Routledge
Taylor & Francis Group

Psychotherapy Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tpsr20

Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts

M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz & A. D. Blackwell

To cite this article: M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz & A. D. Blackwell (2021) Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts, Psychotherapy Research, 31:3, 300-312, DOI: 10.1080/10503307.2020.1788740

To link to this article: https://doi.org/10.1080/10503307.2020.1788740

View supplementary material

Published online: 03 Jul 2020.

Submit your article to this journal

Article views: 1872

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Psychotherapy Research, 2021
Vol. 31, No. 3, 300–312, https://doi.org/10.1080/10503307.2020.1788740

Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts

M. P. EWBANK, R. CUMMINS, V. TABLAN, A. CATARINO, S. BUCHHOLZ, & A. D. BLACKWELL

Clinical Science Laboratory at Ieso, Ieso Digital Health, Cambridge, United Kingdom

(Received 5 December 2019; revised 22 June 2020; accepted 22 June 2020)

Abstract

Objective: Understanding patient responses to psychotherapy is important in developing effective interventions. However, coding patient language is a resource-intensive exercise and difficult to perform at scale. Our aim was to develop a deep learning model to automatically identify patient utterances during text-based internet-enabled Cognitive Behavioural Therapy and to determine the association between utterances and clinical outcomes.

Method: Using 340 manually annotated transcripts we trained a deep learning model to categorize patient utterances into one or more of five categories. The model was used to automatically code patient utterances from our entire data set of transcripts (~34,000 patients), and logistic regression analyses used to determine the association between both reliable improvement and engagement, and patient responses.

Results: Our model reached human-level agreement on three of the five patient categories. Regression analyses revealed that increased counter change-talk (movement away from change) was associated with lower odds of both reliable improvement and engagement, while increased change-talk (movement towards change or self-exploration) was associated with increased odds of improvement and engagement.

Conclusions: Deep learning provides an effective means of automatically coding patient utterances at scale. This approach enables the development of a data-driven understanding of the relationship between therapist and patient during therapy.

Keywords: cognitive behaviour therapy; outcome research; technology in psychotherapy research & training

Clinical or methodological significance of this article: Patient language features have been shown to be associated with clinical outcomes in response to a course of psychotherapy. We believe the application of deep learning enables the accurate measurement of patient utterances during text-based internet-enabled CBT at a scale beyond the scope of previous research. This approach can help provide a data-driven understanding of the relationship between patient language and outcomes in large "real-world" clinical data sets, helping to further our knowledge of how psychotherapy works and improving the effectiveness of talking therapies such as CBT.

Introduction

Cognitive behaviour therapy (CBT) is effective across a wide range of mental disorders, however, response rates vary (Hofmann et al., 2012) and outcomes for many disorders have stagnated or even declined over time (Johnsen & Friborg, 2015). Within the Improving Access to Psychological Therapies (IAPT) programme in the English National Health Service, a rare example of consistent outcome monitoring in routine mental health services, only about one in two people who have a course of treatment will reach recovery (NHS, 2018).

Correspondence concerning this article should be addressed to M. P. Ewbank, Clinical Science Laboratory at Ieso, Ieso Digital Health, The Jeffreys Building, Cowley Road, Milton, Cambridge CB4 0DS, United Kingdom. Email: m.ewbank@iesohealth.com

Identifying possible mediators (e.g., compliance with treatment protocols) and moderators (e.g., patient characteristics) of treatment outcomes is vital to enhancing the effectiveness of treatment and understanding which treatments work and for whom (Kraemer et al., 2002). As CBT requires the patient to be an active participant in a collaborative process (e.g., patients are required to learn new skills and to complete homework assignments that involve applying these skills to everyday life), a patient's willingness and desire to actively engage in treatment is likely to be an important factor in determining outcomes (Drieschner et al., 2004).

Identifying and categorizing patient statements during CBT is an important step towards developing an understanding of when and why patients are unlikely to follow a treatment plan and could help inform the development of more personal and engaging treatments. A number of studies have examined the association between treatment outcomes and measures of patients' desire/willingness to change in the context of motivational interviewing (MI) for addictions (Magill et al., 2018). Most of this research has used the MISC 1.1 coding system (Glynn & Moyers, 2009) to quantify the frequency or averaged strength of patient statements. Within this coding system, any patient language that moves in the direction of change is termed "change talk" and any language indicating a movement away from change is termed "counter-change talk." Research investigating client language in MI has consistently found that lower levels of counter change-talk during early sessions are associated with more positive outcomes, however an association with change talk has not consistently been reported across studies (Magill et al., 2018).

Observational coding methods have also been used to measure patient language in CBT. Westra (2011) coded patient resistance (i.e., behaviour which opposes, blocks, diverts, or impedes the direction set by the therapist) during the first treatment session of CBT for generalized anxiety disorder and found that higher resistance was a predictor of lower homework compliance and greater anxiety symptoms one-year post-treatment. Research adapting the MISC 1.1 for use in CBT has shown that a greater number of counter-change talk statements in the first session and early resistance to therapist direction are both related to poorer outcomes (Button et al., 2015; Lombardi et al., 2014). In addition, increased counter-change talk at the outset of therapy has been shown to be associated with increased rates of alliance rupture later in therapy (Hunter et al., 2014). More recent evidence indicates that observational coding measures of patient language may be a better predictor of outcomes in CBT than self-reported motivation (Poulin et al., 2019).

Observational coding typically involves manual transcription and segmenting of audio and/or video recordings of individual counselling/CBT sessions. The time-consuming and labour-intensive nature of this exercise makes it difficult to scale up such a process, meaning that studies of patient language within CBT have been limited to relatively small samples of patients taking part in randomized controlled trials. As a result, such studies have low statistical power (i.e., a reduced chance of detecting a true effect and lower likelihood that a significant finding reflects a true effect) and the results are difficult to generalize to real-world clinical practice (Bell et al., 2013; Doss & Atkins, 2006). Indeed, compared to pharmacological studies, most studies of psychotherapy are vastly underpowered given typical effect sizes (Bell et al., 2013; Cuijpers, 2016).

