Can a computer detect interpersonal skills?

Using machine learning to scale up the Facilitative Interpersonal Skills task

Simon B. Goldberg

University of Wisconsin-Madison

Michael Tanana

University of Utah

Zac E. Imel

University of Utah

David C. Atkins

University of Washington

Clara E. Hill

University of Maryland

Timothy Anderson

Ohio University

Authors Note: Simon B. Goldberg, Department of Counseling Psychology and Center for Healthy Minds, University of Wisconsin-Madison, Madison, WI, USA; Michael J. Tanana, College of Social Work, University of Utah, Salt Lake City, UT, USA; Zac E. Imel, Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA; David C. Atkins, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA; Clara E. Hill, Department of Psychology, University of Maryland, College Park, MD, USA; Timothy Anderson, Department of Psychology, Ohio University, Athens, OH, 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.

Disclosure statement: Drs. Tanana, Atkins, 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 details: This work was supported by the National Institutes of Health / National Institute on Alcohol Abuse and Alcoholism (NIAAA) under Grant R01/AA018673. Dr. Atkins time was supported in part by the National Institutes of Health / National Institute on Alcohol Abuse and Alcoholism (NIAAA) under Grant K02/AA023814. 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.

Recommended citation: Goldberg, S. B., Tanana, M., Imel, Z. E., Atkins, D. C., Hill, C. E., & Anderson, T. (in press). Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task. *Psychotherapy Research.*

Abstract

Objective: Therapist interpersonal skills are foundational to psychotherapy. However, assessment is labor intensive and infrequent. This study evaluated if machine learning (ML) tools can automatically assess therapist interpersonal skills. **Method:** Data were drawn from a previous study in which 164 undergraduate students (i.e., not clinical trainees) completed the Facilitative Interpersonal Skills (FIS) task. This task involves responding to video vignettes depicting interpersonally challenging moments in psychotherapy. Trained raters scored the responses. We used an elastic net model on top of a term frequency-inverse document frequency representation to predict FIS scores. **Results:** Models predicted FIS total and item-level scores above chance (*rho*s=.27-.53, *p*s<.001), achieving 31-60% of human reliability. Models explained 13-24% of the variance in FIS total and item-level scores on a held out set of data (R^2), with the exception of the two items most reliant on vocal cues (verbal fluency, emotional expression), for which models explained ≤1% of variance. **Conclusion:** ML may be a promising approach for automating assessment of constructs like interpersonal skill previously coded by humans. ML may perform best when the standardized stimuli limit the "space" of potential responses (vs. naturalistic psychotherapy) and when models have access to the same data available to raters (i.e., transcripts).

Keywords: machine learning; interpersonal skills; Facilitative Interpersonal Skills; therapist effects; artificial intelligence

Clinical or Methodological Significance of this Article

We predicted scores on an interpersonal skills performance task using words spoken by participants during the task. The models performed below human inter-rater reliability but generally well above chance, suggesting that machine learning may be a promising methodology for automating human coding of interpersonal skills and other psychotherapy-relevant constructs.

Machine learning (ML) has been 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). While the use of ML to investigate psychiatric conditions and their treatment is not new (e.g., Modai, Stoler, Inbarsaban, & Saban, 1993), ML has become increasingly visible in applied psychology and psychiatry in the past decade due to statistical and technological advances (Bzdok & Meyer-Lindenberg, 2018). This burgeoning interest in ML in psychology and psychiatry may be in part driven by concerns regarding the failing of scientific paradigms built upon null hypothesis significance testing (e.g., lack of reproducibility; Open Science Collaboration, 2015). ML methods are designed to provide optimally accurate and generalizable results (Dwyer, Falkai, & Koutsouleris, 2018).

As ML becomes more and more deeply embedded into our everyday lives (e.g., ML-based search engines, self-driving cars, and automated speech recognition software), the accumulating body of scientific work suggests these methods hold promise for advancing research in psychiatry (see Dwyer et al., [2018] for a review, including a basic introduction to ML). In particular, ML has shown promise for improving psychiatric diagnosis (e.g., differentiating individuals with bipolar vs. unipolar depression using brain imaging data; Redlich et al., 2014), assessment of prognosis (e.g., predicting functional outcomes following a first episode of psychosis using questionnaire data; Koutsouleris et al., 2016), and prediction of response to pharmacotherapy (e.g., of depression to antidepressants; Chekroud et al., 2016).

