modifying these maintaining factors helps produce enduring improvements in patients' presenting symptoms.[7,8] Despite its widespread use, the Improving Access to Psychological Therapies (IAPT) program in England includes no objective measure of treatment integrity for CBT, and it has been proposed that only 3.5% of psychotherapy randomized clinical trials use adequate treatment integrity procedures.[9]

Understanding how CBT works is of particular interest given that the relative effects of different psychotherapeutic interventions appear similar.[10] Thus, whether treatments work through specific factors (eg, CBT change methods) or factors common to most psychotherapies (eg, therapeutic alliance) remains a core issue in the field.[11,12] Studies commonly use observational coding methods (eg, ratings/transcription of recorded therapeutic conversations) to investigate the association between treatment delivered and outcomes.[5] Owing to the resource-intensive nature of this method, studies typically focus on a small number of therapeutic components in a relatively small sample of patients. As with many randomized clinical trials, the results of such interventions are difficult to transfer to real-world psychotherapy[13] and require sample sizes larger than typically used.[14] To determine the most effective components of CBT and whether CBT works via the mechanisms proposed by the approach,[15] quantifiable measures of treatment delivered need to be obtained in a natural clinical context and be gathered from a sufficiently large enough sample to draw meaningful conclusions.

Here, we used a large-scale data set containing session transcripts from more than 14 000 patients receiving internet-enabled CBT (IECBT) (approximately 90 000 hours of therapy). In IECBT, a patient communicates with a qualified CBT therapist using a real-time text-based message system. Internet-enabled CBT has been shown to be clinically effective for the treatment of depression[16] and is currently deployed within IAPT. Using a deep learning approach, we developed a model to automatically categorize therapist utterances according to the role that they play in therapy, generating a quantifiable measure of treatment delivered. We then investigated the association between the quantity of each aspect of therapy delivered and clinical outcomes.

Key Points

Question What aspects of psychotherapy content are significantly associated with clinical outcomes?

Findings In this quality improvement study, a deep learning model was trained to automatically categorize therapist utterances from approximately 90 000 hours of internet-enabled cognitive behavior therapy (CBT). Increased quantities of CBT change methods were positively associated with reliable improvement in patient symptoms, and the quantity of nontherapy-related content showed a negative association.

Meaning The findings support the key principles underlying CBT as a treatment and demonstrate that applying deep learning to large clinical data sets can provide valuable insights into the effectiveness of psychotherapy.

Methods

Design

Data were obtained from patients receiving IECBT for the treatment of a mental health disorder between June 2012 and March 2018. Internet-enabled CBT was delivered using a commercial package currently used in the English National Health Service, provided by Ieso Digital Health (https://www.iesohealth.com/), following internationally recognized standards for information security (ISO 27001; https://www.iesohealth.com/en-gb/legal/iso-certificates). The National Institute for Heath and Care Excellence approved disorder-specific CBT treatment protocols,[17] based on Roth and Pilling CBT competences framework,[18] were delivered in a secure online therapy room via instant synchronous messaging by a British Association for Behavioral and Cognitive Psychotherapies–accredited CBT therapist (see eFigure 1 in the Supplement for a realistic example of a therapy conversation). Patients self-referred or were referred by a primary health care worker directly to the service.

The IAPT program is a large-scale initiative aimed at increasing access to evidence-based psychological therapy for common mental health disorders within the English National Health Service.[19] The information captured through IAPT's minimum data set is intended to support monitoring of implementation and effectiveness of national policy/legislation, performance analysis and benchmarking, and national audit of IAPT services. As determined by the National Health Service, and per The National Institute of Health and Care Excellence principles,[20] clinical audit studies within the IAPT framework do not require additional patient consent or ethical

Box. Feature Categories Used in Transcript Annotation

Therapy Feature Categories

- Hello
- Mood check
- Obtain update
- Bridge
- Risk check
- Set agenda
- Review homework
- Set goals
- Formulation
- Change methods
- Perceptions of change
- Setting homework
- Planning for the future
- Elicit feedback
- Summarize session
- Give feedback
- Arrange next session
- Goodbye
- Socratic questioning[a]
- Therapeutic thanks[a]
- Therapeutic empathy[a]
- Therapeutic praise[a]
- Collaboration[a]
- Other

[a] Features tagged using regular expressions.

approval.[20] When registering to use the Ieso service, patients provide written informed consent as part of a privacy policy agreement, allowing the service to use their anonymized data for audit purposes and to support research, including academic publications.

Clinical Outcomes

Clinical outcomes were defined according to IAPT guidelines[19] and were measured in terms of reliable improvement and IAPT engagement and included as binary measures (ie, 0 or 1). A patient was classed as engaged if they attended 2 or more treatment sessions. Reliable improvement was calculated based on 2 severity measures: Patient Health Questionnaire (PHQ-9)[21] and Generalized Anxiety Disorder 7-item scale (GAD-7),[22] corresponding to depressive and anxiety symptoms respectively, completed by the patient at initial assessment and before every therapy session (see eMethods in the Supplement for details).

