DOI: http://dx.doi.org/10.1037/cou0000382

Machine learning and natural language processing in psychotherapy research: Alliance as example use case

Simon B. Goldberg

University of Wisconsin – Madison

Nikolaos Flemotomos, Victor R. Martinez

University of Southern California

Michael Tanana, Patty Kuo

University of Utah

Brian T. Pace

University of Utah, VA Palo Alto Health Care System

Jennifer L. Villatte

University of Washington

Panayiotis Georgiou

University of Southern California

Jake Van Epps, Zac E. Imel

University of Utah

Shrikanth Narayanan

University of Southern California

David C. Atkins

University of Washington

Authors Note: Simon B. Goldberg, Department of Counseling Psychology, University of Wisconsin – Madison, Madison, WI, USA; Nikolaos Flemotomos, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA; Victor R. Martinez, Department of Computer Science, University of Southern California, Los Angeles, CA, USA; Michael Tanana, College of Social Work, University of Utah, Salt Lake City, UT, USA; Patty Kuo, Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA; Brian T. Pace, Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA and VA Palo Alto Health Care System, Palo Alto, CA, USA; Jennifer L. Villatte, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA; Panayiotis Georgiou, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA; Jake Van Epps, University of Utah Counseling Center, Salt Lake City, UT, USA; Zac E. Imel, Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA; Shrikanth Narayanan, Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA; David C. Atkins, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA.

Please address correspondence to: Simon B. Goldberg, Department of Counseling Psychology, University of Wisconsin – Madison, 335 Education Building, 1000 Bascom Mall Madison, WI, 53703, sbgoldberg@wisc.edu.

Drs. Tanana, Atkins, Narayanan, and Imel are co-founders with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling. The remaining authors report no conflicts of interest.

Funding was provided by the National Institutes of Health / National Institute on Alcohol Abuse and Alcoholism (NIAAA) under award R01/AA018673. Support for this research was also provided by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

Portions of the data presented in this manuscript were reported at the North American Society for Psychotherapy Research meeting in Park City, UT in September, 2018.

Machine learning and natural language processing in psychotherapy research:

Alliance as example use case

Abstract

Artificial intelligence generally and machine learning specifically have become deeply woven into the lives and technologies of modern life. Machine learning is dramatically changing scientific research and industry and may also hold promise for addressing limitations encountered in mental health care and psychotherapy. The current paper introduces machine learning and natural language processing as related methodologies that may prove valuable for automating the assessment of meaningful aspects of treatment. Prediction of therapeutic alliance from session recordings is used as a case in point. Recordings from 1,235 sessions of 386 clients seen by 40 therapists at a university counseling center were processed using automatic speech recognition software. Machine learning algorithms learned associations between client ratings of therapeutic alliance exclusively from session linguistic content. Using a portion of the data to train the model, machine learning algorithms modestly predicted alliance ratings from session content in an independent test set (Spearman's $U = .15$, $p < .001$). These results highlight the potential to harness natural language processing and machine learning to predict a key psychotherapy process variable that is relatively distal from linguistic content. Six practical suggestions for conducting psychotherapy research using machine learning are presented along with several directions for future research. Questions of dissemination and implementation may be particularly important to explore as machine learning improves in its ability to automate assessment of psychotherapy process and outcome.

Public Significance Statement: Our study suggests that client-rated therapeutic alliance can be predicted using session content through machine learning models, albeit modestly.

***Keywords*:** machine learning; natural language processing; methodology; artificial intelligence; therapeutic alliance

Transforming psychotherapy research with machine learning: Alliance as a case in point

"New directions in science are launched by new tools much more than by new concepts. The effect of a concept- driven revolution is to explain old things in new ways. The effect of a tool driven revolution is to discover new things that have to be explained." - Freeman Dyson (1998)

Whether or not we know it, and certainly whether or not we like it, machine learning (ML) is transforming modern life. From eerily prescient Google search suggestions or Amazon product recommendations to iPhones capable of understanding spoken language (i.e., *Siri*), ML undergirds many of the most commonplace technologies of industrialized society. Manifestations range from the seemingly benign or mundane to the perhaps more pernicious (e.g., targeted advertising). These contemporary conveniences are based on a family of quantitative methods that are rapidly changing science and technology and fall under the general umbrella of artificial intelligence. The term artificial intelligence has been defined as "the study of agents that receive percepts from the environment and perform actions" (Russell & Norvig, 2016, p. viii). Early work on artificial intelligence dates back to the 1950s (e.g., Turing, 1950). ML combines pattern recognition and statistical inference and plays an integral role within the inner workings of artificial intelligence. ML can be defined as "the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience" (Mjolsness & DeCoste, 2001, p. 2051).

The ways that ML has impacted scientific research and industry is hard to overstate (Jordan & Mitchell, 2015; Mjolsness & DeCoste, 2001Stead, 2018). Evidence for the widespread relevance of ML dates back several decades (e.g., detecting fraudulent credit card transactions; Mitchell, 1997). More recent ML-based innovations in medicine include detection of diabetic retinopathy (Gulshan et al., 2016), informing cancer treatment decision making

(Bibault, Giraud, & Burgun, 2016), and predicting disease outbreak (Chen, Hao, Hwang, Wang, & Wang, 2017). Innovations based on ML are occurring in basic science as well (e.g., materials science; Butler, Davies, Cartwright, Isayev, & Walsh, 2018). While not all ML applications in science and technology have gone smoothly (e.g., Google Flu consistently overestimating flu occurrence; Lazer, Kennedy, King, & Vespignani, 2014), the potential is unequivocal.

Efforts to apply ML within mental health care are also underway (for a recent scoping review, see Shatte, Hutchinson, & Teague, 2019). Examples include the use of passive sensing to predict psychosis (e.g., data collected from sensors built into modern smartphones; Insel, 2017; Wang et al., 2016), analysis of speech signals to infer symptoms of depression (France, Shiavi, Silverman, Silverman, & Wilkes, 2000; Moore, Clements, Peifer, & Weisser, 2008), prediction of treatment drop-out from ecological momentary assessment (Lutz et al., 2018), and the use of conversational agents (i.e., computers) for clinical assessment and even treatment (Miner, Milstein, & Hancock, 2017). While not incorporated in most settings, these ML-based innovations could dramatically change how mental health treatment and psychotherapy in particular is provided. Importantly, once an ML algorithm has been appropriately trained, it can be deployed at scale without additional human judgment.