ML is increasingly visible in psychotherapy research as well (Figure 1). Recent studies have employed ML to address questions that have proven difficult to resolve relying on traditional statistical approaches. For example, Lutz et al. (2018) used network models and pre-treatment ecological momentary assessment (EMA) data paired with intake variables to predict

treatment dropout among patients with mood and anxiety disorders. LASSO logistic regression models were used to identify the most promising predictors. Their final model explained 32% of the variance in dropout, accurately identifying 81% of patients correctly. Of course, not all applications have been as successful. Rubel, Zilcha-Mano, Giesemann, Prinz, & Lutz (in press) were unable to predict individual patients' within-patient alliance-outcome association using baseline variables shown to moderate this association identified via ML. Rubel and colleagues used the Boruta ML algorithm which is based on random forests.

Other researchers have begun exploring the pairing of ML and natural language processing (NLP; Jurafsky & Martin, 2014) as new tools for conducting psychotherapy research. Given the primary role that language plays in psychotherapy, NLP is a compelling family of techniques. Goldberg et al. (in press) examined the utility of ML for detecting therapeutic alliance, demonstrating that session recordings could be used to predict patient-rated therapeutic alliance modestly above chance (Spearman's *rho* = .15). Goldberg and colleagues used ridge regression with term frequency-inverse document frequency (tf-idf) and sentence embeddings. ML has also been used to investigate therapist factors as well, with Althoff, Clark, and Leskovec (2016) using various ML techniques to identify linguistic features used by successful counselors in text message-based counseling conversations. Atkins, Steyvers, Imel, & Smyth (2014) used topic models to evaluate motivational interviewing fidelity, yielding accuracy similar to human raters (receiver operating curve [ROC] area under the curve [AUC] scores = 0.62 to 0.81).

The successful application of ML and NLP for assessing specific aspects of therapist behavior through evaluation of linguistic content (e.g., motivational interviewing fidelity; Atkins et al., 2014; Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016; Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015) highlights the possibility that time-intensive, human-based coding systems

used for psychotherapy training, quality monitoring, and research could, in the future, be automated using ML. A key objective for psychotherapy researchers working in this area is identifying tasks for which ML may be most promising. Results of recent applications of ML and NLP in psychotherapy provide some candidate answers. In particular, ML is likely to perform best for tasks that humans can perform effectively when models and humans have access to the same data.

As case in point, ML models were able to produce near human reliability for motivational interviewing fidelity assessment (Atkins et al., 2014). ML algorithms effectively learned the primarily language-based rules that govern these ratings (e.g., coding a question as open or closed; Tanana et al., 2016). In contrast, it is unsurprising that ML models had relatively modest performance predicting patient-rated therapeutic alliance from session content (Goldberg et al., in press). This is a task that humans find difficult as well (e.g., weak correlations between observer-rated and patient-rated alliance; Tichenor & Hill, 1989). When patients rate alliance, they process not only the data available to the ML algorithm (i.e., session transcripts), but also data invisible to the algorithm (e.g., their unspoken thoughts or feelings about the therapist and therapy). Thus, the ML algorithm does not have access to the same set of information as the human rater. Moreover, it may be that ML algorithms, like human raters, perform better when the rating task is primarily behavioral and therefore more objective (e.g., assessing cognitive rationale, assigned homework) versus abstract and therefore more subjective (e.g., facilitative conditions). For example, the ICC (3,8) for the Cognitive-Behavior Therapy Scale was .92, whereas it was .58 for the Facilitative Conditions Scale in Hill, O'Grady, and Elkin (1992).