Therapy Feature Categories

We defined a total of 24 feature categories (**Box**), informed by the CBT competences framework[18] and the Revised Cognitive Therapy Scale.[23] A research psychologist (M.P.E.) annotated 290 therapy session transcripts, under the guidance of a qualified clinical therapist (S.B.), tagging each therapist text-message utterance as belonging to 1 (or more) of 19 features, with 5 features tagged using regular expressions (see eTable 1 in the Supplement for a full description). A deep learning model (see eMethods in the Supplement) was trained on the annotated utterances and then used to automatically classify all utterances in the full data set into 1 or more of 24 feature categories. Model accuracy is detailed in eTable 2 in the Supplement. To obtain a measure of interrater agreement, a second psychologist (S.B.) annotated a subsample of the transcripts. The interrater reliability was $\kappa = 0.54$ (a value of 0.4-0.6 is considered moderate agreement, with zero equaling chance agreement[24]; eTable 3 in the Supplement).

Statistical Analysis

Using the output of the model, the mean number of words for each feature, averaged across all sessions, was calculated for each case. The final treatment session was excluded because outcome measures are taken prior to the commencement of each treatment session. The initial sample comprised a total of 90 934 session transcripts taken from 17 572 patients, with a reliable improvement rate of 63.4% and IAPT engagement rate of 87.3%.

All analyses were performed in R (the R Foundation). Cases with missing start or end PHQ-9 or GAD-7 scores (n = 1338) were excluded from the analysis. We performed 3 multivariable logistic regression analyses. First, a multivariable logistic regression was performed to investigate the association between session features and reliable improvement. Predictor variables were the mean number of words for each feature across sessions plus patient demographics: starting PHQ-9 and GAD-7 scores, sex (male, female, or unstated/unknown), age, whether the patient had a long-term physical condition (yes, no, or unstated/unknown), and whether the patient was taking psychotropic medication at the start of treatment (prescribed not taking, prescribed taking, not prescribed, or unstated/unknown). The number of sessions completed and the mean duration of sessions were also included. Cases with a mean of fewer than 50 patient words were excluded (n = 16), leaving a total of 13 073 patients (at a clinical caseness threshold and engaged in treatment) in the analysis.

We also investigated the association between first-session features and IAPT engagement. Predictor variables were the number of each therapy feature in the first session, patient demographics, and duration of first session. Sessions with a total of fewer than 50 patient words were excluded (n = 121) making a total of 14 899 patients, at caseness.

Details of a logistic regression analysis investigating the association between first-session features and outcomes can be found in eResults and eTable 5 in the Supplement. Details of diagnoses for patients included in the analysis can be found in eTable 4 in the Supplement. Patient demographic information is shown in **Tables 1** and **2** and eTable 5 in the Supplement.

For all analyses, continuous predictor variables were scaled and centered to the mean. Statistical significance was defined as *P* less than .05 two-tailed, uncorrected. Multicollinearity analyses revealed that variance inflation factors were smaller than 2 for all predictor variables, confirming that regression models were not affected by the presence of multicollinearity.

Table 1. Factors Associated With Reliable Improvement—All Sessions[a]