The need for innovation in psychotherapy

Psychotherapy is in need of innovation. For one, mental health care matters: mental health conditions are extremely common and associated with enormous economic and social costs (Substance Abuse and Mental Health Services Administration, 2014; Whiteford et al., 2013). Psychotherapy is a frontline treatment approach (Cuijpers et al., 2014), with efficacy similar to psychotropic medications and with potentially longer lasting benefits and fewer side effects (Berwian, Walter, Seifritz, & Huys, 2017). Yet despite enormous investment in

psychotherapy in terms of therapist and client time and health care dollars (Olfson & Marcus, 2010), what actually happens in psychotherapy is largely unknown (i.e., is unobserved). Psychotherapy research remains heavily reliant on retrospective client or therapist self-report (e.g., Elliott, Bohart, Watson, & Murphy, 2018; Flückiger, Del Re, Wampold, & Horvath, 2018), limiting our understanding of actual therapist-client interactions that drive treatment. We do know that treatment outcomes vary widely, related to client (Lambert & Barley, 2001; Thompson, Goldberg, & Nielsen, 2018), therapist (Baldwin & Imel, 2013; Johns et al., 2018), relationship (e.g., therapeutic alliance; Flückiger, Del Re, Wampold, & Horvath, 2018), and treatment-specific factors.

One source of variability may be treatment quality. To date, however, there are no established and routinely implemented methods for quality control. The absence of quality control limits clinical training, supervision, and the development of therapist expertise (Tracey, Wampold, Lichtenberg, & Goodyear, 2014); decreases the ability to demonstrate quality to payers (Fortney et al., 2017); slows scientific progress in determining which treatments are likely to succeed and why; and restricts efforts to improve service delivery (Fairburn & Cooper, 2011). For these reasons, psychotherapy researchers have developed numerous observer rating systems to evaluate aspects of treatment quality (e.g., adherence and competence; Goldberg et al., in press; Webb, DeRubeis, & Barber, 2010). Behavioral coding has been invaluable in allowing researchers to understand what occurs in the moment between therapists and clients that may contribute to therapeutic change. However, human-coded rating systems are labor intensive, expensive to implement, and not widely used in community-based therapy (Fairburn & Cooper, 2011). Clients may also be asked to provide evaluation of treatment quality (e.g., measures of satisfaction, therapeutic alliance; Flückiger et al., 2018). Regular use of these kinds

of measures, while robust predictors of outcome (Flückiger et al., 2018), increase burden on clients and providers, are at risk for response set biases (e.g., social desirability) and random error, and have known psychometric limitations (e.g., ceiling effects; Tryon, Blackwell, & Hammel, 2008).

The new tools of psychotherapy research

Recent methodological advances may be quickly changing our ability to process the complex data of psychotherapy (Imel, Caperton, Tanana, & Atkins, 2017) and could allow automated assessment of treatment quality along with other outcome and process variables. Two related innovations include the development of natural language processing (NLP) and ML. As spoken language forms a key component of most psychotherapies, the ability to rapidly and reliably process speech (or text) data may allow routine assessment of treatment quality and evaluation of numerous other constructs of interest. Several recent proof-of-concept examples have appeared in the literature, including using NLP and ML to reliably code motivational interviewing treatment fidelity (Atkins, Steyvers, Imel, & Smyth, 2014; Imel et al., in press), to differentiate classes of psychotherapy (e.g., cognitive behavioral therapy and psychodynamic psychotherapy; Imel, Steyvers, & Atkins, 2015), and to identify linguistic behaviors of effective counselors in text-based crisis counseling (Althoff, Clark, & Leskovec, 2016).

The current study extends these efforts further by employing NLP and ML to predict one of the most studied process variables in psychotherapy: the therapeutic alliance (Flückiger et al., 2018). This was examined within the context of a large, naturalistic psychotherapy dataset drawn from a university counseling center. Sessions recordings were available for 1,235 sessions of 386 clients seen by 40 therapists. NLP and ML methods were used to predict client-rated alliance from session recordings.

Alliance is used as a test case to demonstrate the potential applicability of NLP and ML for several reasons. First, alliance is important for effective psychotherapy, based on its robust relationship with outcome (Flückiger et al., 2018). Second, alliance, unlike other more objective linguistic features (e.g., ratio of open and closed questions in motivational interviewing adherence coding; Miller, Moyers, Ernst, & Amrhein, 2003), requires a potentially higher-order of processing to assess (e.g., through the cognitive and affective system of a client, therapist, or observer providing alliance ratings). This additional level of abstraction likely makes automated prediction more difficult, but also more widely relevant if it can be accomplished. Third, alliance represents a relatively old concept (Bordin, 1979; Greenson, 1965) that may be less viable for concept-driven innovations (Dyson, 1998). New tools, however, could drive innovation in this area. There are also important open questions related to alliance, such as the proportion and cause of therapist and client contributions to alliance (Baldwin, Imel, & Wampold, 2007), the source of unreliability in alliance ratings across rating perspectives (i.e., client, therapist, and observer; Tichenor & Hill, 1989), the state- versus trait-like qualities of alliance (Zilcha-Mano, 2017), the potentially causal nature of alliance as a driver of symptom change (Falkenström, Granström, & Holmqvist, 2013; Flückiger et al., 2018; Zilcha-Mano & Errázuriz, 2017), and ways to include alliance assessment in routine clinical care without increasing participant burden (Duncan et al., 2003; Goldberg, Rowe, Ruan, Owen, & Miller, 2019). While NLP and ML are likely not panacea for resolving all outstanding debates regarding alliance, they may be useful research tools. Theoretically, these questions could be addressed more thoroughly if ML enabled alliance assessment on a much larger scale, particularly if ML models were built in a way to minimize construct irrelevant variance (e.g., social desirability). Ultimately, assessment of alliance could be automated using ML, providing clients and therapists

with ongoing information about this aspect of therapeutic process without the drawbacks (e.g., time required, psychometric issues) of repeated self-report assessment. Such technology could also be used to assess alliance directly from session transcripts or recordings.

Prior to presenting a preliminary attempt at assessing alliance using NLP and ML, it is worth introducing basic concepts involved in each methodology. This is, of course, intended to be a cursory treatment and interested readers are encouraged to review sources cited below.

Basics of NLP. NLP is a subfield of computer science and linguistics focused on the interaction between machines and humans through language (Jurafsky & Martin, 2014). NLP aims to understand human communication by processing and analyzing large quantities of textual data. Popular applications of NLP include machine translation (e.g., Google Translate), question-answering systems, or sentiment analysis (e.g., extraction of sentiments within social media).