One way to overcome the limitations inherent to manual transcription is to use an automated approach to coding patient language. These approaches segment dialogues into therapist and patient utterances (units of speech that typically perform one function). In one of the first approaches to this task, Can et al. (2015) applied a conditional random field approach to the task of predicting utterance level MISC codes during MI. Although they allow their model to have access to some preceding context, it was reported to not increase accuracy for the task. Early work using neural networks (Xiao et al., 2016) applied recurrent neural networks to the task of observational coding. Hasan et al. (2016) conducted an extensive study of different machine learning methods and feature sets applied to the same MI task, finding that support vector machines outperformed other algorithms and finding that their Convolutional neural network (CNN) using general domain word embeddings did not seem to perform well for tasks with a large number of classes. Others (Gibson et al., 2019) have developed a multi-label multi-task neural network in which the task of observational coding in both MI and CBT domains were learned jointly.

More recently, a hierarchical deep learning approach (Cao et al., 2019), somewhat similar to the architecture outlined here, has been published for the tasks of classifying patient and therapist utterances in motivational interviewing. However, in that work the authors train two different models for the two different tasks with only therapist information used as context for classifying therapist utterances, and only patient information used as context for classifying patient utterances.

The aim of the current study was to train a deep learning model to automatically code patient utterances during text-based internet-enabled CBT (IECBT). This approach differs from recent research (Can et al., 2015; Cao et al., 2019; Hasan et al., 2016) in a number of respects. First, our deep learning approach optimizes for the task of classifying therapist and patient language jointly. It is likely that the type of therapist utterance preceding the patient response has information that can disambiguate patient categories. For example, a response of "yes, definitely" after setting homework might be categorized differently than if it appeared after a statement that arranged the next session. Second, we use word-embeddings that were pre-trained on a large corpus of text from the domain of text-based IECBT. It is therefore much more likely that the semantics of the word-embeddings used in our model are more appropriate for our task than word-embeddings trained on, for example, a News corpus. Finally, and more importantly, whereas previous research on automatically coding transcripts has used the predicted output codes to measure therapist competencies only, here we also applied the deep learning model to a large dataset of therapy transcripts and used the model output to determine the association between patient utterance codes and clinical outcomes using logistic regression.

In a previous study, we trained a deep learning model to categorize therapist utterances from approximately 90,000 h of text-based IECBT (Ewbank et al., 2019). In text-based IECBT, a patient communicates with a qualified CBT therapist using a real-time text-based message system. In the present study, we use a similar approach to train a deep learning model to categorize patient utterances taken from a data set of ~34,000 patients receiving text-based IECBT (~188,000 h of therapy). Our aim was to determine whether a deep learning model could achieve human-level agreement on the task of automatically coding patient utterances, and to enter the output of this model into a logistic regression analysis to determine the association between outcomes and patient responses in a real-world clinical data set.

Similar to previous work, we coded patient language that expressed a movement away from or resistance to change as "counter-change talk," and patient language that expressed a desire or commitment to change as "change-talk." Given that CBT is informed by the principle that mental disorders are maintained by cognitive and behavioural phenomena (Beck, 1970; Ellis, 1962), we included an additional category of change-talk relating to self-exploration (i.e., where the patient explores the links between thoughts, emotions and behaviour) (see Table I for details). Based on previous research, we predicted that counter-change talk would be associated with poorer treatments outcomes (i.e., lower rates of improvement and treatment engagement) and hypothesized that change-talk reflecting an active movement towards change or self-exploration would be positively associated with outcomes.

Methods

Experimental Design

All patients received text-based IECBT for the treatment of a mental health disorder between June 2012 and October 2019. Patients were being treated for a broad range of mood and anxiety disorders, most commonly depressive episode and generalized anxiety disorder (see Supplementary Table S1). Text-based IECBT was delivered using a commercial package currently used in the English NHS, provided by Ieso Digital Health (https://www.iesohealth.com/en-gb), following internationally recognized

Table I. Response categories and guidelines (definitions and examples) used to tag patient utterances.

Category	Definition	Examples
Change-Talk Active	Patient responses that reflect a desire or commitment to change (including evidence that they have enacted some change such as doing their homework).	*"I don't want to live like this anymore"* *"Yes, I did fill in my thought record. I found it really helpful."*
Change-Talk Exploration	Patient responses that reflect a past, present or future state of mind and show evidence that the patient is using self-exploration and reflection of their problems.	*"So, I think that if I did that my heart would be racing, and I would get very nervous."*
Counter Change-Talk	Patient responses that move away from the target behaviour (including evidence of not-engaging in therapy).	*"I was unable to do the homework"* *"I don't think that would work"* *"I won't be able to"*
Follow/Neutral	Patient responses which do not favour either change or counter-change talk (including simple clarifications and non-therapeutic chit-chat).	*"Hello," "Ok," "I see"* *"Could you put that another way?" "What do you mean?"*
Describing Problems	Patient responses that describe external problems in their lives (i.e., not directly related to presenting problem).	*"I'm in a lot of debt and can't pay it off"* *"My mother has been very ill"*

standards for information security (ISO 27001; https://www.iesohealth.com/en-gb/legal/iso-certificates). NICE-approved disorder-specific CBT treatment protocols (National Institute for Health and Clinical Excellence, 2011), based on Roth and Pilling's CBT competences framework (Roth & Pilling, 2008), were delivered in a secure online therapy room via instant synchronous messaging, by a qualified CBT therapist accredited by the British Association for Behavioural & Cognitive Psychotherapies (BABCP). Patients self-referred or were referred by a primary healthcare worker directly to the service. All patients who referred to the service and were suitable (over 18 years old, registered with a GP and not at significant risk of suicide) were offered treatment. Each treatment session was scheduled for a duration of one hour. On average each hour-long session consisted of 70 utterances (34 patient and 36 therapist) and 1387 words (601 patient and 786 therapist).

Text-based IECBT is available to England NHS patients through the IAPT programme, a large-scale initiative aimed at increasing access to evidence-based psychological therapy for common mental health, whilst controlling costs (Clark et al., 2018). The information captured through IAPT's minimum dataset, including IECBT, is intended to support monitoring of implementation and effectiveness of national policy/legislation, policy development, performance analysis and benchmarking, national analysis and statistics, and national audit of IAPT services. Clinical audit studies do not require additional patient consent or ethical approval. When registering to use the Ieso service, patients are given a clear description of how their data may be used, including use of de-identified therapy transcripts to support research, and agree to such use by accepting the services' terms and conditions.

Patient Utterance Categories

We defined a total of five patient response categories informed by the MISC 1.1 guidelines (Glynn & Moyers, 2009) (see Table I). The MISC 1.1 codes patient statements that argue for change or against change. Here, we defined two types of change talk: "Change-Talk Active," in which a patient expresses a desire or commitment to change, including having enacted some change (e.g., completing homework), and "Change-Talk Exploration," in which a patient demonstrates self-exploration by reflecting on the processes that maintain their mental health problem. "Counter-Change Talk" included any statement that moved away from the target behaviour (including, for example, non-completion of homework). We included the category of "Describing Problems" to capture utterances in which the patient describes problems considered external to their presenting problem (e.g., being in debt). A final category of "Follow/Neutral" was included, representing patient responses which do not favour either change-talk or counter-change talk and do not fit in any other category.