Feature	No. of Words, Mean (SD)	Sessions, %	Odds Ratio (95% CI)	*z* Value	*P* Value
Hello	12 (22.7)	99.6	0.92 (0.88-0.96)	−3.57	<.001
Mood check	5.6 (7)	97.9	0.99 (0.95-1.03)	−0.34	.73
Obtain update	16.4 (14.5)	59.0	1.03 (0.99-1.08)	1.56	.12
Bridge	12.2 (17.9)	27.9	0.95 (0.91-0.98)	−2.76	.006
Risk check	13.6 (31.5)	21.0	0.85 (0.81-0.89)	−7.54	<.001
Set agenda	47.2 (43.5)	71.3	1.08 (1.02-1.14)	3.02	.002
Review homework	18.5 (19.2)	44.5	1.04 (1.00-1.09)	2.00	.04
Set goals	15.9 (30.8)	19.4	1.00 (0.96-1.05)	0.40	.69
Formulation	30.3 (63.9)	18.2	0.96 (0.92-1.00)	−1.89	.06
Give feedback	33.6 (40)	52.1	1.05 (1.00-1.10)	2.20	.02
Change methods	477.1 (236)	97.9	1.11 (1.06-1.17)	4.37	<.001
Perceptions of change	1.6 (4.8)	5.8	1.11 (1.06-1.16)	4.59	<.001
Set homework	63.2 (48.9)	69.1	0.96 (0.92-1.00)	−1.68	.09
Planning for future	1.1 (6)	2.4	1.12 (1.06-1.19)	4.01	<.001
Elicit feedback	15.3 (16.4)	55.3	1.06 (1.02-1.11)	2.82	.004
Summarize session	0.25 (2.6)	0.4	0.99 (0.95-1.03)	−0.52	.60
Arrange next session	30 (21.3)	82.5	1.00 (0.96-1.04)	0.05	.96
Goodbye	15.4 (10.4)	90.7	0.95 (0.91-0.99)	−2.34	.02
Socratic questioning	24.1 (31.1)	47.4	1.02 (0.98-1.06)	0.95	.34
Therapeutic thanks	5.4 (13.3)	13.3	0.97 (0.93-1.01)	−1.48	.14
Therapeutic empathy	21 (31.3)	38.0	0.84 (0.81-0.88)	−8.21	<.001
Therapeutic praise	30.6 (39.4)	52.6	1.21 (1.15-1.27)	7.18	<.001
Collaboration	41 (45.9)	61.9	0.97 (0.93-1.02)	−1.09	.27
Other	121.1 (81)	96.0	0.88 (0.85-0.92)	−5.82	<.001
Variable, mean/prevalence (SD)					
Total sessions, No.	6.2 (2.9)	NA	1.22 (1.17-1.27)	9.01	<.001
Session duration, min	62.4 (7.5)	NA	0.95 (0.91-0.99)	−2.34	.02
Start PHQ-9	14.7 (5.4)	NA	0.95 (0.91-0.99)	−2.41	.03
Start GAD-7	8.3 (5.7)	NA	1.29 (1.23-1.34)	11.8	<.001
Patient age, y	34.8 (12.0)	NA	1.16 (1.12-1.22)	7.47	<.001
Patient sex, No. (%)					
Male	3493 (26.7)	NA	0.96 (0.88-1.05)	−0.89	.50
Female	9537 (72.9)	NA			
Unknown/not stated	43 (0.4)	NA	0.92 (0.49-1.78)	−0.24	.74
Long-term condition, No. (%)					
No	6056 (46.4)	NA			
Yes	3632 (27.8)	NA	0.72 (0.66-0.80)	−6.55	<.001
Unknown/not stated	3383 (25.8)	NA	0.78 (0.71-0.86)	−5.08	<.001
Psychotropic medication, No. (%)					
Prescribed not taking	1116 (8.6)	NA			
Not prescribed	5971 (45.7)	NA	1.23 (1.06-1.41)	2.84	.004
Prescribed taking	5535 (42.3)	NA	0.98 (0.84-1.13)	−0.27	.78
Unknown/not stated	451 (3.4)	NA	0.85 (0.67-1.08)	−1.28	.20

Abbreviations: GAD-7, Generalized Anxiety Disorder 7-item scale; NA, not applicable; PHQ-9, Patient Health Questionnaire.

[a] Output of logistic regression investigating association between reliable improvement and mean number of words per feature across treatment. Standardized odds ratios indicate the association of an increase of 1 SD of a feature with the odds of improvement. Percentage of sessions indicates the percentage of the total number of sessions that contained utterances categorized as that feature. Female sex, no long-term conditions, and prescribed not taking psychotropic medication were reference classes for the categorical variables.

Results

Factors Associated With Reliable Improvement Across Treatment

Figure 1 shows the standardized odds ratios (ORs) for each therapy feature included in the multivariable logistic regression (Table 1). The results revealed increased quantities of "therapeutic praise" (OR, 1.21; 95% CI, 1.15-1.27), "planning for the future" (OR, 1.12; 95% CI, 1.06-1.19), "perceptions of change" (OR, 1.11; 95% CI, 1.06-1.16), "change methods" (OR, 1.11; 95% CI, 1.06-1.17), "set agenda" (OR, 1.08; 95% CI, 1.02-1.14), "elicit feedback" (OR, 1.06; 95% CI, 1.02-1.11), "give feedback" (OR, 1.05; 95% CI, 1.00-1.10), and "review homework" (OR, 1.04; 95% CI, 1.00-1.09) were all associated with greater odds of reliable improvement. By contrast, increases in nontherapy-related content ("other" [OR, 0.89; 95% CI, 0.85-0.92], "hello" [OR, 0.92; 95% CI, 0.88-0.96], and "goodbye" [OR, 0.95; 95% CI, 0.91-

Table 2. First-Session Factors Associated With IAPT Engagement[a]