Typically, NLP applications start with a collection of raw text documents (i.e., a language corpus). From this corpus, the first step is to extract or estimate quantitative features from the text. One of the most widely used NLP features is the bag-of-words representation (BoW). In BoW, each document is represented by counts of its unique words, without regard to the ordering of these words. Conceptually, BoW is a large crosstabulation table of words by documents. Other common text features include N-grams (Shannon, 1948), which are short multi-word phrases with *N* elements (e.g., bi-grams include two word phrases); dictionary-based features, such as those provided by Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Boyd, Jordan & Blackburn, 2015) or the General Inquirer (Stone, Bales 1962); and dialogue acts (Okada et al, 2016), which try to capture a high-level interaction between participants in a conversation (i.e. “statement”, “question”, etc.). More recently, linguistic units are converted to

a vector-space representation of either word (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013, 2013; Pennington, Socher & Manning, 2013) or sentence (Pagliardini, Gupta & Jaggi, 2018) embeddings, which capture the semantic context. Words (or sentences) that appear in similar contexts appear closer to each other in vector space, and semantic relationships are represented by the operations of addition and subtraction (e.g., v(king) – v(man) + v(woman) = v(queen) where v(w) is represents the vector for word w).

Basics of ML. The human brain has a remarkable ability to learn and recognize patterns from its surrounding environment. Machine Learning (ML) comprises a set of computational techniques simulating this capability (Haykin, 2009). As opposed to knowledge-based approaches, where a human designs an algorithm having specific rules in mind, ML is typically based on data-driven methods and on statistical inference. ML algorithms derive prediction rules from (typically) large amounts of data.

Two major paradigms in ML are unsupervised and supervised learning (Murphy, 2012). Similar to cluster analysis, unsupervised learning does not involve an outcome to predict but rather focuses on finding structure within a given set of data. Supervised learning is similar to regression modeling, in which an outcome (either discrete or continuous) is associated with a set of input data, and the ML algorithm is tasked with finding an optimal mapping function between the input data and the outcome (e.g., linking linguistic content with alliance ratings). Once such a mapping has been learned, it can be used to predict outcomes for new data. Since the goal of ML is to apply the algorithm on previously unseen data, ML analyses train algorithms on a subset of "training data" but are evaluated on a separate subset of "test data." Typical supervised learning algorithms include Support Vector Machines (SVMs), regularized linear or logistic regression, and decision trees (Murphy, 2012). Recently, there has been rapid

development and increased focus on artificial neural networks and deep learning techniques (Goodfellow, Bengio & Courville, 2016).

Method

Participants and Setting

Data were collected at the counseling center of a large, Western university. The counseling center provides approximately 10,000 sessions per year, with treatment focused on concerns common among undergraduate and graduate students (e.g., depression, anxiety, substance use, academic concerns, relationship concerns; Benton et al., 2003). Treatment is provided by a combination of licensed permanent staff (including social workers, psychologists, and counselors) as well as trainees pursuing masters- or doctoral-level mental health degrees (e.g., masters of social work, doctorate in counseling/clinical psychology).

Data were collected between September 11th, 2017 and December 11th, 2018. Both clients and therapists provided consent for audio recording of sessions and for use of recordings for the current study. Recordings were made from microphones installed in clinic offices and archived on clinic servers. Two microphones were hung from the ceiling in each room. One cardioid choir mic was hung to capture voice anywhere in the room and a second choir mic pointed in the direction where the therapist generally sits. In order for sessions to be recorded, clinicians had to start and stop recordings (i.e., sessions were not recorded automatically). All recordings were from individual therapy sessions (approximately 50 minutes in length). All audio recordings with associated alliance ratings were used (i.e., no exclusions were made). Alliance is assessed routinely in the clinic, with no standardized instructions regarding how therapists use these ratings in therapy.

The current study was integrated into the partner clinic with minimum modifications to the existing clinic workflow. One feature of the workflow is collecting alliance ratings *prior* to sessions, rather than asking clients to complete measures both before (e.g., symptom ratings) and after (e.g., alliance ratings) session. When making alliance ratings prior to session, clients were asked to reflect on their experience of alliance at their previous session (i.e., time – 1). In all models, alliance ratings were associated with the session they were intended to represent (e.g., ratings made prior to session 2 were associated with session 1). No alliance ratings were made prior to the initial session. Study procedures were approved by the relevant Institutional Review Board.

Clients were on average 23.77 years old (*SD* = 4.86). The majority of the sample identified as female (*n* = 214, 55.4%), with the remainder identifying as male (*n* = 158), non-binary (*n* = 5), genderqueer (*n* = 1), gender neutral (*n* = 3), female-to-male transgender (*n* = 1), and questioning (*n* = 2), with two choosing not to respond. The client sample predominantly identified as white (*n* = 294, 76.2%), with the remainder identifying as Latinx (*n* = 33), Asian American (*n* = 28), African American (*n* = 5), Pacific Islander (*n* = 2), Middle Eastern (*n* = 1), and multiracial (*n* = 21), with two choosing not to respond.

Demographic data were available from 26 of the 40 included therapists. Therapists were on average 35.15 years old (*SD* = 14.04). The majority identified as female (*n* = 17, 65.4%), with the remainder identifying as male (*n* = 7), or genderqueer (*n* = 1). The majority identified as white (*n* = 15, 57.7%), with the remainder identifying as Latinx (*n* = 4), Asian American (*n* = 3), African American (*n* = 2), Middle Eastern (*n* = 1), and multiracial (*n* = 1).

Measures

Therapeutic alliance. Therapeutic alliance was assessed using a previously validated (Imel, Hubbard, Rutter, & Simon, 2013) four-item version of the Working Alliance Inventory – Short Form Revised (Hatcher & Gillaspy, 2006) representing the bond, task, and goal dimensions of alliance. Items included "_____ and I are working towards mutually agreed upon goals" (goal), "I believe the way we are working on my problem is correct" (task), "I feel that "_____ appreciates me" (bond), and "_____ really understands me" (bond). Items were rated on a 1 (Never) to 7 (Always) scale. A total score was computed by averaging across the four items. Internal consistency reliability was adequate in the current sample ($\alpha = .90$). As noted above, ratings were made prior to each session (starting with the second session) asking clients to reflect back on their experience of alliance in the previous session. Although alliance can be rated from various perspectives (e.g., client, therapist, observer; Flückiger et al., 2018), the current study employed client-rated alliance due to its robust link with treatment outcome, ease of data collection, and ecological validity (i.e., the experience of alliance largely exists in the subjective experience of the client).