Annotated Dataset

We obtained the dataset used in prior work (Ewbank et al., 2019) in which therapist utterances had been manually annotated according to a coding system developed for CBT. This dataset has since been extended and now consists of 483 transcripts with annotated therapist utterances. We extracted transcripts from this dataset to be annotated by four research scientists (two with postdoctoral training in psychology, two with postdoctoral training in computer science) according to the patient codes outlined in Table I. A training stage consisted of the annotators coding 10 transcripts according to the guidelines. The annotators met to discuss the agreements and disagreements for those 10 transcripts and the guidelines were updated to reduce the potential for future disagreements. After this round of discussions, each annotator was given 100 transcripts to annotate with no further meetings. Each scientist annotated 80 non-overlapping hours of therapy (12,668 patient utterances in 320 transcripts) and in order to estimate inter-annotator agreement, all scientists annotated a common pool of 20 h of therapy (657 patient utterances). Each patient utterance was tagged with one or more of the five categories, although it was found that annotators very rarely used multiple tags. The resultant dataset consists of transcripts in which both the therapist and the patient side of the conversation have been manually annotated.

Deep Learning Model

We modified the architecture of the deep learning model used in previous work (Cummins et al., 2019; Ewbank et al., 2019) to also automatically classify each patient utterance into one or more of the five categories previously outlined. The existing deep learning model classifies therapist utterances into one (or more) of 24 categories, and therefore we adapted this model to perform the patient utterance classification task in conjunction. Figure S2 outlines the architecture of the model used in this paper.

In our model, words are represented as word embeddings (dense vectors) that capture the semantics of words in a general sense. Most approaches to generating word embeddings utilize the *distributional hypothesis* in distributional semantics. The hypothesis posits that words that have similar neighbours tend to have similar semantics. For example, the words *dog* and *hound* might be considered semantically similar as they both appear proximal to such words as *bark*, *bite*, *tail*, and *paws,* in a large corpus of text. This understanding has led to the development of a number of unsupervised algorithms that aim to learn word embeddings using only a large corpus of natural language text as input (Baroni et al., 2014; Turney & Pantel, 2010).

We applied word2vec (Mikolov et al., 2013), a neural network approach to computing word-embeddings, to a pre-processed version of our entire therapy corpus (300,000 h of therapy). We pre-processed the corpus by tokenizing the text according to whitespace and punctuation and then by lower-casing all tokens (keeping punctuation as individual tokens). We ran word2vec using the skip-gram model with a context window size of 10 and a token frequency threshold of 5 to generate 91,470 word embeddings of length 100. These 91,470 words define the vocabulary of our system.

We modelled each utterance in a transcript as a sequence of these word embeddings and fed these into a bidirectional LSTM (Long Short-Term Memory) (Hochreiter & Schmidhuber, 1997). An LSTM is a popular deep-learning architecture that has been applied to many sequence labelling tasks in natural language processing. They are particularly well-suited to natural language tasks because they can encode context and word order when they are important for the task (e.g., although containing the same words, the sentence *the dog bit the man* is semantically different to *the man bit the dog*). For each token, the LSTM outputs a representation of the token in context. We used a hidden dimension of length 400 and then used max-pooling over all these tokens to result in a fixed-length vector of 400 to represent the utterance. These fixed-length utterance representations are fed into another LSTM that encodes these utterances in the context of the entire transcript (again of dimension 400). We encoded the patient or therapist role as a binary variable and concatenated this to the fixed-length utterance-in-context representation. Each utterance is now modelled as a vector of length 401 that has access to information about the tokens in the utterances, all utterances that preceded it, and whether it is a patient or therapist utterance.

This representation is then fed into a single output layer that maps the utterance-in-context representations to a fixed set of non-mutually exclusive classes. As in previous research, a sigmoid activation function on the output layer ensures a multilabel output vector. The output vector consists of a vector of length 4 (four patient classes with "follow-neutral" modelled as all zeros) concatenated with a vector of length 23 (23 therapist classes with the "Other" class modelled as all zeros). The output labels are masked by a role mask, which means that therapist labels are never predicted for patient utterances and vice versa. Furthermore, a loss mask ensures that gradients for patient labels are only backpropagated on four of the outputs, and similarly gradients for the therapist labels are only backpropagated on 23 of the outputs. All deep learning code was implemented in Tensorflow.

Model Training

As information about the categories of therapist utterances preceding a patient utterance are likely to be useful when classifying the patient utterance, we trained the deep learning model to optimize for both patient and therapist utterance classification tasks simultaneously. We used randomly selected training, validation, and test sets for both tasks (i.e., randomly selected at the outset). The splits for the patient language classification task were 270, 35, and 20 (where 15 of the singly annotated transcripts were retained for a future study). The test set of 20 transcripts is the multiply annotated dataset outlined earlier. For training the therapist language classification task, we used splits of 401 and 45 for training and validation respectively as we had more data available (we do not report test results for the therapist language task as it lies outside the scope of this work). We corrected for the imbalance in the amount of data available for each task by reweighting the loss instances of the patient language task (increasing the loss during training by 1.48 such that both tasks were equally represented).

Due to the large class imbalance for patient categories, we used a weighted sigmoidal cross-entropy loss. In preliminary experiments, models trained without a weighted loss predicted minority classes less frequently than expected. In order to give more importance to minority patient classes, we weighted the classes inversely related to the square-root of class support (Cui et al., 2019), a common heuristic used in the literature. For example, if a particular class only accounted for 5% of instances seen, its weighting for positive examples would be $0.95^{0.5}$, while the weighting for negative examples would be $0.05^{0.5}$.

We selected the model with the best macro averaged F1-score (which is the harmonic mean of precision and recall). All hyperparameters used in this work (embedding dimensionality, hidden layers, dropout, pooling) are identical to previous work (Cummins et al., 2019; Ewbank et al., 2019). None of the transcripts used in the training or validation step had multiple human annotations. We used the 20 transcripts with multiple human annotations as a test set to report the inter-annotator results and final model accuracy.