Feature	No. of Words, Mean (SD)	Sessions, %	Odds Ratio (95% CI)	z Value	*P* Value
Hello	14.4 (34.7)	99.7	0.93 (0.88-0.99)	−2.45	.01
Mood check	5.6 (10.5)	48.1	0.98 (0.93-1.03)	−0.96	.33
Obtain update	12.3 (19.6)	46.1	0.96 (0.92-1.01)	−1.54	.11
Bridge	9.7 (24.8)	22.7	0.94 (0.90-0.98)	−2.63	.008
Risk check	22.8 (54.7)	30.4	0.98 (0.94-1.03)	−0.69	.48
Set agenda	61.3 (68.7)	74.9	0.99 (0.94-1.05)	−0.27	.79
Review homework	15.2 (27.3)	39.4	0.96 (0.91-1.01)	−1.47	.14
Set goals	28.3 (57.9)	35.9	1.03 (0.98-1.09)	1.07	.28
Formulation	53.2 (126)	30.4	1.10 (1.04-1.17)	3.33	<.001
Give feedback	17.4 (57.2)	49.3	1.00 (0.95-1.07)	0.31	.75
Change methods	426.5 (279.5)	97.6	1.20 (1.12-1.27)	5.56	<.001
Perceptions of change	1.13 (7.4)	3.6	0.97 (0.93-1.01)	−1.42	.14
Set homework	75.8 (74.4)	78.4	1.09 (1.03-1.16)	2.97	<.002
Planning for future	0.56 (8.5)	1.0	0.93 (0.89-0.96)	−3.77	<.001
Elicit feedback	17.4 (25)	60.9	1.09 (1.03-1.16)	2.97	.002
Summarize session	0.24 (4.67)	0.3	1.00 (0.94-1.09)	0.01	.98
Arrange next session	33.1 (32.6)	84.0	1.17 (1.10-1.24)	5.30	<.001
Goodbye	16.2 (15.6)	90.9	1.02 (0.97-1.08)	0.83	.40
Socratic questioning	20 (39.5)	40.8	0.94 (0.89-0.99)	−2.28	.02
Therapeutic thanks	8.5 (24.3)	19.4	1.13 (1.06-1.20)	3.73	<.001
Therapeutic empathy	25.5 (51.1)	44.0	0.93 (0.88-0.97)	−3.20	.001
Therapeutic praise	23.3 (47)	41.8	1.05 (0.98-1.11)	1.47	.15
Collaboration	45.2 (72.8)	60.4	1.01 (0.94-1.07)	0.26	.79
Other	141.1 (117.4)	96.9	0.88 (0.84-0.92)	−5.12	<.001
Variable, mean/prevalence (SD)					
Session duration, min	63.1 (9.9)	NA	1.26 (1.20-1.33)	8.89	<.001
Start PHQ-9	14.9 (5.5)	NA	0.87 (0.82-0.92)	−4.81	<.001
Start GAD-7	8.8 (5.9)	NA	1.00 (0.95-1.06)	−0.01	.99
Patient age	34.8 (12.0)	NA	1.07 (1.02-1.13)	2.64	.008
Patient sex, %					
Male	3967 (26.7)	NA	1.02 (0.91-1.01)	0.28	.78
Female	10 882 (73.0)	NA	NA	NA	NA
Unknown/not stated	50 (0.3)	NA	0.95 (0.45-2.34)	−0.11	.91
Long-term condition, %					
No	6860 (46.0)	NA			
Yes	4129 (27.7)	NA	1.02 (0.90-1.15)	0.24	.81
Unknown/not stated	3910 (26.3)	NA	0.90 (0.80-1.02)	−1.68	.09
Psychotropic medication, %					
Prescribed not taking	1304 (8.8)				
Not prescribed	6755 (45.3)	NA	1.21 (1.02-1.44)	2.19	.03
Prescribed taking	6320 (42.4)	NA	1.20 (1.01-1.47)	2.06	.04
Unknown/not stated	520 (3.5)	NA	1.10 (1.01-1.43)	0.64	.52

Abbreviations: GAD-7, Generalized Anxiety Disorder 7-item scale; IAPT, Improving Access to Psychological Therapies; NA, not applicable; PHQ-9, Patient Health Questionnaire.

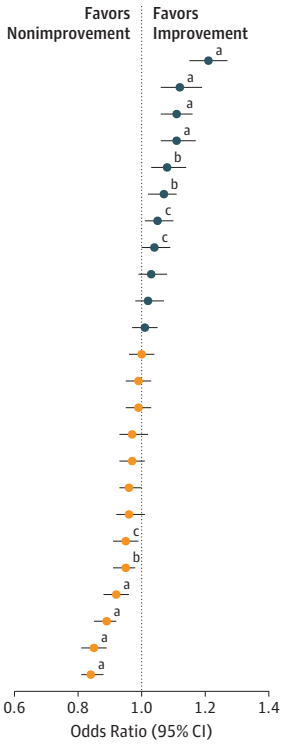
[a] Output of logistic regression investigating association between patient engagement and number of words per feature in the first treatment session. Standardized odds ratios indicate the effect of an increase of 1 SD of a feature on the odds of engagement. Percentage of sessions indicates the percentage of the total number of first treatment sessions that contained utterances categorized as that feature. Female sex, no long-term conditions, and prescribed not taking psychotropic medication were reference classes for the categorical variables.

0.99]), along with "therapeutic empathy" (OR, 0.84; 95% CI, 0.81-0.88), "risk check" (OR, 0.85; 95% CI, 0.81-0.89), and "bridge" (OR, 0.95; 95% CI, 0.91-0.98) were negatively associated with improvement.