The Facilitative Interpersonal Skills Task: A Promising Candidate for ML

The Facilitative Interpersonal Skills (FIS; Anderson, Ogles, Patterson, Lambert, & Vermeersch, 2009) task is a measure reliant on human coders that may be amenable to ML-based scoring. In theory, the FIS lies someplace between a fairly subjective task requiring substantial internal processing (e.g., patients rating alliance) and an objective task that is predominantly language based (e.g., coding an utterance as an open or closed question for motivational interviewing fidelity assessment). The FIS is a performance-based task designed to assess interpersonal skills relevant to the provision of psychotherapy (Anderson et al., 2009). These are termed *facilitative* to reflect their capacity for facilitating a collaborative relationship between therapist and patient. The task involves the participant responding to a series of video-based vignettes displaying interpersonally challenging patients (e.g., aggressive, avoidant) within the context of a psychotherapy session. Participants are asked to respond to the vignette as if they were the patient's therapist. Responses are then coded by humans based on an eight-item scoring scheme reflecting the degree to which the participant displayed various interpersonal skills in their response (e.g., verbal fluency, emotional expression; Anderson et al., 2019). The FIS has shown acceptable inter-rater reliability (ICCs > .70; Anderson et al., 2019).

In efforts to understand therapists' contribution to patient outcomes (i.e., therapist effects; Baldwin & Imel, 2013), the FIS has emerged as one of the most robust therapist-level predictors of outcomes (Heinonen & Nissen-Lie, in press). FIS scores have predicted therapists' outcomes in routine, outpatient psychotherapy (Anderson et al., 2009) and to prospectively predict outcomes for trainee therapists a year after the FIS assessment (Anderson, McClintock, Himawan, Song, & Patterson, 2016). Despite clear theoretical and empirical relevance of therapists' interpersonal skills for providing and studying psychotherapy (Heinonen & Nissen-Lie, in press; Schöttke, Flückiger, Goldberg, Eversmann, & Lange, 2017), use of the FIS has as

yet not been widespread. As with many resource-intensive methodologies, this may in part be due to the specialized training required for scoring the FIS along with the time required for conducting human coding. However, a scalable FIS scoring methodology could make this assessment tool available for the variety of stakeholders interested in assessing this outcome-relevant therapist variable (e.g., psychotherapy researchers, graduate admissions committees, clinical supervisors, clinic hiring committees).

The FIS may be a prime candidate for ML-based scoring for two key reasons. First, similar to the assessment of motivational interviewing fidelity, the algorithm would have access to the same information provided to humans (i.e., verbal responses to FIS vignettes). Second, similar to the use of standardized patients (e.g., for assessing clinical skills in medicine; Barrows, 1993), the nature of the task limits the "space" of potential responses. All participants respond to the same video, thereby reducing variance in the stimuli relative to a naturalistic context (e.g., outpatient psychotherapy; Goldberg et al., in press).

The Current Study

The current study examined the use of ML to automate FIS scoring using FIS ratings drawn from a previous study ([omitted for masked review]). Model inputs included only the transcript of FIS responses paired with human-coded FIS ratings. As the ML algorithm did not have access to the vocal and nonverbal cues available to the human rater, this approach theoretically provides a more conservative test of the potential of ML.

Method

Participants and Setting

FIS task data were drawn from a previous study ([omitted for masked review]) that investigated the effectiveness of a helping skills training for undergraduate students delivered

through a semester-long, 4-credit psychology laboratory course at a large, public, Mid-Atlantic university. The original sample included 191 participants (141 female, 50 male; 115 European American, 28 Asian, 20 multi-ethnic or other, 19 African American, nine Latino, mean age = 21.35, *SD* = 2.00 years; see [omitted for masked review] for additional demographics). The majority of participants (56%) had previously taken introductory courses in counseling and/or clinical psychology, although they did not have formal training in psychotherapy. Participation was voluntary and confidential; students received extra credit for participating. Participants completed the FIS at the beginning of the semester.