Outcomes

Clinical outcomes were measured in terms of reliable improvement and IAPT-engagement and were included as binary measures (i.e., 0 or 1). Following IAPT guidelines, a patient was classed as engaged if they attended two or more treatment sessions (Clark et al., 2018). This is the minimum amount of therapy a patient must receive such that pre- and post-treatment measures can be collected and clinical outcomes estimated. Reliable improvement was calculated based on two severity measures: PHQ-9 (Kroenke et al., 2001) and GAD-7 (Spitzer et al., 2006), corresponding to depressive and anxiety symptoms respectively. Both measures were completed by the patient at initial assessment and before every therapy session. Other symptom severity measures were not examined as only PHQ-9 and GAD-7 are mandatorily collected within the IAPT framework. A reduction of 6 points or more on the PHQ-9 scale or a reduction of 4 points or more on the GAD-7 scale between two-time points (while not showing a significant increase in the other outcome measure) is indicative of statistically reliable improvement in symptom severity. For IAPT-engaged patients, the difference between scores at initial assessment and last treatment session for PHQ-9 and GAD-7 was used to determine patients' improvement status. Consistent with IAPT guidelines, outcomes measures were based on scores when patients were discharged from treatment, irrespective of whether they were coded by their therapist as having completed or dropped out of treatment.

According to IAPT, a patient scoring 10 or more in the PHQ-9 (range 0–27) or 8 or more in the GAD-7 (range 0–21) is considered to be suffering from clinically significant symptoms (clinical caseness). A patient is classed as recovered if they are above the clinical caseness threshold for either the PHQ-9 or GAD-7 metrics at start of treatment, and below the clinical caseness threshold at the end of treatment. As the definition of recovery is based on a patient crossing the clinical caseness threshold for these measures, patients whose initial scores are closer to that threshold have higher chances of recovery. Furthermore, by definition, recovery does not take into account whether the reduction in scores is greater than the measurement error of the scales (Gyani et al., 2013), as opposed to reliable improvement (Jacobson & Truax, 1991). Therefore, we chose to use reliable improvement as an outcome measure in this study.

Statistical Analysis

All analyses were performed in R (R Core Team, 2017). The original dataset comprised a total of 33,943 patients who had attended at least one treatment session of text-based IECBT. Cases starting below the clinical caseness threshold, with missing start or end PHQ-9, GAD-7 or Work and Social Adjustment Scale (WSAS) scores, and with sessions of fewer than 50 patient words or a session duration of less than five minutes were excluded from the analysis. This left a total of 28,809 patients (21,065 female) aged between 18 and 94 years old (median age = 32) included in the analysis. For the logistic regressions predicting reliable improvement, only patients who had engaged in treatment were included, leaving a total of 25,366 patients in each analysis.

Using the output of the deep learning model, the number of utterances of each category, averaged across all sessions, were calculated for each case. The final treatment session was excluded as outcome measures are taken prior to the commencement of each treatment session. A logistic regression analysis was performed to investigate the association between patient utterance categories and reliable improvement. Predictor variables were the mean number of patient utterances for each category across sessions and patient demographics of starting PHQ-9, GAD-7 and WSAS scores, Gender (Male, Female, Unstated/Unknown), Age, Ethnicity (White, Non-white), whether the patient suffered from a long-term physical condition (Yes, No, Unstated/Unknown), whether the patient was taking psychotropic medication at the start of treatment (Prescribed Not taking, Prescribed Taking, Not Prescribed, Unstated/Unknown), the patient's employment status (Employed, Unemployed, Other, Unstated/Unknown), whether the patient suffered from a disability (Yes, None), the patient's sexual orientation (Straight, Bi/Homosexual), and whether the patient was currently in a perinatal period (Yes/No) The total number of treatment sessions and mean duration of sessions were also included. The mean number of treatment sessions was 5.58 (range 1–39; SD = 3.42).

To determine whether outcomes were predictable based on patient utterances at the start of treatment, we performed analogous logistic regressions to determine the association between both reliable improvement and IAPT-engagement and patient utterance categories in the first treatment session. Predictors included in the models were identical to the first analysis, the exception being that the number of

Table II. F1-Score, precision, and recall for our deep learning model (M) on the subset of the test utterances where all four human annotators agreed (339 utterances out of 657).

Label	# actual	# predicted	F1-Score	Precision	Recall
Change-Talk Active	15	16	0.645	0.625	0.666
Change-Talk Explore	67	73	0.843	0.808	0.881
Neutral/Follow	250	235	0.940	0.970	0.912
Describing Problems	4	9	0.462	0.333	0.750
Counter Change-Talk	3	6	0.222	0.167	0.333

utterances for each category was taken from the first session only and total number of sessions was not included.

For all analyses, continuous predictor variables were scaled and standardized, such that odds ratios indicate the effect of an increase in one standard deviation of the predictor variable. Statistical significance was defined as p<.05 two-tailed, uncorrected. Multicollinearity analyses revealed that variance inflation factors were smaller than two for all predictor variables, confirming that regression models were not affected by the presence of multicollinearity.

Results

Deep Learning Model

The distribution of labels applied by the annotators is shown in Figure S1. The most frequent category used by the annotators was "Follow-Neutral" with the least frequent being "Describing Problems" and "Counter-Change Talk." Table II shows the performance of the model on the subset of utterances from the test data on which all human annotators agreed. This subset represents the most unambiguous cases of each class in the test data. The model shows good performance on three of the five classes. We can see that the hardest classes (lowest F1 score) for the model are "Describing Problems" and "Counter-Change Talk."

We also performed an inter-rater analysis of the four human annotators (A, B, C, and D) and our deep learning model (M) on the 20 multiply annotated transcripts. Supplementary Table S5 shows Cohen's kappa, a chance-corrected agreement metric, between each annotator pair on the 657 test utterances. In general, there is moderate agreement (0.4–0.6). Also shown is the agreement between the model and all annotators. These results tend to indicate that the model is approaching human-level performance for this task. Furthermore, Figure 1 shows the inter-rater agreement of individual patient categories for both human-human and human-model pairs. The deep learning model performs as good as a human annotator for three of the categories but is substantially worse on "Counter-Change Talk." However, the model performs better