Patient variables of starting GAD-7 score (OR, 1.29; 95% CI, 1.23-1.34), not being prescribed medication (OR, 1.23; 95% CI, 1.06-1.41), patient age (OR, 1.16; 95% CI, 1.12-1.22), and total number of treatment sessions (OR, 1.22; 95% CI, 1.17-1.27) were also associated with increased odds of improvement. Starting PHQ-9 score (OR, 0.95; 95% CI, 0.91-0.99), the presence of a long-term medical condition (OR, 0.72; 95% CI, 0.66-0.88), and longer session durations (OR,

Figure 1. Factors Associated With Reliable Improvement—All Sessions

Reliable Improvement, All Treatment Sessions	OR (95% CI)
Therapeutic praise	1.21 (1.15-1.27)
Planning for future	1.12 (1.06-1.19)
Perceptions of change	1.11 (1.06-1.16)
Change methods	1.11 (1.06-1.17)
Set agenda	1.08 (1.03-1.14)
Elicit feedback	1.07 (1.02-1.11)
Give feedback	1.05 (1.01-1.10)
Review homework	1.04 (1.00-1.09)
Obtain update	1.03 (0.99-1.08)
Socratic questioning	1.02 (0.98-1.07)
Set goals	1.01 (0.97-1.05)
Arrange next session	1.00 (0.96-1.04)
Mood check	0.99 (0.95-1.03)
Summarize session	0.99 (0.95-1.03)
Collaboration	0.97 (0.93-1.02)
Therapeutic thanks	0.97 (0.93-1.01)
Formulation	0.96 (0.93-1.00)
Set homework	0.96 (0.92-1.01)
Goodbye	0.95 (0.91-0.99)
Bridge	0.95 (0.91-0.98)
Hello	0.92 (0.88-0.96)
Other	0.89 (0.85-0.92)
Risk check	0.85 (0.81-0.89)
Therapeutic empathy	0.84 (0.81-0.88)



Forest plot of logistic regression model investigating association between mean number of words per feature across treatment and reliable improvement. Standardized odds ratios and 95% confidence intervals are shown (and listed in the right column). Adjusted for total number of sessions, symptom severity, patient sex, age, medication status, presence of long-term condition, and session duration.

[a] $P < .001$.

[b] $P < .01$.

[c] $P < .05$.

0.95; 95% CI, 0.91-0.99) were associated with reduced odds of improvement.

Factors Associated With IAPT Engagement in First Treatment Session

Figure 2 shows the standardized ORs for each session feature included in the multivariable logistic regression (Table 2). We found that "change methods" (OR, 1.20; 95% CI, 1.12-1.27), "elicit feedback" (OR, 1.09; 95% CI, 1.03-1.16), "set homework" (OR, 1.09; 95% CI, 1.03-1.16)," arrange next session" (OR, 1.17; 95% CI, 1.10-1.24), "therapeutic thanks" (OR, 1.13; 95% CI, 1.06-1.20), and "formulation" (OR, 1.10; 95% CI, 1.04-1.17) were associated with increased odds of IAPT engagement. By contrast, nontherapy-related content ("other" and "hello") showed a negative association ("other" OR, 0.88; 95% CI, 0.84-0.92; "hello" OR, 0.93; 95% CI, 0.88-0.99), as did "therapeutic empathy" (OR, 0.93; 95% CI, 0.88-0.97), "Socratic questioning" (OR, 0.94; 95% CI, 0.89-0.99), "bridge" (OR, 0.94; 95% CI, 0.90-0.98), and "planning for the future" (OR, 0.93; 95% CI, 0.89-0.96). Patient age (OR, 1.07; CI, 1.02-1.13), not being prescribed medication (OR, 1.21; 95% CI, 1.02-1.44), being prescribed and taking medication (OR, 1.20; 95% CI, 1.01-1.47), and duration of the first session (OR, 1.26; CI, 1.20-1.33) were positively associated with IAPT engagement, while starting PHQ-9 score (OR, 0.87; CI, 0.82-0.92) was negatively associated.

Discussion

Improving the quality and efficacy of psychotherapy requires that treatment be delivered as intended; however, monitoring and measuring treatment delivered presents a substantial challenge. We developed a method of objectively quantifying psychotherapy using a deep learning approach to automatically categorize therapist utterances from approximately 90 000 hours of IECBT. We find that factors specific to CBT, as well as factors common to most psychotherapies, are associated with increased odds of reliable improvement in patient symptoms.

The results revealed a positive association between the quantity of CBT change method-related content and both reliable improvement and IAPT engagement. This finding supports the key principles underlying CBT and provides validation for CBT as a treatment (ie, modifying cognitive and behavioral factors produces improvements in patient symptoms). Here, the category of "change methods" included any example of cognitive or behavioral reattribution, skill-teaching, conceptualization, or psychoeducation. Thus, further research is needed to determine the association between different types of change method and outcomes.[15]

Homework in CBT is used to help patients practice skills learned in therapy and generalize these skills to the real world.[25] Increased content related to reviewing homework was positively associated with symptom improvement, while setting homework in the first session was associated with increased engagement. It is unclear whether an increase in reviewing homework plays a causal role in symptom change or whether it reflects a patient who has completed homework; however, these findings accord with evidence that out-of-session homework is important in determining outcomes in CBT.[26] The results show that agenda setting is also positively associated with reliable improvement. Agenda setting involves the therapist and patient deciding on the topics to be discussed during the session. However, we are unable to determine whether the agenda was adhered to in the session. The results also support the principle that giving and eliciting feedback helps both the therapist and patient develop a greater understanding of key issues and possibly strengthens the therapeutic alliance.[27]