Data Analysis

For this study, we used 1,235 recorded sessions together with client-reported alliance, assessed prior to the subsequent session occurring between the same therapist and client. Audio recordings were processed through a speech pipeline to generate automatic speech-to-text transcriptions. The automatic speech recognition made use of the open-source, freely available Kaldi software (Povey et al., 2011). Components of the pipeline along with their corresponding accuracy (vs. human transcription) using data from the current study include: (a) a voice activity detector, where speech segments are detected over silence or noise (unweighted average recall = 82.7%); (b) a speaker diarization system, where the speech is clustered into speaker-

homogeneous groups (i.e., speaker A, speaker B) (diarization error rate = 6.4%); (c) a speaker role recognizer, where each group is assigned the label ‘therapist’ or ‘client’ (misclassification rate = 0.0%); and (d) an automatic speech recognizer, which transduces speech to text (word error rate = 36.43%). The modules of the speech pipeline have been adapted with the Kaldi speech recognition toolkit (Povey et al., 2011) using psychotherapy sessions provided by the same counseling center, but not used for the alliance prediction, thus not inducing bias. A similar system architecture is described in Xiao et al. (2016) and Flemotomos et al. (2019).

Linguistic features were extracted from resulting transcripts, independently for therapist and client text. We report results using unigrams and bigrams (i.e., one- and two-word pairings) weighted by the term frequency-inverse document frequency (tf-idf) (Salton & McGill, 1986) or sentence (Sent2vec) embeddings (Pagliardini, Gupta & Jaggi, 2018). Tf-idf weighting accounts for the frequency with which words appear within a given document (i.e., session), while also considering its frequency within the larger corpus of text (i.e.., all sessions). This allows less commonly used words (e.g., suicide) more weight than commonly used words (e.g., the). Thus, less common words are treated as more important. Tf-idf weighting was calculated across all sessions in the train set and applied to the test set. As described earlier, Sent2vec maps sentences to vectors of real numbers. Using Sent2vec, the session is represented as the mean of its sentence embeddings. Models used linear regression with L2-norm regularization (i.e., ridge regression, see Hoerl & Kennard, 1970), which is a method designed for highly correlated features, which is often the case for NLP data.

To estimate the performance of our method, experiments were run using a 10-fold cross-validation: data is split into ten parts, with nine parts used for training at each iteration (Train), and one for evaluation (Test). This is commonly used in ML and allows estimation of the extent

to which model results based on the training set (Train) will generalize to an independent sample (Test). Train and Test sets were constructed so as not to share therapists between them, as shared therapists could artificially inflate the model's accuracy. The algorithm is therefore expected to learn patterns of words related to alliance ratings in general instead of capitalizing on therapist-specific characteristics.

We employed two commonly used metrics of accuracy: mean squared error (MSE) and Spearman's rank correlation (*U*). These metrics reflect the accuracy of the ML algorithm when applied to the test set. Specifically, mean squared error is the average of the squared differences between the predictions and the true values and is useful for comparing models, though its absolute value is not interpretable. Spearman's rank correlation measures the strength of association between two variables, ranging from -1 to 1, with higher values preferred.

Computer Software

Self-report data were processed within the R statistical environment (R Core Team, 2018). NLP and ML was conducted using the Python programming language (Python Software Foundation, 2019). Models used the 'scikit-learn' toolkit (Pedregosa et al., 2011) and the 'sklearn.linear_model.Ridge' function (see Supplemental Materials Table 1 for syntax). Sent2Vec was implemented using the method developed by Pagliardini, Gupta, and Jaggi (2018) and N-grams obtained using the text feature extraction in 'scikit.'[1] The time required for running the speech pipeline and ML models can vary. In the current data, the speech pipeline required approximately 30 minutes per 50-minute session using one core of an AMD Opteron Processor

[1] Readers interested in working with text data in Python are encouraged to read the 'scikit' and Kaldi tutorials: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html; https://kaldi-asr.org/doc/kaldi_for_dummies.html

6276 (2.3 GHz). The 10-fold cross-validation models took approximately 10 minutes on a MacBook Pro with 2.8 GHz Intel Core i7, 16 GB RAM, and 2133 MHz LPDDR3.

Results

The sample included a total of 1,235 sessions with recordings and associated alliance ratings (provided at the subsequent session; n = 386 clients; 40 therapists). Clients had on average 3.20 sessions in the data set (SD = 2.50, range = 1 to 13) and therapists had 30.88 (SD = 32.97, range = 1 to 131). Sessions represented a variety of points in treatment, with a mean session number of 5.31 (SD = 3.37, range = 1 to 23). Across the 1,235 alliance ratings, the mean rating was 5.47 (SD = 0.83, median = 5.5, range = 1.75 to 6.50; Supplemental Materials Figure 1). Ratings showed the typical negative skew found in the assessment of alliance (Tryon, Blackwell, & Hammel, 2008).

ML model results are presented in Table 1. Models are shown using either therapist or client text as the input. Results are also separated by feature extraction method (tf-idf, Sent2vec). The baseline model reflects accuracy of the average rating (i.e., 5.47) and is useful to evaluate model performance.

The predictions of three out of the four models are significantly better than chance (Spearman´s $U > .00$, $p < .01$). The model that used therapist text and extracted features using tf-idf performed best overall, with MSE = 0.67 and $U = 0.15$, $p < .001$. For illustrative purposes only, we extracted the 15 unigrams/bigrams that were most positively or negatively correlated with alliance ratings in our best performing model. As these features represent only a small portion of the corresponding model, they should not be viewed as a replacement for the full model. The 15 most positively correlated unigrams/bigrams were: group, really, husband, right, think, phone, values, maybe, divorce, got, yeah, situation, um right, don think, max. The 15

most negatively correlated unigrams/bigrams were: counseling, yeah yeah, going, sure, coping, just want, friends, motivation, feeling, Monday, huh yeah, oh, physical, pretty, time.