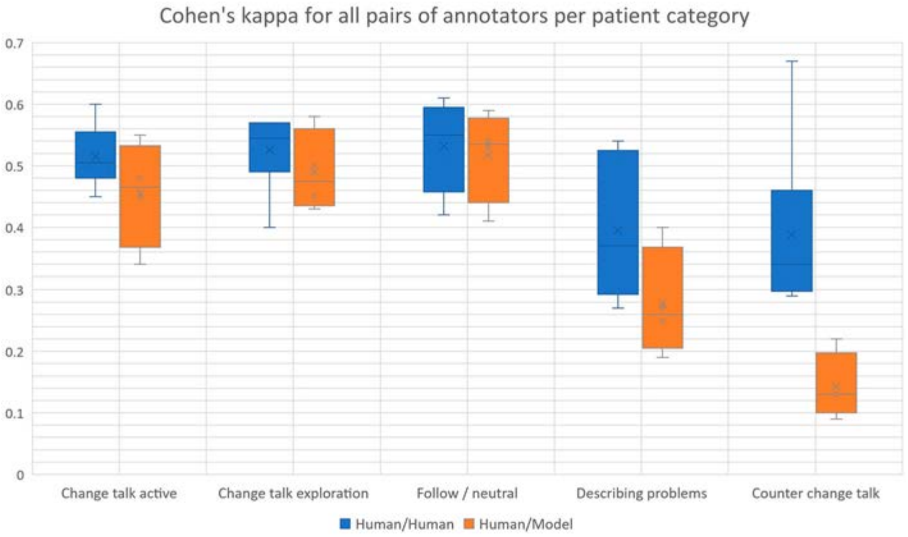

Figure 1. **Inter-annotator agreement for individual patient categories for all human-human and human-model pairs of annotators**. Cohen's kappa for all six human-human pairs and all four human-model pairs of annotators. As Cohen's kappa is chance corrected where zero indicates chance agreement, the deep learning model is significantly better than chance for all categories.

Table III. Predictors of reliable improvement across all treatment sessions.

Predictors	Mean (sd)	Odds Ratio	95% CI	z	*p*
Change-Talk Active	3.59 (1.75)	1.38	1.34–1.43	19.13	<0.001
Change-Talk Explore	11.41 (6.14)	1.14	1.10–1.19	6.40	<0.001
Neutral/Follow	21.53 (13.17)	0.90	0.87–0.94	−5.27	<0.001
Describing Problems	1.67 (1.73)	0.94	0.91–0.97	−4.26	<0.001
Counter-Change Talk	1.07 (0.96)	0.80	0.77–0.82	−15.11	<0.001

Note. Output of logistic regression investigating association between reliable improvement and the number of patient utterances for each of the five categories averaged across treatment sessions, adjusting for patient demographics. Odds ratios indicate the effect of an increase in one standard deviation of the number of utterances on the odds of reliable improvement in patient symptoms. Mean and standard deviation of number of patient utterances for each category averaged across treatment sessions are shown.

than chance-agreement (i.e., Cohen's Kappa greater than zero) for all categories demonstrating that the model has predictive power for all five categories.

Logistic Regression: Association Between Outcomes and Patient Utterance Categories

We performed a logistic regression to determine the association between reliable improvement in symptoms and patient utterances across the course of treatment (Table III). As hypothesized, the results revealed that the quantity of Counter-Change Talk utterances across treatment was negatively associated with reliable improvement (Odds Ratio [OR], 0.80; 95% CI, 0.77–0.82). Describing Problems utterances (OR, 0.94; 95% CI, 0.91–0.97 and Neutral/Follow utterances also showed a negative association with improvement (OR, 0.90; 95% CI, 87–93).). By contrast, the quantity of Change-Talk Active utterances (OR, 1.38; 95% CI, 1.34–1.43) as well as Change-Talk Explore utterances (OR, 1.14; 95% CI, 1.10–1.19) were positively associated with improvement. Odds ratios and CIs for all patient demographic variables included in the regression model can be found in Supplementary Table S2.

Next, we performed a logistic regression to determine whether patient utterances at the start of treatment (first treatment session) were predictive of reliable improvement at the end of treatment (Table IV). Again, we found that the quantity of Counter-Change Talk utterances was negatively associated with improvement (OR, 0.94; 92% CI, 0.92–0.97). There was also a negative association between improvement and Neutral/Follow utterances (OR, 0.93; 95% CI, 0.90–0.96) and Describing Problems utterances (OR, 0.97; 95% CI, 0.94–0.99). The quantity of Change-Talk Active utterances (OR, 1.20; 95% CI, 1.17–1.24) and Change-Talk Explore utterances (OR, 1.12; 95% CI, 1.08–1.16) were both positively associated with reliable improvement. Odds ratios and CIs for all patient demographic variables included in the regression model can be found in Supplementary Table S3.

Finally, we investigated whether patient utterances in the first treatment session were predictive of IAPT-engagement (i.e., a patient returning for a second treatment session). The results of a logistic regression (Table V) revealed that Counter-Change Talk utterances (OR, 0.92; 95% CI, 0.89–0.96) and Follow/Neutral utterances (OR, 0.90; 95% CI, 0.86–0.94) were both negatively associated with engagement. There was a positive association between IAPT-engagement and both Change-Talk Explore (OR, 1.26; 95% CI, 1.20–1.32) and Change-Talk Active utterances (OR, 1.14; 95% CI, 1.10–1.19). There was also a positive association with Describing Problems utterances (OR, 1.07; 95% CI, 1.03–1.12), which contrasts with the negative association found

Table IV. First session predictors of reliable improvement.

Predictors	Mean (sd)	Odds Ratio	95% CI	z	*p*
Change Talk Active	2.79 (2.45)	1.20	1.17–1.24	11.60	<0.001
Change Talk Explore	11.78 (8.02)	1.12	1.08–1.16	6.25	<0.001
Neutral Follow	21.74 (14.08)	0.93	0.90–0.96	−4.31	<0.001
Describing Problems	2.14 (3.14)	0.97	0.94–0.99	−2.38	0.018
Counter Change Talk	0.90 (1.40)	0.94	0.92–0.97	−4.24	<0.001

Note. Output of logistic regression investigating association between reliable improvement and number of patient utterances for each of the five categories occurring in the first treatment session, adjusting for patient demographics. Odds ratios indicate the effect of an increase in one standard deviation of the number of utterances on the odds of reliable improvement in patient symptoms. Mean and standard deviation of number of patient utterances for each category in the first treatment session are shown.

Table V. First session predictors of IAPT-engagement.