FIS responses were video recorded and coded by a team of 11 trained raters (three male, eight female; 10 European American, one African American). The coding team always included 8 individuals, although 11 were involved over the course of the original study due to rater turnover. Videos were coded by non-overlapping teams of at least 3-4 raters. Four clinical psychology doctoral students served as co-leaders of the coding teams; seven undergraduate students served as coding team members. FIS ratings were available for 164 participants. The procedure required participants to attend a separate, individual session for the FIS task where their verbal responses to the stimulus clips were recorded in a private room and 27 of the 191 participants could not be scheduled. FIS responses were transcribed for use in ML models.

Measures

The FIS performance task (Anderson et al., 2009) involves participants responding to seven brief video clips (i.e., approximately one minute each) depicting interpersonally challenging moments in therapy (i.e., patients referring directly to troubling aspects of the therapist-patient relationship). Vignettes include patients reflecting a range of interpersonal

styles based on the interpersonal circumplex (e.g., being overly friendly, hostile, or submissive; Leary, 1957). Each video includes an actor reenacting transcripts from actual therapy sessions.

Participants completed the assessment individually. After viewing each clip, they were instructed to respond as if they were the therapist for the particular patient. Participant responses were video recorded for analysis.

Video-recorded responses were coded by the rating teams on eight standard FIS items (Anderson et al., 2019): (a) verbal fluency, (b) hope and positive expectations, (c) persuasiveness, (d) emotional expression, (e) warmth, acceptance, and understanding, (f) empathy, (g) alliance-bond capacity, (h) alliance rupture-repair responsiveness. Ratings were made on a 5-point Likert-type scale ranging from 1 (skill deficits) to 5 (optimal presence of skill). The items provided operational definitions for subjective judgments about interpersonal qualities and raters were trained in how to use the manual through calibrated examples of the levels of the scale. Weekly rater meetings were held over the course of an academic year. These meetings included discussions of ratings of divergent codes (+/- 1 point) and selected examples. Raters were instructed to use "3" as a baseline and increase or decrease the rating based on the presence or absence of skills. An overall FIS total score was computed by taking the average of the eight items. Inter-rater reliability was high in the current sample (ICCs = .91 to .95 within the six teams of raters). FIS ratings were averaged across all coders for each item for a given participant. Internal consistency reliability for the total score was high (α = .97).

Data Analysis

Transcripts of each participant's FIS responses were paired with FIS total and item-level scores. ML models were constructed predicting either the total or item-level scores. Inclusion of item-level scores allowed assessment of potential variation in ML model performance across FIS

subdomains. ML models used unigrams (i.e., appearance of single words as predictors) that were weighted using term frequency-inverse document frequency (tf-idf; Salton & McGill, 1986). Tf-idf weighting accounts for how frequently a given word appears within a given document (i.e., within a given participant's set of FIS responses), while simultaneously considering its frequency within the larger text corpus (i.e., across all participants' FIS responses). Tf-idf allows less commonly used words (e.g., rupture) more weight than commonly used words (e.g., the). In essence, this allows less common words to be treated as more important within the models. These word-level weights were entered into a regularized linear regression (which imposes a penalty on the size of coefficients to avoid overfitting a training dataset; Tibshirani, 1996). The regularized regression used both L1 and L2 regularizers, also known as elastic net regression (Tibshirani, 1996).

Following typical ML practice, cross-validation was used to assess prediction accuracy (Dwyer et al., 2018). Specifically, 10-fold cross-validation was used, in which models are iteratively fitted to 90% of the available data (i.e., training set) and then evaluated on the 10% not included in the model training (i.e., test set). Prediction on the test set was evaluated using R^2 and Spearman's rank order correlation (*rho*), which provides unbiased estimates of association even in the presence of skewed, non-normal distributions (Cohen, Cohen, West, & Aiken, 2003).

Results

Descriptive statistics for the FIS total and item-level scores are provided in Table 1. Consistent with scoring guidelines (i.e., starting with a score of 3 and increasing or decreasing based on response content), the mean FIS total and item-level scores were close to 3.00 (2.53 to 3.18). The range was somewhat restricted (all ranges < 3.0). Item-level ICCs ranged from .86 (Empathy) to .90 (Emotional expression, Alliance bond capacity).