Discussion

The current study introduces two related quantitative methods – NLP and ML – that have the potential to significantly expand methodological tools available to psychotherapy researchers and clinicians. The prediction of client-rated therapeutic alliance from session recordings was used as a test case for these methods due to the importance of alliance in psychotherapy and the potential contribution of technologies able to reliably automate alliance assessment. Results presented here suggest that ML models modestly predict alliance ratings ($U = .15$). That is to say, there was linguistic signal indicative of the strength of the alliance that is detectable through ML, supporting the notion that ML may be a useful tool for examining alliance in future studies.

It is worth contextualizing these results within the broader field of speech signal processing and NLP as well as prior work specifically within the domain of psychotherapy research. An important feature of the alliance, and part of the motivation to examine alliance, is its greater degree of abstraction from the actual linguistic context of a psychotherapy session. Compare alliance with another commonly studied psychotherapy process variable – motivational interviewing fidelity codes. Motivational interviewing codes are primarily linguistic in nature (e.g., open versus closed question; Miller et al., 2003) and can be reliably coded by trained human raters and ML-algorithms at approximately similar levels (e.g., $\kappa s > .75$ for use of open questions over a session of motivational interviewing; Atkins et al., 2014). Importantly, aspects of motivational interviewing fidelity that show lower inter-rater reliability among human raters (e.g., empathy) are also more difficult to predict via ML (e.g., $\kappa s \approx .25$ for talk turns and .00 for sessions; Atkins et al., 2014). Alliance, in contrast to most

motivational interviewing fidelity dimensions, requires in-depth processing by humans (i.e., client, therapist, or observer) and is presumably influenced by a variety of unobservable, non-linguistic factors. It is exactly this non-linguistic, internal processing that may be more difficult for ML models to replicate. This highlights a truism of NLP methodologies: behaviors more distal from linguistic content that are more difficult for human raters to rate reliably will also be more difficult for ML models to predict. This may make predicting even more abstracted aspects of treatment, such as treatment outcome, yet more challenging to predict using ML.

Practical Suggestions

Given these potential limitations, there are six practical considerations offered here that may increase the viability of ML to contribute to psychotherapy research. Several of these are fundamental principles of ML reviewed previously but are worth highlighting due to the possibility that many readers may not be familiar.

1. **ML may be most promising for predicting observable linguistic behaviors.** For efforts employing ML using text data, it may be valuable to start with observable behaviors that humans can code reliably using only text data (e.g., treatment fidelity; Atkins et al., 2014). Human reliability provides an estimate of the upper limit to reliability likely to be achieved using ML models. Behaviors for which humans have difficulty reaching consensus will likely be more challenging for ML models as well.
2. **ML models should be trained using human coding as the gold standard.** Related to the previous suggestion, it may be prudent to develop ML models based on behaviors that are observable and to use human-based ratings as the standard for training ML algorithms. Thankfully, promising observer-rated measures of alliance and other psychotherapy processes (e.g., empathy, treatment fidelity) have been developed that may

serve as a basis for future ML psychotherapy research. While this has been done in previous work on motivational interviewing (Atkins et al., 2014; Xiao et al., 2015), this was not used in the current study due both to resource limitations and an interest in attempting to predict client (rather than observer) ratings. However, ML models could be constructed predicting observer rated alliance which may be less prone to client response set biases (e.g., social desirability). While models using human coding as the basis are a promising starting point, it may also be useful to develop models attempting to predict more diffuse constructs that are not reliably rated by observers (e.g., treatment outcome).

3. **ML models should be tested using large data sets.** One of the distinct advantages of ML is its potential to process large amounts of data, an impractical task when using human coders. However, for the development of reliable ML algorithms, large amounts of training data are ideal. The actual amount of data necessary varies widely depending on the nature of the ML task, but data sets of 10,000 cases or more are commonly used in NLP applications. Given advances in NLP, researchers and clinicians who have access to high fidelity session recordings may be able to convert existing recordings to text data for ML models.
4. **Develop models using a training set and test models using a test set.** Similar to the rationale for employing separate sample for exploratory and confirmatory factor analysis (Gerbing & Hamilton, 1996), evaluation of ML algorithms requires separate samples. It is possible to get perfect accuracy within a training set, but this in no way indicates that results will be perfectly accurate in a future data set (i.e., for prediction). The need for separate samples echoes the need for large data sets when conducting ML.

5. **Develop interdisciplinary collaborations.** Most psychotherapy researchers are not trained in ML during graduate school. As these models depart in some important ways from traditional quantitative methods used in psychology (e.g., regression and analysis of variance), it may be vital for researchers interested in ML to build collaborations with colleagues more versed in the intricacies of ML. Researchers with expertise in processing linguistic data, with backgrounds in computer science and engineering, for example, may be ideal compliments to the clinical and context expertise brought by psychologists. Of course, interdisciplinary collaborations involve their own complexity, with researchers working across disciplinary cultures, practices, and standards.
6. **Have reasonable expectations and avoid the risk of alchemy.** A final suggestion is that those interested in pursuing ML-based psychotherapy research have reasonable expectations about the promise of these methods, and the speed with which they will become viable tools. One concern is that ML-based models simply replicate the human biases in the patient rated measures: if the model accurately learns the human rating, it will also include ceiling effects, social desirability, and other potentially construct irrelevant variance. In addition, it is encouraged that ML not be viewed as form of "alchemy" (Hutson, 2018) in which ML becomes a quasi-magical black box for researchers and consumers of research. ML research, like other research methodologies, is likely to benefit from transparency, humility, and replication (Open Science Collaboration, 2015) along with a healthy dose of skepticism.

Future Directions

Consistent with these practice suggestions, future work should continue to explore important psychotherapy process and outcome variables using linguistic, paralinguistic (e.g.,

prosody, pitch), and non-verbal therapy behaviors. Ideally this is done using large data sets (e.g., *N*s > 10,000 sessions). The current study focused on alliance, but future work could use similar methods to predict treatment outcome (e.g., Hamilton Rating Scale of Depression; Hamilton, 1960), multicultural competence (Tao, Owen, Pace, & Imel, 2015), empathy (Imel et al., 2014), interpersonal skill (Anderson et al., 2009), treatment fidelity (e.g., Cognitive Therapy Rating Scale; Creed et al., 2016; Goldberg et al., in press), and other variables previously assessed using observer ratings (e.g., innovative moments; Gonçalves, Ribeiro, Mendes, Matos, & Santos, 2011).