Session content related to planning for the future after therapy and discussing perceptions of change was also positively associated with improvement. A discussion of perceptions of change is only likely to occur following some degree of change; similarly, planning for a future most likely occurs when patients are close to completing treatment and/or have moved toward improvement. As such, the increased occurrence of both features is likely to be reflective of treatment progressing well. Consistent with this, neither feature was sig-

nificantly associated with outcomes in the first treatment session (eTable 5 in the Supplement). By contrast, goal setting in the first session was positively associated with improvement, supporting the goal-directed nature of CBT.[27] Content associated with formulation (ie, the beliefs and behavioral strategies that characterize a disorder)[28] in the first session also showed a positive association with IAPT engagement (and a borderline significant association with improvement), suggesting that placing patients' experiences within a cognitive behavioral framework early in therapy is beneficial.

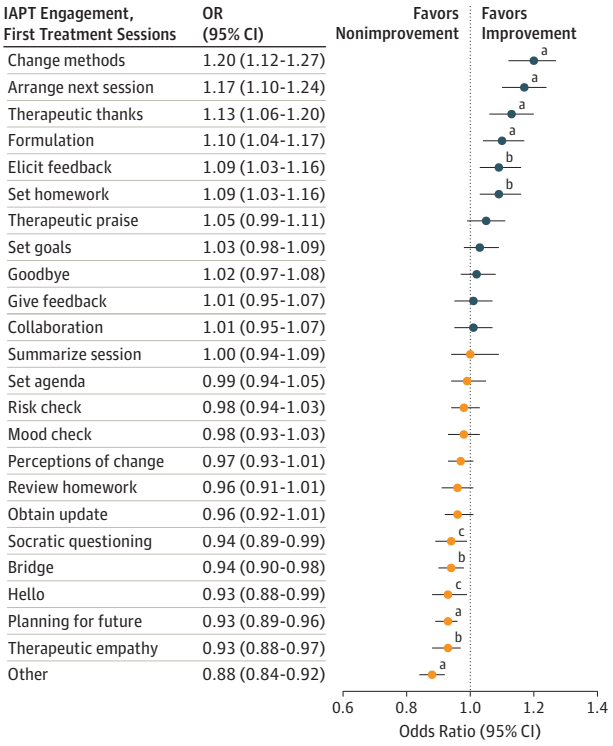
Several features were found to be negatively associated with outcomes, in particular nontherapy-related content. Content that did not fall within any of the other 23 categories ("other") includes utterances related to technical/practical matters or nontherapeutic advice/conversations. While greetings and goodbyes are essential to the structure of a therapy session, our results indicate that, when aggregated across sessions, an excessive or disproportionate amount of time spent on such nontherapeutic aspects may reduce the quantity of active intervention. Importantly, this suggests that rather than the quantity of conversation, it is the therapeutic nature of conversation and/or the dosage of therapy delivered in a session that is associated with improvement in patient symptoms.

Risk checking also showed a strong negative association with reliable improvement. We believe this is likely to be reflective of patients with more complex problems who report more thoughts of self-harm. The quantity of risk checks will increase if a patient confirms that they feel at risk; thus, it is important to recognize that increased risk-checking content is essential and unavoidable. An extended period focused on risk is also likely to cause a deviation in the structure of the session and a subsequent reduction in the dosage of active therapy delivered.

A central issue in psychotherapy research is whether different approaches work through specific factors or factors that are common to most psychotherapies. Here, we find a positive association between improvement and/or IAPT engagement for each of 6 techniques identified as distinguishing CBT from psychodynamic therapy.[11] Common factors, such as therapeutic alliance, are thought to play a role in all psychotherapeutic treatments[29] and show a moderate association with outcomes.[30] Here, we found that "therapeutic praise" was positively associated with improvement, whereas "therapeutic empathy" showed a negative association. Rather than playing a causal role in outcomes, we believe increased empathy is likely to be indicative of a patient reporting a greater number of problems. Similarly, increased praise may be reflective of a patient responding well to treatment. Further research is required to determine the causal association between therapeutic alliance and outcomes, although previous work indicates therapeutic alliance may be reflective of a change in symptoms.[31]

We also investigated the association between patient variables and outcomes. Patient age (older patients showing better outcomes), absence of a long-term medical condition, not being prescribed psychotropic medication, and severity of anxiety symptoms were all positively associated with reliable improvement. By contrast, severity of depressive symptoms, the presence of a long-term medical condition, and being prescribed psychotropic medication were negatively associated. These results accord with previous work investigating treatment outcomes in a sample of approximately 3000 patients receiving IECBT.[32] Both studies report a positive association between GAD-7 scores and reliable improvement. Further work is needed to determine whether this reflects a greater association of CBT with short-term symptoms of anxiety and/or whether this effect may be specific to IECBT.