Predictors	Mean (sd)	Odds Ratio	95% CI	z	*p*
Change-Talk Active	2.75 (2.42)	1.14	1.10–1.19	6.42	<0.001
Change-Talk Explore	11.58 (8.01)	1.26	1.20–1.32	9.34	<0.001
Neutral/Follow	21.73 (14.01)	0.90	0.86–0.94	–4.92	<0.001
Describing Problems	2.12 (3.11)	1.07	1.03–1.12	3.51	<0.001
Counter Change-Talk	0.92 (1.43)	0.92	0.89–0.96	–4.52	<0.001

Note. Output of logistic regression investigating association between IAPT-engagement and number of patient utterances for each of the five categories occurring in the first treatment session, adjusting for patient demographic. Odds ratios indicate the effect of an increase of one standard deviation in the number of utterances on the odds of engagement. Mean and standard deviation of number of patient utterances for each category in the first treatment session are shown.

with reliable improvement. Odds ratios and CIs for all patient demographic variables included in the regression model can be found in Supplementary Table S4.

To demonstrate the clinical importance of changes in the number of utterances on outcomes, we provide an example of a positive and negative predictor from the analyses of first session utterances. We split patients into tertiles based on the mean number of utterances for each category and excluded the middle tertile in order to clearly separate the groups. Figure 2 shows the improvement rate for patients in the first and third tertile for the number of Change-Talk Active utterances in the first session. The results show a 10% increase in improvement rates for those patients expressing a greater number of Change-Talk Active utterances. Figure 3 shows the engagement rate for patients in the first and third tertile of Counter-Change Talk utterances in the first session. This reveals a 3% decrease in engagement rates for patients expressing a greater number of Counter-Change Talk utterances. We note that representing the difference in improvement and engagement rates in this manner does not account for differences in other variables included in the regression model, however we include these figures to provide an easily interpretable representation of the clinical importance of patient language in predicting outcomes.

Discussion

In this study, we developed a deep learning model that was able to automatically categorize patient utterances from a large unstructured clinical data set of text-based internet-enabled CBT session transcripts. While previous work has applied machine

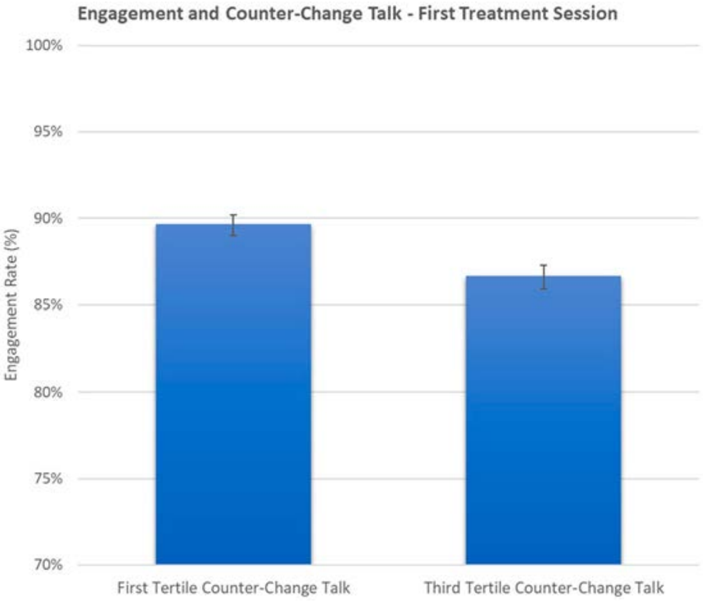

Figure 2. **Reliable improvement rates (%) for patients with a low and high number of Change-Talk Active utterances in the first treatment session**. Error bars represent 95% CIs. Patients are split into tertiles based on the number of Change-Talk Active utterances in the first treatment session only.

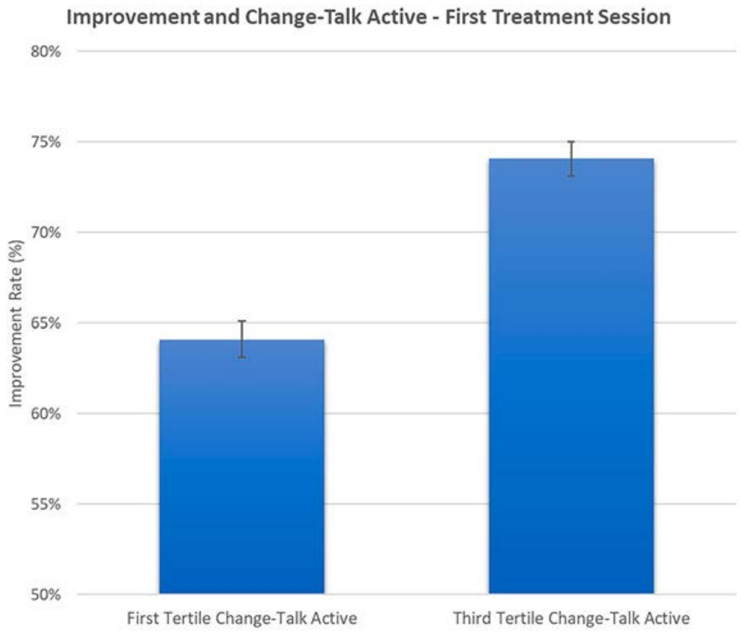

Figure 3. **IAPT-engagement rates (%) for patients with a low and high number of Counter-Change Talk utterances in the first treatment session**. Error bars represent 95% CIs. Patients are split into tertiles based on the number of Counter-Change Talk utterances in the first treatment session only.

learning to automatically annotate MI transcripts (Can et al., 2015; Cao et al., 2019; Hasan et al., 2016), here we provide the first example of the use of machine learning to automatically categorize patient utterances in text-based IECBT as well as an investigation into the association between patient responses and outcomes. The time consuming and labour-intensive nature of manual transcription means that previous studies investigating patient language during CBT have been limited to relatively small sample sizes. Here, we demonstrate that deep learning provides a promising new direction for annotating therapeutic conversations at a large scale (an initial sample of ~34,000). Taken together with previous work in which we used deep learning to develop a model categorizing therapist utterances (Ewbank et al., 2019), these results suggest that an utterance level analysis of therapy content is predictive of outcomes in text-based IECBT.

The deep learning model developed here achieves moderate agreement levels with human annotators, which is similar to the agreement levels between humans. Previous studies (Lombardi et al., 2014; Westra, 2011) measuring inter-rater agreement of MISC and modified MISC coding schemes tend to do so at the transcript level. However, the most comprehensive study of inter-rater agreement using the MISC system (Lord et al., 2015) has shown that inter-rater agreement at the transcript level is consistently higher than at the utterance level, and that some theoretically important codes (e.g., types of change talk) have much lower human agreement levels due to their lower frequency (a result consistent with our findings in Figure 1). Despite these limitations the inter-rater agreement levels for the five patient categories in this paper are broadly in line with previous research measuring patient MISC categories at the utterance level (Lord et al., 2015).