Development will also ideally occur in tandem with attention to measurement and known issues in psychotherapy research. For example, future work should consider likely bias in the measurement of alliance. Clients whose ratings are invariant across sessions (e.g., consistently provided alliance ratings at the ceiling of the measure) could be removed from ML models, perhaps eventually providing models that better predict the correlates of alliance (e.g., treatment retention) than self-report. Or ML models could be used to determine when collecting self-report alliance data would provide information beyond what analysis of session content could provide (e.g., models predicting discrepancies between ML-based and self-report alliance ratings). It also may be worthwhile attempting to predict therapist-level alliance scores using session content and ratings aggregated across multiple clients.

The current cross-validation design allowed no therapist to appear in both the Train and Test sets. Conceptually, this ML approach is trying to discover a universal model for mapping language to alliance, and as such, it is the hardest and most conservative modeling approach. Alternative strategies would allow therapists to be in both Train and Test sets, which allows a model to learn individual-specific mappings of text to alliance to support prediction of future

alliance scores for either therapist or client. It could be valuable to explore these additional models in future work.

Provided ML models continue to improve in their ability to detect important aspects of psychotherapy, questions of dissemination and implementation will become increasingly central. Many potentially valuable technologies have existed for years (e.g., models detecting depression symptoms via speech features; France et al., 2000), yet are not widely implemented. There are, of course, numerous reasons that innovations may not be adopted, and considerable scholarship focused on precisely this research-to-practice impasse (e.g., Wandersman et al., 2008). Part of the solution to bringing ML-based technologies to market may require researchers moving outside of the traditional academic boundaries and developing collaborations with industry. For clinicians and researchers alike, there may be discomfort with the notion of partnering with for-profit entities with fears of disruptions in objectivity that form the theoretical backbone of both science and practice (DeAngelis, 2000). While these concerns may be valid, these partnerships may play a central role in bringing novel technologies such as those based on ML to the therapists and clients who could benefit from them.

Gaining buy-in from clinicians is another dissemination and implementation barrier. Clinician discomfort discussed in relation to measurement-based care (e.g., Boswell, Kraus, Miller, & Lambert, 2015; Fortney et al., 2017; Goldberg et al., 2016) may very well be magnified when clinicians are asked to routinely record therapy sessions. Discomfort may be further magnified knowing that these recordings will subsequently be analyzed by a computer algorithm to determine treatment quality, therapeutic alliance, or outcome. Sensitivity to these and other dissemination and implementation issues will be crucial for moving this work forward.

A final future direction to mention is the importance of ultimately evaluating whether ML-based feedback – be it focused on alliance, fidelity, or any other aspect of treatment – actually provides benefits. The benefit of interest may depend on the stakeholder: for payers, this may involve demonstrating the quality of services; for clinicians, this may involve demonstrating improved client outcomes; for researchers, this may involve demonstrating reliability and validity with reduced cost of research team time and money. It is likely these metrics will ultimately determine whether ML can transform psychotherapy.

Limitations

While promising, the current study has several important limitations. The first is the relatively modest sample size. While large by human coding standards, the current number of sessions evaluating is well below the samples often used for ML. As noted previously, ML models improve with larger amounts of training data. Thus, the available sample size may have reduced the ability to predict alliance ratings from session recordings.

Another limitation is related to the available speech signal processing technology. In particular, existing NLP technologies have known limitations, including inaccuracy in transcription (i.e., misinterpreting spoken words) and errors in assigning speech to a given speaker (i.e., diarization). These factors introduce error variability into the text data which functions to reduce statistical power and the accuracy of the ML models.

A third key limitation is related to the assessment of alliance. For one, ratings were made retrospectively (i.e., about a prior session). Collecting ratings at time points more distant from the actual session may have reduced linkages between ratings and session content and thereby decreased the signal available for detection (i.e., exerting a conservative rather than liberal bias on our ability to predict alliance ratings from session content). Similarly, there was evidence that

alliance ratings in the current study suffered from range restriction due to the well-documented ceiling effects for ratings of alliance (Tryon et al., 2008). Range restriction also may have decreased statistical power and the ability to reliably predict alliance ratings (Cohen, Cohen, West, & Aiken, 2003). For this reason, it may be useful to examine alliance in other contexts in which ratings may be more variable (e.g., clients with more severe personality psychopathology). Lastly, alliance was assessed only by clients. While relevant and ecologically valid, accuracy may have been improved for predicting observer rated alliance in which observers and ML algorithms had access to the same information (i.e., session text).

Clinical Vignette

The algorithm developed in the current study is only a first attempt at predicting alliance ratings using ML, but these initial results suggest a potential future for using these technologies in clinical research and practice. We imagine a future application in the following vignette. This example indicates how machine learning-generated analytics derived directly from the session encounter can be used as another source of information for the therapist to reflect on their work and potentially improve the process of therapy.

Sandra is a 43-year old, married, African American, cisgender female who has been struggling with social anxiety since adolescence. She is a school librarian and the mother of two teenage sons. She has recently begun working with a psychologist, Dr. Martinez, due to "increasing stress and anxiety" at work which is beginning to spill-over into Sandra's family life. She reports she has trouble "asserting herself and expressing her needs" at home and at work.

During the intake session, Dr. Martinez shares with Sandra that the clinic has been using a recording platform that can provide Dr. Martinez with information about how therapy might be going, in particular feedback on the therapy "relationship." Sandra provides her consent for use

of the platform. Therapy starts out smoothly, with Sandra sharing more about the difficulties she is experiencing, which in recent months have included periodic panic attacks in social situations. Dr. Martinez, who primarily operates from a cognitive-behavioral therapy perspective, introduces exposure therapy as a treatment approach for reducing her symptoms.

During the fifth session, Dr. Martinez initiates a conversation about Sandra's progress in treatment. Sandra reports that therapy is going "just fine" and she apologies for not having had the time to complete the exposure exercises Dr. Martinez had recommended. Dr. Martinez reflects that she knows it can be challenging to make the time for engaging in therapy "homework" and that the exposures themselves can be unpleasant. Sandra quickly assures Dr. Martinez that she will try to do a better job making time for exposures.