ML model performance results are shown in Table 2. Spearman's *rho* values ranged from .27 (verbal fluency) to .53 (hope and positive expectations), all indicating prediction of FIS scores above chance (*p*s < .001). These *rho* values indicate that the ML models were able to achieve 52% of human reliability on the total score and between 31 and 60% at the item level. R^2 values ranged from .00 to .24, indicating that the model performed poorly when predicting verbal fluency and emotional expression (R^2s = .01 and .00, respectively). Relative to traditional R^2 in psychology research in which there is no separation of training data and test data, the cross-validated R^2 provides a more conservative estimate of model performance as values can be negative.

The ten unigrams most positively and the ten unigrams most negatively associated with FIS total scores were extracted from the model for descriptive purposes only. As the models include many more predictors than the ten unigrams listed in each category, the unigrams should not be interpreted in isolation. The top ten positively correlated unigrams were: "specific", "insight", "plan", "behavior", "let's", "end", "something", "responsibility", "explore", and "into". The top ten negatively correlated unigrams were: "perhaps", "guess", "though", "transition", "never", "today", "schedule", "pressure", "responsive", and "head".

Discussion

We evaluated the possibility of automating the scoring of a performance task designed to assess interpersonal skills (the FIS) using ML. On the whole, results were promising. Using a relatively modest sample of transcripts (modest by ML standards; Chekroud et al., 2016) and associated FIS scores (*n* = 164), models predicted FIS total and item-level scores above chance. Associated effect sizes were generally in the moderate to large range (Cohen, 1988). For the FIS total score, models predicted 19% of variance in scores (i.e., cross-validated R^2 = .19), with an

associated Spearman's *rho* = .48 representing 52% of human reliability. Moreover, as would be expected, the model performed worst when predicting the domains that rely explicitly on vocal components of the response (e.g., prosody, fluency), to which the ML models did not have access ($R^2 \leq 1\%$ for verbal fluency and emotional expression scores). Rating instructions for verbal fluency and emotional expression items explicitly reference vocal qualities (e.g., "the verbal quality of the response may have a 'melodic,' rhythmical quality" and "there is affect and prosody in the participant's voice," for verbal fluency and emotional expression, respectively; Anderson et al., 2019, pp. 17, 20).

It is worth considering these results within the context of other recent applications of ML and NLP in psychotherapy research. One relevant point of comparison is the study conducted by Goldberg et al. (in press) that found ML models were only modestly able to predict patient-rated alliance from psychotherapy session content (*rho* = .15). The effect size obtained in the current study (*rho* = .48 for FIS total score) indicates markedly improved performance. As noted previously, there are several plausible explanations for this discrepancy. For one, in the current study, both the ML models and the FIS raters had access to most of the same material (with the exception of the vocal and nonverbal behavior). Therefore, the ML models were able to use relevant information from the FIS transcripts in order to "learn" to mimic the human coding. In contrast, the patients in Goldberg et al.'s study presumably used large amounts of information for determining their alliance ratings that were not available to the algorithm (e.g., idiosyncratic thoughts and feelings associated with the particular therapist, session, moment in time). A second key feature was that the FIS task constrained the linguistic space by providing each participant with the same prompts (i.e., video vignettes) to which to respond. These uniform stimuli are in stark contrast, of course, with naturalistic psychotherapy, a context that is more likely to contain

much wider linguistic variability. In principle, the constrained linguistic space makes it more likely that linguistic variation (e.g., in the FIS responses) provides signals relevant for predicting scores (e.g., FIS ratings).

Model performance in the current study more closely replicates the successful applications of ML to predict motivational interviewing fidelity. Indeed, similar to FIS ratings, motivational interviewing fidelity is based largely on linguistic context that is available to both human raters and ML algorithms. In a recent study, Xiao et al. (2015) predicted human empathy ratings made using the Motivational Interviewing Treatment Integrity 3.0 (MITI3.0; Moyers, Martin, & Manuel, 2005) from ML models based on 200 motivational interviewing sessions. The best performing model showed a correlation of $r = .71$ with human coded empathy. Importantly, this value is similar to the inter-rater reliability of empathy ratings made by observers in the study (ICC = .60, *Kappa* = .74). This performance on par with human inter-rater reliability is consequential as human reliability theoretically provides an upper bound on the accuracy of ML models (i.e., ML models are unlikely to perform more accurately than humans are able to agree).