Using the output of the deep learning model applied to a large corpus of text-based IECBT transcripts, we performed logistic regression analyses to determine the association between patient utterances and clinical outcomes. As hypothesized, the results revealed that Counter-Change Talk utterances measured in the first treatment session, and across treatment sessions, showed a negative association with reliable improvement and with IAPT-engagement (i.e., a patient returning for a second treatment session). These findings accord with previous work showing that counter-change talk in the first session of CBT predicts poorer outcomes (Lombardi et al., 2014; Poulin et al., 2019).

Although previous work has not consistently found a positive association between change-talk and outcomes in MI (Magill et al., 2018) or in CBT (Hunter et al., 2014; Lombardi et al., 2014), here we found that the quantity of both Change-Talk Active and Change-Talk Exploration was significantly associated with greater odds of reliable improvement and engagement. It has been proposed that the reason why previous work has found counter –change talk to be a more consistent predictor of outcomes than change-talk is that increased change-talk (and restricted expressions of counter-change talk) may reflect patients' interest in making a good impression when starting treatment (Lombardi et al., 2014). We speculate that the positive association between change-talk and outcomes in the current study is likely to reflect increased power due to the large sample size. This emphasizes the importance of adequate statistical power to detect smaller effects that may not be apparent in sample sizes typically employed in psychotherapy research (Bell et al., 2013).

We also defined an utterance category entitled "Describing Problems" to capture patient utterances that did not suggest a move towards or away from change but instead reflected an element of complexity or external difficulties relating to a patient (e.g., financial problems). As predicted, we found that increased Describing Problems utterances across treatment sessions was associated with lower odds of improvement and showed a borderline negative association with improvement based on the first treatment session. Interestingly, we found an increased quantity of Describing Problems utterances in the first session was positively associated with engagement. We speculate that patients who express more external problems in the first session may feel a greater initial benefit from therapy and therefore be more motivated to return for a second treatment session, however increased talk of external problems appears to be an indicator that a patient is less likely to show improvement at the end of treatment.

The odds ratios reported here indicate the effect of an increase in one standard deviation of the number of utterances for each category on the odds of reliable improvement in patient symptoms. While a small increase in the number of utterances is likely to have a relatively small effect on outcomes (after accounting for an additional 14 treatment and patient variables), we note that the odds ratio for change-talk active utterances is larger than that of most patient variables, including the presence of a long-term physical condition and the total number of treatment sessions (Table III and Table S2). Thus, the association between utterances and outcomes appears comparable to the effects of patient variables typically used to predict outcomes. In addition, given differences in patient prognosis, the relative effects of therapeutic interventions may be obscured by patients who are expected to improve irrespective of treatment (Derubeis et al., 2014; Lorenzo-Luaces et al., 2017)

Limitations

A limitation of these data is that the class Follow/Neutral accounts for the majority (58%) of utterances. We found that the quantity of Follow/Neutral utterances was negatively associated with outcomes, suggesting that simply following or agreeing with the therapist is less likely to lead to improvements in patient symptoms. However, it is also possible that this category comprises a number of sub-types of response that may themselves be important predictors of outcomes. Thus, further work could identify distinct sub-categories within Follow/Neutral to obtain a more detailed/accurate representation of patient language. Furthermore, while the performance of our deep learning model on the least frequent class, Counter-Change Talk, is significantly better than chance agreement (i.e., it has predictive power), it is worse than human-level agreement (See Figure 1). This is likely due to the low number of examples of Counter-Change Talk that are presented to the model during training and due to the fact that this is the category with the lowest human inter-rater agreement. Future work will look at ways of overcoming this limitation by over-sampling or by further annotation. Given that the data used in the current study was obtained using text-based internet-enabled CBT, the extent to which our findings generalize to other types of therapy (e.g., psychodynamic) are unclear. However, our findings accord with previous studies investigating client language in both motivational interviewing and face-to-face CBT, which found lower levels of counter-change talk during early sessions are associated with more positive outcomes.

Our analysis also does not account for qualitative variation in how treatment was delivered, meaning we are unable to determine to what extent a Counter-Change Talk response can be attributed to an inappropriately applied therapeutic technique or to the patient's opposition to therapy and/or resistance to change. However, when coupled with measures of therapist language (Ewbank et al., 2019), future work will be able to determine the interactive nature of the therapist/patient relationship during a therapy session, and this is something we are currently investigating in our lab. It should also be noted that the patients included in this study were being treated for a range of conditions, including depression and anxiety disorders, thus future work is needed to determine whether the association between outcomes and patient language is dependent upon diagnosis/treatment delivered or is predicted by specific symptom profiles.

Conclusion

Improving the effectiveness of CBT depends on an understanding not only of what aspects of therapy are effective in changing symptoms, but also an understanding of why and how patients do not engage with therapy. Observational coding measures of patient language have been shown to be a reliable predictor of outcomes in MI and CBT, however the resource-intensive nature of this exercise means studies are limited to relatively small sample sizes. We demonstrate that the application of deep learning-facilitated automatic annotation provides an effective means of obtaining categorization of patient utterances during text-based IECBT, at a scale previously beyond the scope of psychotherapy research, providing evidence of both positive and negative associations between patient utterances categories and outcomes. Coupled with an automated understanding of therapist language, deep learning can be used to enable a data-driven understanding of the relationship between therapeutic interventions, patient language, and clinical outcomes.

Acknowledgements

We would like to thank Rebecca Baker-Munton for her work in the early stages of this project. This study was funded by Ieso Digital Health and Innovate UK: Project No. 104544.

Supplemental data

Supplemental data for this article can be accessed at https://doi.org/10.1080/10503307.2020.1788740

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. Context-predicting semantic vectors. In *52nd Annual Meeting of the association for Computational Linguistics, ACL 2014—Proceedings of the Conference* (Vol. 1). https://doi.org/10.3115/v1/P14-1023

Beck, A. T. (1970). Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy*, *1*(2), 184–200. https://doi.org/10.1016/S0005-7894(70)80030-2

Bell, E. C., Marcus, D. K., & Goodlad, J. K. (2013). Are the parts as good as the whole? A meta-analysis of component treatment studies. *Journal of Consulting and Clinical Psychology*, *81*(4), 722–736. https://doi.org/10.1037/a0033004

Button, M. L., Westra, H. A., Hara, K. M., & Aviram, A. (2015). Disentangling the Impact of resistance and ambivalence on therapy outcomes in cognitive behavioural therapy for generalized anxiety disorder. *Cognitive Behaviour Therapy*, *44*(1), 44–53. https://doi.org/10.1080/16506073.2014.959038

Can, D., Atkins, D. C., & Narayanan, S. S. (2015). *A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations*. Proceedings of the Annual Conference of the International speech Communication association, INTERSPEECH, 2015-Janua06-10-September.