Through the treatment, Dr. Martinez has been reviewing sessions and automated feedback on the quality of his relationship with Sandra and has noticed that the alliance scores generated by the system have been low in the past two sessions. Although Sandra indicated in session that treatment was going fine, the alliance algorithm was built using observer-rated alliance that is less contaminated with self-report biases (e.g., social desirability). Dr. Martinez uses this opportunity to discuss the automated feedback with Sandra, "You know Sandra, I was reviewing some feedback I received on our sessions last week, and it suggested that it might be smart for me to check in with you again on how things are going. I know you said, things are fine, but I can't help wonder if there's something I'm missing. I'd really like to know." At this point, Sandra notes that she has been having trouble with Dr. Martinez's therapeutic approach. Sandra shares that she has been having significant difficulties in her marriage recently and has experienced several racial microaggressions at work that have contributed to her anxiety. Sandra notes that she was hoping to discuss these events in therapy but was not sure how to bring them

up given Dr. Martinez's emphasis on exposure therapy and Sandra's difficulty completing her exposure exercises. Dr. Martinez expresses her appreciation to Sandra for sharing this. They begin a discussion of ways to refocus treatment to include these additional dimensions.

Conclusion

The current study introduced and attempted to model ML as a statistical approach that may be relevant for addressing important questions about psychotherapy. Just as ML is centrally involved in numerous cultural, technological, and social changes, it may also play a leading role in future innovation within psychotherapy research and practice. Our prediction of therapeutic alliance discussed here is one of several recent examinations of potential synergy between ML and psychotherapy. As available sample sizes grow and technology evolves, it may well be that ML algorithms can be developed to even more reliably detect treatment features like alliance from session recordings. Clearly such technologies could dramatically revolutionize training and provision of clinical services. In a way, these methods, while heavily reliant on computers and artificial intelligence, may prove crucial in helping human researchers and clinicians unravel the dizzying complexity of the human interaction that is psychotherapy.

References

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association of Computational Linguistics, 4*, 463-476.

Anderson, T., Ogles, B.M., Patterson, C.L., Lambert, M.J., & Vermeersch, D.A. (2009). Therapist effects: Facilitative interpersonal skills as a predictor of therapist success. *Journal of Clinical Psychology, 65*(7), 755-768. doi: 10.1002/jclp.20583

Atkins, D.C., Steyvers, M., Imel, Z.E., & Smyth, P. (2014). Scaling up the evaluation of

psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science, 9*(49). doi:10.1186/1748-5908-9-49

Baldwin, S.A., & Imel, Z.E. (2013). Therapist effects: Findings and methods. In M.J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change (6th ed.)* (p. 258-297). Hoboken, NJ: Wiley & Sons.

Baldwin, S.A., Wampold, B.E., & Imel, Z.E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology, 75*(6), 842-852. 852.

Benton, S.A., Robertson, J.M., Tseng, W.C., Newton, F.B., & Benton, S.L. (2003). Changes in counseling center client problems across 13 years. *Professional Psychology: Research and Practice, 34*(1), 66-72. doi: 10.1037/0735-7028.34.1.66

Berwian, I. M., Walter, H., Seifritz, E., & Huys, Q. J. (2017). Predicting relapse after antidepressant withdrawal: A systematic review. *Psychological Medicine, 47*(3), 426-437. doi:10.1017/S0033291716002580

Bibault, J. E., Giraud, P., & Burgun, A. (2016). Big data and machine learning in radiation oncology: state of the art and future prospects. *Cancer Letters, 382*(1), 110-117.

Bordin, E.S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research and Practice, 16*(3), 252-260.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19. doi:10.1080/10503307.2013.817696

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature, 559*(7715), 547-555.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access, 5*, 8869-8879.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences (3rd ed.).* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Creed, T. A., Frankel, S. A., German, R. E., Green, K. L., Jager-Hyman, S., Taylor, K. P., ... & Beck, A. T. (2016). Implementation of transdiagnostic cognitive therapy in community behavioral health: The Beck Community Initiative. *Journal of Consulting and Clinical Psychology, 84*(12), 1116. doi: 10.1037/ccp0000105

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge: Cambridge University Press.

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., & Reynolds III, C. F. (2014). Adding psychotherapy to antidepressant medication in depression and anxiety disorders: A meta- analysis. *World Psychiatry, 13*(1), 56-67.

DeAngelis, C. D. (2000). Conflict of interest and the public trust. *JAMA, 284*(17), 2237-2238.

Duncan, B.L., Miller, S.D., Sparks, J.A., Claud, D.A., Reynolds, L.R.,...& Johnson, L.D. (2003). The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of Brief Therapy, 3*(1), 3-12.

Dyson, F. J. (1998). *Imagined worlds (Vol. 6).* Cambridge, MA: Harvard University Press.

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy, 55*(4), 399-410.

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy, 49*(6-7), 373-378.

Falkenström, F., Granström, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *Journal of Counseling Psychology, 60*(3), 317-328. doi: 10.1037/a0032258

Flemotomos, N., Martinez, V., Chen, Z., Singla, K., Peri, R., Ardulov, V.,...& Narayanan S. (2019). *A speech and language pipeline for quality assessment of recorded psychotherapy sessions.* Manuscript in preparation.

Flűckiger, C., Del Re, A.C., Wampold, B.E., & Horvath, A.O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy, 55*(4), 316-340.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering, 47*(7), 829-837.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*(1), 62-72.

Goldberg, S.B., Babins-Wagner, R., Rousmaniere, T., Berzins, S., Hoyt, W.T., Whipple, J.L., Miller, S.D., & Wampold, B.E. (2016). Creating a climate for therapist improvement: A case study of an agency focused on outcomes and deliberate practice. *Psychotherapy*, 53(3), 367-375. doi: 10.1037/pst0000060

Goldberg, S. B., Baldwin, S. A., Merced, K., Caperton, D., Imel, Z. E., Atkins, D. C., & Creed, T. (in press). The structure of competence: Evaluating the factor structure of the Cognitive Therapy Rating Scale. *Behavior Therapy*. doi: 10.1016/j.beth.2019.05.008

Goldberg, S. B., Rowe, G., Ruan, H., Owen, J. J., & Miller, S. D. (2019). Routine outcome monitoring of therapeutic alliance to predict treatment engagement in a Veteran Affairs substance use disorders clinic. *Psychological Services.* doi: 10.1037/ser0000337

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Greenson, R. R. (1965). The working alliance and the transference neuro- sis. *Psychoanalytic Quarterly, 34*, 155–181.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Kim, R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402-2410.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 24*, 56-62.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research, 16*(1), 12-25.

Haykin, S. S. (2009). *Neural networks and learning machines* (3rd Ed.). Upper Saddle River, NJ: Pearson.

Hutson, M. (2018). Has artificial intelligence become alchemy? *Science, 360*(6388), 478.

Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Baer, J. S., Kircher, J. C., & Atkins, D. C. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology, 61*(1), 146-153.