Figure 2. First-Session Factors Associated With IAPT Engagement



Forest plot of logistic regression model investigating association between mean number of words per feature in the first treatment session and patient engagement. Standardized odds ratios and 95% confidence intervals are shown (and listed in the right column). Adjusted for symptom severity, patient sex, age, medication status, presence of long-term condition, and session duration.

[a] $P < .001$.

[b] $P < .01$.

[c] $P < .05$.

Limitations

A limitation of our approach is that it is not possible to determine whether a therapeutic feature is applied in an appropriate manner or whether a therapist adheres to the CBT protocol. It should be noted that the model provides a measure of the association between features and outcomes across sessions rather than measuring the quality of an individual session. Thus, future work needs to build on this approach to generate a validated model of session quality/adherence, alongside further refinement of the annotation guidelines and pooling of annotations. In addition, the model does not assess how the treatment was received by patients. To partly address this, we

are currently developing procedures to quantify patient utterances, enabling us to determine, for example, how use of change methods are associated with a change in patient's cognitions and whether therapeutic empathy is positively associated with outcomes after adjusting for the number of problems expressed by the patient.

We emphasize that our results only reveal the presence of an association between therapy content and outcomes, although some aspects of therapy (eg, change methods) are typically initiated by the therapist and appear likely to play a causal role. Further work is needed to determine the casual relationship between therapy features and outcomes by focusing on the temporal association between content and symptom change. Given the limited outcomes measures available, we are also unable to address the association between therapy content and long-term improvements in symptoms. In addition, other patient factors not included are likely to play a role in determining outcomes. Finally, it should be noted that for large data sets, the ORs and confidence intervals should be considered more informative of the clinical importance of a feature than statistical significance alone.

Conclusions

At present, the detailed monitoring of therapist performance requires expensive and time-consuming procedures. We believe that this work represents a first step toward a practicable approach for quality controlled behavioral health care. Such monitoring could help arrest therapist drift, ie, the failure to deliver treatments a therapist has been trained to deliver, which may be one of the biggest factors contributing to poor delivery of treatment.[33] Monitoring may help reverse the lower improvement rates observed in more experienced therapists.[34] We note that while a typical IAPT therapist may accrue substantial experience throughout a career (approximately 30 000 therapy hours), this data set represents an accumulation of knowledge from more than 90 000 hours of CBT. Deep leaning allows us to extract this knowledge to provide valuable insights into therapy that were previously unavailable to an individual therapist. As such, we believe this approach represents an important step in developing a data-driven understanding of mental health treatment and in improving the efficacy of psychotherapy.

ARTICLE INFORMATION

Accepted for Publication: July 18, 2019.

Published Online: August 22, 2019. doi:10.1001/jamapsychiatry.2019.2664

Author Contributions: Dr Ewbank had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Mr Martin created the utterance annotation software. Dr Cummins led the deep learning modeling.
Concept and design: All authors.
Acquisition, analysis, or interpretation of data: Ewbank, Cummins, Catarino, Martin, Blackwell.
Drafting of the manuscript: Ewbank.
Critical revision of the manuscript for important intellectual content: Cummins, Tablan, Bateup, Catarino, Martin, Blackwell.
Statistical analysis: Ewbank, Cummins.
Obtained funding: Tablan, Blackwell.
Administrative, technical, or material support: Cummins, Tablan, Bateup, Catarino, Martin, Blackwell.
Supervision: Tablan, Catarino, Blackwell.

Conflict of Interest Disclosures: All authors are employees of Ieso Digital Health. All authors report a patent to Methods and Systems for Improved Therapy Delivery and Monitoring pending.

Funding/Support: This study was funded by Ieso Digital Health.

Role of the Funder/Sponsor: As employees of Ieso Digital Health, the authors were responsible for the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Meeting Presentation: This paper was presented at the World Congress of Psychiatry; August 22, 2019; Lisbon, Portugal.

REFERENCES

1. Kilbourne AM, Beck K, Spaeth-Rublee B, et al. Measuring and improving the quality of mental health care: a global perspective. *World Psychiatry*. 2018;17(1):30-38. doi:10.1002/wps.20482

2. Johnsen TJ, Friborg O. The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: a meta-analysis. *Psychol Bull*. 2015;141(4):747-768. doi:10.1037/bul0000015

3. Holmes EA, Ghaderi A, Harmer CJ, et al. The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *Lancet Psychiatry*. 2018;5(3):237-286. doi:10.1016/S2215-0366(17)30513-8

4. Waller G, Turner H. Therapist drift redux: why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behav Res Ther*. 2016;77:129-137. doi:10.1016/j.brat.2015.12.005

5. Perepletchikova F. On the topic of treatment integrity. *Clin Psychol (New York)*. 2011;18(2):148-153. doi:10.1111/j.1468-2850.2011.01246.x

6. Essock SM, Covell NH, Weissman EM. Inside the black box: the importance of monitoring treatment implementation. *Schizophr Bull*. 2004;30(3):613-615. doi:10.1093/oxfordjournals.schbul.a007107