The promising results in the current study imply several potentially fruitful future directions. Based on performance in the current sample, this particular FIS scoring algorithm is not ready for widespread use. Performance was below recommended reliability cut-offs and below human inter-rater reliability for the FIS (i.e., < .70; Anderson et al., 2019). Nonetheless, this proof-of-concept initial attempt suggests it is reasonable to expect that ML models may be able to provide near-human performance in the future, if provided the appropriate input. Given FIS raters typically have access to audiovisual information, future modeling could use lexical (i.e., text), paralinguistic (i.e., prosody), and visual information. This would allow the ML algorithm access to all the same data available to human raters. However, it is not guaranteed

that the model will improve with paralinguistic or visual information included. Some ML applications perform better when provided with a smaller amount of less noisy data. Another potentially promising avenue would be examining the model's ability to differentiate between high and low FIS scores (i.e., extreme groups approach; Preacher, Rucker, MacCallum, & Nicewander, 2005). ML models often show higher performance when predicting scores drawn from extreme groups (e.g., Xiao et al., 2015). Another potential route to improve model performance is by providing a larger number of responses per participant (i.e., more than eight vignettes). Just as adding items tends to improve internal consistency reliability (Crocker & Algina, 2008), a greater number of behavioral samples can improve inter-rater reliability. Perhaps most importantly, future studies should use larger samples. It is very likely that performance will improve simply by having access to more data. Akin to human learning, ML models tend to improve with more data from which to learn (Hastie, Tibshirani, & Friedman, 2017). Finally, even if future models do not outperform the current model, the possibility of scaling up FIS assessment and implementing this tool in real world settings may outweigh the limitations of attenuated reliability.

Several limitations are important to note. While the study included a reasonably large sample for some null hypothesis testing purposes (e.g., detecting correlations or mean differences; Cohen, 1992), model performance was likely reduced due to the sample size. A second key limitation was that participants completing the FIS were drawn from an undergraduate convenience sample and were not in fact psychotherapists; this may limit generalizability to actual therapists. Consistent with this possibility, FIS total scores in the current sample were below those from a previous study that assessed therapists-in-training at the beginning of graduate school (means = 2.96 vs. 3.40, respectively, $d = 0.93$; Anderson et al.,

2016). The relatively low representation of racial/ethnic minorities among participants (60% white) and a restricted age range leaves open the question of whether the current model would perform similarly well in a more heterogeneous sample. Although the models showed some promise, the limited linguistic space of the FIS task may have contributed to this; future research should examine the degree to which similar models can reliably replicate human ratings (e.g., of FIS constructs, observer-rated therapeutic alliance, etc.) drawn from naturalistic psychotherapy. Lastly, the models did not employ paralinguistic and visual data that were available to human raters, potentially limiting performance. This also leaves open the question of the degree to which ML model performance is related to models having access to precisely the same data as human raters, or whether certain elements of the data (e.g., linguistic content) may be most important.

We report a first attempt at detecting psychotherapy-relevant interpersonal skills, operationalized as performance on the FIS, using ML. While results are encouraging, there is substantial room for improvement. Nonetheless, ML is a highly promising avenue for automating the assessment of key process and outcome variables in psychotherapy that have previously relied on time-intensive human coding (e.g., observer-rated alliance, clinician-rated depression, innovative moments; Gonçalves, Ribeiro, Mendes, Matos, & Santos, 2011; Hamilton, 1960; Tichenor & Hill, 1989). Continued development of these technologies may provide the new tools for discovering "new things that have to be explained" (Dyson, 1998, p. 51) in psychotherapy research and ultimately for improving the efficacy and efficiency of 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

Anderson, T., McClintock, A. S., Himawan, L., Song, X., & Patterson, C. L. (2016). A prospective study of therapist facilitative interpersonal skills as a predictor of treatment outcome. *Journal of Consulting and Clinical Psychology, 84*(1), 57-66. doi: 10.1037/ccp0000060

Anderson, T., Patterson, C., McClintock, A. S., McCarrick, S. M., Song, X., & The Psychotherapy and Interpersonal Lab Team (2019). *Facilitative Interpersonal Skills Task and Rating Manual.* Ohio University, Athens, Ohio.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage, 145*, 137-165. doi: 10.1016/j.neuroimage.2016.02.079

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.