Imel, Z. E., Caperton, D. D., Tanana, M., & Atkins, D. C. (2017). Technology-enhanced human interactions in psychotherapy. *Journal of Counseling Psychology, 64*(4), 385-393.

Imel, Z. E., Hubbard, R. A., Rutter, C. M., & Simon, G. (2013). Patient-rated alliance as a measure of therapist performance in two clinical settings. *Journal of Consulting and Clinical Psychology, 81*(1), 154-165. doi: 10.1037/a0030903

Imel, Z.E., Pace, B.T., Soma, C.S., Tanana, M., Gibson, J., Hirsch, T., Georgiou, P.G.,...& Atkins, D.A. (in press). Initial development and evaluation of an automated, interactive,

web-based therapist feedback system for motivational interviewing fidelity. *Psychotherapy.*

Imel, Z.E., Steyvers, M., & Atkins, D.C. (2015). Computation psychotherapy research: Scaling up the evaluation of patient-provider interactions. *Psychotherapy, 52*(1), 19-30.

Insel, T.R. (2017). Digital phenotyping: Technology for a new science of behavior. *JAMA, 318*(13), 1215-1216. doi:10.1001/jama.2017.11295

Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (in press). A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review.*

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (2nd Ed.). London: Pearson.

Lambert, M.J., Barley, D.E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory/Research/Practice/Training, 38*(4), 357-361.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203-1205.

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports, 8*(1), 7819. doi:10.1038/s41598-018-25953-0

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed

representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2003). *Manual for the Motivational Interviewing Skills Code v. 2.0.* Retrieved from http://casaa.unm.edu/codinginst.html.

Miner, A.S., Milstein, A., & Hancock, J.T. (2017). Talking to machines about personal mental health problems. *JAMA, 318*(13), 1217-1218. doi:10.1001/jama.2017.14151

Mitchell, T. M. (1997). Does machine learning really work? *AI Magazine, 18*(3), 11-20.

Mjolsness, E., & DeCoste, D. (2001). Machine learning for science: State of the art and future prospects. *Science, 293*(5537), 2051-2055.

Moore, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Biomedical Engineering, 55*(1), 96-107.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT press.

Okada, S., Ohtake, Y., Nakano, Y. I., Hayashi, Y., Huang, H. H., Takase, Y., & Nitta, K. (2016, October). Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 169-176). ACM.

Olfson, M., & Marcus, S.C. (2010). National trends in outpatient psychotherapy. *American Journal of Psychiatry, 167*(12), 1456-1463.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi: 10.1126/science.aac4716

Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised learning of sentence embeddings

using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Vol. 1, pp. 528-540).

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin, Austin, TX, Technical Report.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach (3rd ed.).* Essex, England: Pearson Education Limited.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval.* New York: McGaw-Hill, Inc.

Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine,* 1-23.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.

Stead, W. W. (2018). Clinical implications and challenges of artificial intelligence and deep learning. *JAMA, 320*(11), 1107-1108.

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484-498.

Substance Abuse and Mental Health Services Administration. (2014). *Projections of national expenditures for treatment of mental and substance use disorders, 2010–2020*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Tao, K. W., Owen, J., Pace, B. T., & Imel, Z. E. (2015). A meta-analysis of multicultural competencies and psychotherapy process and outcome. *Journal of Counseling Psychology, 62*(3), 337-350. doi: 10.1037/cou0000086

Thompson, M. N., Goldberg, S. B., & Nielsen, S. L. (2018). Patient financial distress and treatment outcomes in naturalistic psychotherapy. *Journal of Counseling Psychology, 65*(4), 523-530. doi: 10.1037/cou0000264

Tichenor, V., & Hill, C.E. (1989). A comparison of six measures of working alliance. *Psychotherapy, 26*(2), 195-199.

Tracey, T.J.G., Wampold, B.E., Lichtenberg, J.W., & Goodyear, R.K. (2014). Expertise in psychotherapy: An elusive goal? *American Psychologist, 69*(3), 218-229.

Tryon, G. S., Blackwell, S. C., & Hammel, E. F. (2008). The magnitude of client and therapist working alliance ratings. *Psychotherapy: Theory, Research, Practice, Training, 45*(4), 546-551. doi: 10.1037/a0014338

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460. Wampold, B., & Imel, Z.E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work (2nd ed.).* New York: Routledge.

Wandersman, A., Duffy, J., Flaspohler, P., Noonan, R., Lubell, K., Stillman, L., ... & Saul, J.

(2008). Bridging the gap between prevention research and practice: The interactive systems framework for dissemination and implementation. *American Journal of Community Psychology, 41*(3-4), 171-181.

Wang, R., Aung, M. S., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., ... & Tseng, V. W. (2016, September). CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 886-897). ACM.

Webb, C.A., DeRubeis, R.J., & Barber, J.P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 200-211. doi: 10.1037/a0018912

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., ... & Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet, 382*(9904), 1575-1586.

Xiao, B., Huang, C., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. S. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science, 2*, e59. doi: 10.7717/peerj-cs.59

Zilcha-Mano, S. (2017). Is the alliance really therapeutic? Revising this question in light of recent methodological advances. *American Psychologist, 72*(4), 311-325.

Zilcha-Mano, S., & Errázuriz, P. (2017). Early development of mechanisms of change as a predictor of subsequent change and treatment outcome: The case of working alliance. *Journal of Consulting and Clinical Psychology, 85*(5), 508-520.

Table 1. Results from machine learning prediction model

Model	Feature extraction method	MSE	*U*	*p*
Therapist	tf-idf	0.67	.15*	<.001
	Sent2Vec	3.34	.08*	.003
Client	tf-idf	0.69	.11*	<.001
	Sent2Vec	3.67	.01	.800
Baseline	Average	0.69	.00	n/a

Note: Therapist = therapist speech; Client = client speech; baseline = model results if model always predicts the mean alliance rating (i.e., 5.47); MSE = mean square error; *U* = Spearman's rank order correlation; tf-idf = term frequency-inverse document frequency weighting based on (inverse) frequency of occurrence within the document and larger corpus; Sent2vec = sentence embeddings used to map sentences to vectors of real numbers. Models employed unigrams and bigrams (i.e., one- and two-word pairings) and a linear regression with L2-norm regularization (i.e., ridge regression; Hoerl & Kennard, 1970). Models were evaluated using 10-fold cross-validation with nine parts used for model training and one used for evaluation.