Cao, J., Tanana, M., Imel, Z. E., Poitras, E., Atkins, D. C., & Srikumar, V. (2019). Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *Proceedings of the 57th conference of the association for computational linguistics* (pp. 5599–5611). https://doi.org/10.18653/v1/p19-1563

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): An analysis of public data. *The Lancet*, *391*(10121), 679–686. https://doi.org/10.1016/S0140-6736(17)32133-5

Cui, Y., Jia, M., Lin, T., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. *2019 IEEE/CVF conference on computer Vision and pattern recognition (CVPR)* (pp. 9260–9269). https://doi.org/10.1109/CVPR.2019.00949

Cuijpers, P. (2016). Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. *Evidence Based Mental Health*, *19*(2), 39–42. https://doi.org/10.1136/eb-2016-102341

Cummins, R., Ewbank, M. P., Martin, A., Tablan, V., Catarino, A., & Blackwell, A. D. (2019). TIM: A tool for gaining insights into psychotherapy. *The World Wide Web Conference*, 3503–3506. https://doi.org/10.1145/3308558.3314128

Derubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014). Understanding processes of change: How some patients reveal more than others-and some groups of therapists less-about what matters in psychotherapy. *Psychotherapy Research : Journal of the Society for Psychotherapy Research*, *24*(3), 419–428. https://doi.org/10.1080/10503307.2013.838654

Doss, B. D., & Atkins, D. C. (2006). Investigating treatment mediators when simple random assignment to a control group is not possible. *Clinical Psychology: Science and Practice*, *13*(4), 321–336. https://doi.org/10.1111/j.1468-2850.2006.00045.x

Drieschner, K. H., Lammers, S. M. M., & van der Staak, C. P. F. (2004). Treatment motivation: An attempt for clarification of an ambiguous concept. *Clinical Psychology Review*, *23*(8), 1115–1137. https://doi.org/10.1016/j.cpr.2003.09.003

Ellis, A. (1962). *Reason and emotion in psychotherapy*. Lyle Stuart.

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2019). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, https://doi.org/10.1001/jamapsychiatry.2019.2664

Gibson, J., Atkins, D., Creed, T., Imel, Z., Georgiou, P., & Narayanan, S. (2019). Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, *1*. https://doi.org/10.1109/TAFFC.2019.2952113

Glynn, L. H., & Moyers, T. B. (2009). *Manual for the Motivational Interviewing Skill Code (MISC), Version 1.1: Addendum to MISC 1.0*. (Vol. 0, Issue April, pp. 1–15).

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606. https://doi.org/10.1016/j.brat.2013.06.004

Hasan, M., Kotov, A., Idalski Carcone, A., Dong, M., Naar, S., & Brogan Hartlieb, K. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of Biomedical Informatics*, *62*, 21–31. https://doi.org/10.1016/j.jbi.2016.05.004

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1144/GSL.MEM.1999.018.01.02

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The Efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/s10608-012-9476-1

Hunter, J. A., Button, M. L., & Westra, H. A. (2014). Ambivalence and alliance ruptures in cognitive behavioral therapy for generalized anxiety. *Cognitive Behaviour Therapy*, *43*(3), 201–208. https://doi.org/10.1080/16506073.2014.899617

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. American Psychological Association. https://doi.org/10.1037/0022-006X.59.1.12

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling : a meta-analysis. *Psychological Bulletin*, *141*(4), 747–768. https://doi.org/10.1037/bul0000015

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *JAMA Psychiatry*, *59*(10), 877–883. https://doi.org/10.1001/archpsyc.59.10.877

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lombardi, D. R., Button, M. L., & Westra, H. A. (2014). Measuring motivation: Change talk and counter-change talk in cognitive behavioral therapy for generalized anxiety. *Cognitive Behaviour Therapy*, *43*(1), 12–21. https://doi.org/10.1080/16506073.2013.846400

Lord, S. P, Can, D, Yi, M, Marin, R, Dunn, C. W, Imel, Z. E, Georgiou, P, Narayanan, S, & Steyvers, M. (2015). Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of Substance Abuse Treatment*, *49*, 50–57. https://doi.org/10.1016/j.jsat.2014.08.005

Lorenzo-Luaces, L., DeRubeis, R. J., van Straten, A., & Tiemens, B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders*, *213*(January), 78–85. https://doi.org/10.1016/j.jad.2017.02.010

Magill, M., Apodaca, T. R., Borsari, B., Gaume, J., Hoadley, A., Gordon, R. E. F., Tonigan, J. S., & Moyers, T. (2018). A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology*, *86*(2), 140–157. https://doi.org/10.1037/ccp0000250

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119. https://doi.org/10.1162/jmlr.2003.3.4-5.951

National Institute for Health and Clinical Excellence. (2011). Common mental health disorders: The NICE guideline on identification and pathways to care. *The British Psychological Society and The Royal College of Psychiatrists*, https://doi.org/clinical guideline CG123.2011

NHS. (2018). *The Improving Access to Psychological Therapies Manual* (Issue June).

Poulin, L. E., Button, M. L., Westra, H. A., Constantino, M. J., & Antony, M. M. (2019). The predictive capacity of self-reported motivation vs. early observed motivational language in cognitive behavioural therapy for generalized anxiety disorder. *Cognitive Behaviour Therapy*, *48*(5), 369–384. https://doi.org/10.1080/16506073.2018.1517390

R Core Team. (2017). *R: A language and environment for statistical computing*.

Roth, A. D., & Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *36*(February), 129–147. https://doi.org/10.1017/S1352465808004141

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Westra, H. A. (2011). Comparing the predictive capacity of observed in-session resistance to self-reported motivation in cognitive behavioral therapy. *Behaviour Research and Therapy*, *49*(2), 106–113. https://doi.org/10.1016/j.brat.2010.11.007

Xiao, B., Can, D., Gibson, J., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. (2016). Behavioral coding of therapist language in addiction counseling using recurrent neural networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08-12-Sept* (pp. 908–912). https://doi.org/10.21437/interspeech.2016-1560