7. Beck AT. Cognitive therapy: nature and relation to behavior therapy. *Behav Ther*. 1970;1:184-200. doi:10.1016/S0005-7894(70)80030-2

8. Ellis A. *Reason and Emotion in Psychotherapy*. New York: Lyle Stuart; 1962.

9. Perepletchikova F, Treat TA, Kazdin AE. Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *J Consult Clin Psychol*. 2007;75(6):829-841. doi:10.1037/0022-006X.75.6.829

10. Barth J, Munder T, Gerger H, et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS Med*. 2013;10(5):e1001454. doi:10.1371/journal.pmed.1001454

11. Blagys MD, Hilsenroth MJ. Distinctive activities of cognitive-behavioral therapy: a review of the comparative psychotherapy process literature. *Clin Psychol Rev*. 2002;22(5):671-706. doi:10.1016/S0272-7358(01)00117-9

12. Mulder R, Murray G, Rucklidge J. Common versus specific factors in psychotherapy: opening the black box. *Lancet Psychiatry*. 2017;4(12):953-962. doi:10.1016/S2215-0366(17)30100-1

13. Doss BD, Atkins DC. Investigating treatment mediators when simple random assignment to a control group is not possible. *Clin Psychol Sci Pract*. 2006;13:321-336. doi:10.1111/j.1468-2850.2006.00045.x

14. Bell EC, Marcus DK, Goodlad JK. Are the parts as good as the whole? a meta-analysis of component treatment studies. *J Consult Clin Psychol*. 2013;81(4):722-736. doi:10.1037/a0033004

15. Lorenzo-Luaces L, German RE, DeRubeis RJ. It's complicated: the relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clin Psychol Rev*. 2015;41:3-15. doi:10.1016/j.cpr.2014.12.003

16. Kessler D, Lewis G, Kaur S, et al. Therapist-delivered internet psychotherapy for depression in primary care: a randomised controlled trial. *Lancet*. 2009;374(9690):628-634. doi:10.1016/S0140-6736(09)61257-5

17. National Institute for Health and Clinical Excellence. Common mental health disorders: the NICE guideline on identification and pathways to care. London, England: RCPsych Publication; 2011.

18. Roth AD, Pilling S. Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders.. *Behav Cogn Psychother*. 2008;36(February):129-147. doi:10.1017/S1352465808004141

19. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *Int Rev Psychiatry*.

2011;23(4):318-327. doi:10.3109/09540261.2011.606803

20. N.I.C.E. Principles for Best Practice in Clinical Audit. Radcliffe Medical Press; 2002. http://www.uhbristol.nhs.uk/files/nhs-ubht/best_practice_clinical_audit.pdf. Accessed January 31, 2019.

21. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606-613. doi:10.1046/j.1525-1497.2001.016009606.x

22. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. 2006;166(10):1092-1097. doi:10.1001/archinte.166.10.1092

23. Blackburn I, James IA, Milne DL, et al. The revised cognitive therapy scale (CTS-R): psychometric properties. *Behav Cogn Psychother*. 2001;29:431-446. doi:10.1017/S1352465801004040

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174. doi:10.2307/2529310

25. Alford BA, Beck AT. Cognitive therapy of delusional beliefs. *Behav Res Ther*. 1994;32(3):369-380. doi:10.1016/0005-7967(94)90134-1

26. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton PJ, Hofmann SG. Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. *Behav Ther*. 2016;47(5):755-772. doi:10.1016/j.beth.2016.05.002

27. Beck JS. *Cognitive Behavior Therapy: Basics and Beyond*. 2nd ed. New York, NY: Guilford Press; 2011.

28. Alford BA, Beck AT. *The Integrative Power of Cognitive Therapy*. New York, NY: Guilford Press; 1997. doi:10.1891/0889-8391.11.4.309

29. Wampold BE. How important are the common factors in psychotherapy? an update. *World Psychiatry*. 2015;14(3):270-277. doi:10.1002/wps.20238

30. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *J Consult Clin Psychol*. 2000;68(3):438-450. doi:10.1037/0022-006X.68.3.438

31. Strunk DR, Brotman MA, DeRubeis RJ. The process of change in cognitive therapy for depression: predictors of early inter-session symptom gains. *Behav Res Ther*. 2010;48(7):599-606. doi:10.1016/j.brat.2010.03.011

32. Catarino A, Bateup S, Tablan V, et al. Demographic and clinical predictors of response to internet-enabled cognitive-behavioural therapy for depression and anxiety. *BJPsych Open*. 2018;4(5):411-418. doi:10.1192/bjo.2018.57

33. Waller G. Evidence-based treatment and therapist drift. *Behav Res Ther*. 2009;47(2):119-127. doi:10.1016/j.brat.2008.10.018

34. Shapiro DA, Shapiro D. Meta-analysis of comparative therapy outcome studies: a replication and refinement. *Psychol Bull*. 1982;92(3):581-604. doi:10.1037/0033-2909.92.3.581