Barrows, H. S. (1993). An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine, 68,* 443-443.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223-230.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H.,...& Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry, 3*(3), 243-250. doi: 10.1016/ S2215-0366(15)00471-X

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

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.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91-118.

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

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M., Kuo, P., Pace, B. T.,…& Atkins, D. C. (in press). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*.

Gonçalves, M. M., Ribeiro, A. P., Mendes, I., Matos, M., & Santos, A. (2011). Tracking novelties in psychotherapy process research: The innovative moments coding system. *Psychotherapy Research, 21*(5), 497-509.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction. (2nd ed.).* New York: Springer.

Heinonen, E., & Nissen-Lie, H. A. (in press). The professional and personal characteristics of effective psychotherapists: a systematic review. *Psychotherapy Research*. doi: 10.1080/10503307.2019.1620366

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology, 60*(1), 73-79. doi: 10.1037/0022-006X.60.1.73

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., ... & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry, 3*(10), 935-946. doi: 10.1016/ S2215-0366(16)30171-7

Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald Press.

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

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

Modai, I., Stoler, M., Inbar-Saban, N., & Saban, N. (1993). Clinical decisions for psychiatric inpatients and their evaluation by a trained neural network. *Methods of Information in Medicine, 32*(05), 396-399.

Moyers, T.B., Martin, T., & Manuel, J. (2005). *Motivational Interviewing Treatment Integrity (MITI) coding system.* The University of New Mexico Center on Alcoholism. Retrieved from: http://casaa-0031.unm/edu

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

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178-192. doi: 10.1037/1082-989X.10.2.178

Redlich, R., Almeida, J. R., Grotegerd, D., Opel, N., Kugel, H., Heindel, W., ... & Dannlowski, U. (2014). Brain morphometric biomarkers distinguishing unipolar and bipolar depression: a voxel-based morphometry–pattern classification approach. *JAMA Psychiatry, 71*(11), 1222-1230. doi:10.1001/jamapsychiatry.2014.1100

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., & Lutz, W. (in press). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches: A demonstration. *Psychotherapy Research*.

doi: 10.1080/10503307.2019.1597994

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

Schöttke, H., Flückiger, C., Goldberg, S.B., Eversmann, J., & Lange, J. (2017). Predicting psychotherapy outcome based on therapist interpersonal skills: A five-year longitudinal study of a therapist assessment protocol. *Psychotherapy Research, 27*(6), 642-652. doi:10.1080/10503307.2015.1125546

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43-50. doi: 10.1016/j.jsat.2016.01.006

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267-288.

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

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). " Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PloS ONE, 10*(12), e0143055.

Table 1. Descriptive statistics for FIS scores

FIS domain	Mean	*SD*	Min	Max	ICC
Verbal fluency	3.18	0.50	1.55	4.36	.87
Hope and positive expectations	3.11	0.36	2.21	4.2	.88
Persuasiveness	2.97	0.52	1.46	4.34	.87
Emotional expression	3.11	0.53	2.04	4.38	.90
Warmth, acceptance, and understanding	3.08	0.46	1.86	4.15	.89
Empathy	2.78	0.47	1.71	4.36	.86
Alliance bond capacity	2.95	0.45	1.95	4.04	.90
Alliance rupture-repair responsiveness	2.53	0.50	1.38	3.84	.87
Total score	2.96	0.43	1.99	4.21	.93

Note: n = 164; FIS = Facilitative Interpersonal Skills task (Anderson et al., 2009); *SD* = standard deviation; Min = minimum value; Max = maximum value; ICC = intraclass correlation (ICC[3,8]).

Table 2. Results of machine learning models predicting FIS scores from transcripts

FIS domain	R_2	Spearman's *rho*	*p*	% human ICC
Verbal fluency	.01	.27	< .001	.31
Hope and positive expectations	.24	.53	< .001	.60
Persuasiveness	.18	.47	< .001	.54
Emotional expression	.00	.31	< .001	.34
Warmth, acceptance, and understanding	.16	.46	< .001	.52
Empathy	.18	.47	< .001	.55
Alliance bond capacity	.15	.47	< .001	.52
Alliance rupture-repair responsiveness	.13	.47	< .001	.54
Total score	.19	.48	< .001	.52

Note: FIS = Facilitative Interpersonal Skills task (Anderson et al., 2009); R_2 = cross-validated R_2 that allows for negative values; *p* = *p*-value for Spearman's *rho*; % human ICC = machine learning model performance relative to human coders' intraclass correlation (ICC[3,8]).

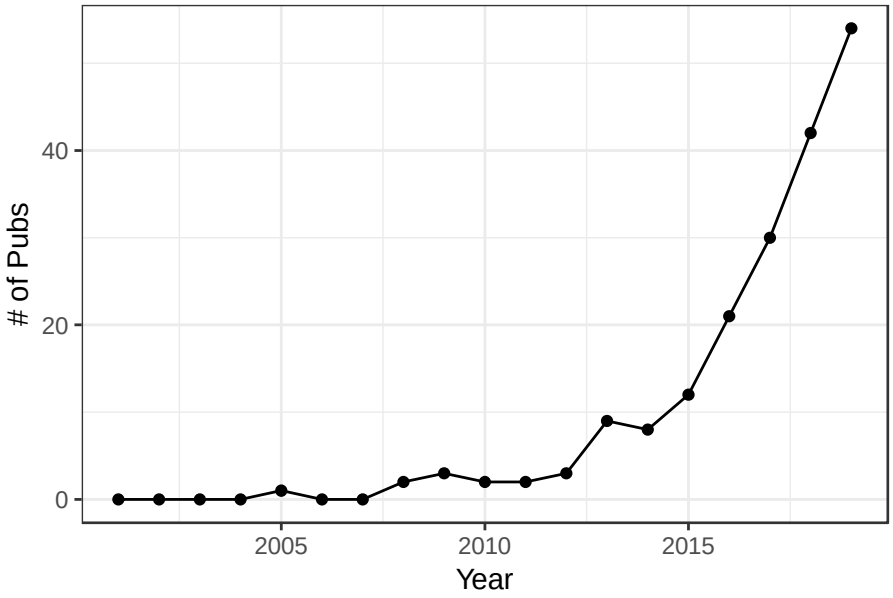

Figure 1. Publications indexed in PubMed using the search terms “machine learning” and “psychotherapy” from 2001 to 2019.

Supplemental Materials Table 1. Sample Python code for machine learning models

This appendix details some of the key implementation decisions for our machine learning models to facilitate replication. It is intended to capture all of the key featurization and modeling decisions. All models were run in Python 3.5 and used the 'sklearn' machine learning packages along with the 'pandas' data model.

First, we preprocessed the text into unigram features, with all non-ascii characters removed. The TF-IDF processor used was from the sklearn package (note- lowercase is set to false since we have already converted the data to lowercase at this step):

```
from sklearn.feature_extraction.text import TfidfVectorizer
def ident_func(x):
    return x;
TfidfVectorizer(tokenizer=ident_func, preprocessor=None, lowercase=False)
```

The 'elastinet' model used the following configuration:

```
from sklearn import linear_model
linear_model.ElasticNet(alpha = 0.0005, l1_ratio=.5)
```

We assessed the model performance using metrics from sklearn and scipy:

```
from scipy.stats import spearmanr
from sklearn import metrics
```

```
rho, p_val = spearmanr(y_test, predictions)

r2 = metrics.r2_score(y_test, predictions)
```