Psychosocial Intervention (2021) 30(3) 139-153

Psychosocial Intervention

https://journals.copmadrid.org/pi

A Systematic Review of Machine Learning for Assessment and Feedback of Treatment Fidelity

Asghar Ahmadi[a], Michael Noetel[a, b], Melissa Schellekens[a], Philip Parker[a], Devan Antczak[a], Mark Beauchamp[c], Theresa Dicke[a], Carmel Diezmann[d], Anthony Maeder[e], Nikos Ntoumanis[f, g, h], Alexander Yeung[a], and Chris Lonsdale[a]

[a]*Institute for Positive Psychology and Education, Australian Catholic University, Australia;* [b]*School of Health and Behavioural Sciences, Australian Catholic University, Australia;* [c]*School of Kinesiology, University of British Columbia, Canada;* [d]*Faculty of Education, Queensland University of Technology, Australia;* [e]*College of Nursing and Health Sciences, Flinders University, Australia;* [f]*Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Denmark;* [g]*School of Psychology, Curtin University, Australia;* [h]*Halmstad University, Sweden*

ARTICLE INFO

Received 6 December 2020
Accepted 13 May 2021
Available online 21 June 2021

Keywords:
Machine learning
Treatment fidelity
Treatment integrity
Clinical supervision
Feedback

A B S T R A C T

Many psychological treatments have been shown to be cost-effective and efficacious, as long as they are implemented faithfully. Assessing fidelity and providing feedback is expensive and time-consuming. Machine learning has been used to assess treatment fidelity, but the reliability and generalisability is unclear. We collated and critiqued all implementations of machine learning to assess the verbal behaviour of all helping professionals, with particular emphasis on treatment fidelity for therapists. We conducted searches using nine electronic databases for automated approaches of coding verbal behaviour in therapy and similar contexts. We completed screening, extraction, and quality assessment in duplicate. Fifty-two studies met our inclusion criteria (65.3% in psychotherapy). Automated coding methods performed better than chance, and some methods showed near human-level performance; performance tended to be better with larger data sets, a smaller number of codes, conceptually simple codes, and when predicting session-level ratings than utterance-level ones. Few studies adhered to best-practice machine learning guidelines. Machine learning demonstrated promising results, particularly where there are large, annotated datasets and a modest number of concrete features to code. These methods are novel, cost-effective, scalable ways of assessing fidelity and providing therapists with individualised, prompt, and objective feedback.

Revisión sistemática del aprendizaje automático para la evaluación y *feedback* de la fidelidad al tratamiento

Palabras clave:
Aprendizaje automático
Fidelidad al tratamiento
Integridad del tratamiento
Supervisión clínica
Feedback

R E S U M E N

Se ha puesto de manifiesto que muchos tratamientos psicológicos tienen un coste efectivo y son eficaces siempre que se apliquen con fidelidad. La evaluación de esta y el *feedback* son caros y exigen mucho tiempo. El aprendizaje automático se ha utilizado para evaluar la fidelidad al tratamiento, aunque su fiabilidad y capacidad de generalización no estén claras. Recopilamos y analizamos todas las aplicaciones de aprendizaje automático con el fin de evaluar el comportamiento verbal de todos los profesionales de ayuda, con el acento particular en la fidelidad al tratamiento de los terapeutas. Llevamos a cabo búsquedas en nueve bases de datos electrónicas para enfoques automáticos de codificación de comportamiento verbal en terapia y contextos semejantes. Llevamos a cabo el cribado, la extracción y la evaluación de la calidad por duplicado. Cincuenta y dos estudios cumplían nuestros criterios de inclusión (el 65.3% en psicoterapia). Los métodos de codificación automática resultaban mejor que el azar y algunos de ellos mostraban un desempeño casi al nivel humano, que tendía a ser mejor con conjuntos más grandes de datos, un número de códigos menor, códigos conceptualmente simples y cuando predecían índices al nivel de sesión que los de tipo declaración. Escasos estudios cumplían las directrices de buena praxis en aprendizaje automático. Este presentó unos resultados alentadores, sobre todo donde había conjuntos de datos grandes y anotados y un escaso número de características concretas que codificar, modos expansibles de evaluar la fidelidad y facilitar a los terapeutas un *feedback* individualizado, rápido y objetivo.

Cite this article as: Ahmadi, A., Noetel, M., Schellekens, M., Parker, P., Antczak, D., Beauchamp, M., Dicke, T., Diezmann, C., Maeder, A., Ntoumanis, N., Yeung, A., & Lonsdale, C. (2021). A systematic review of machine learning for assessment and feedback of treatment fidelity. *Psychosocial Intervention, 30*(3), 139-153. https://doi.org/10.5093/pi2021a4

Corresponding author: michael.noetel@acu.edu.au (M. Noetel)

When implemented faithfully, psychological treatments are powerful (Barth et al., 2013; Blanck et al., 2018; Kazdin, 2017; Öst & Ollendick, 2017). But, a major problem with both researching and implementing psychological treatments is fidelity (Bellg et al., 2004; Perepletchikova & Kazdin, 2005). Ensuring that treatments are implemented faithfully is important for a few reasons. First, when training practitioners on evidence-based interventions, prompt clinician feedback can facilitate skill acquisition and faithful implementation (Prowse & Nagel, 2015; Prowse et al., 2015). Second, without assessing fidelity we cannot determine whether effects from intervention studies are due to a homogenous treatment (Prowse & Nagel, 2015; Prowse et al., 2015). However, treatment fidelity is rarely well assessed—fewer than 10% of studies adequately assess fidelity (Perepletchikova & Kazdin, 2005; Perepletchikova et al., 2007). Cost and time are significant barriers (Borrelli, 2011). In psychotherapy, technology has become a well-established method of reducing costs of treatment by creating, for example, online interventions (Fairburn & Patel, 2017; Kazdin, 2017). But, the use of technologies for assessment and training is comparatively nascent (Fairburn & Cooper, 2011; Fairburn & Patel, 2017). This paper presents a systematic review of machine learning strategies to assess the fidelity of psychological treatments.

Fidelity encompasses three core components: adherence, differentiation, and competence (Rodriguez-Quintana & Lewis, 2018). Adherence describes a therapist's use of methods proposed by the guiding framework (e.g., using cognitive defusion while delivering Acceptance and Commitment Therapy). Differentiation is the avoidance of methods not proposed by that theory (e.g., using thought stopping while delivering Acceptance and Commitment Therapy). Competence is the skill with which the therapist implements the intervention (e.g., demonstrating a strong therapeutic alliance; Kazantzis, 2003). As a result, treatment fidelity is important both in the content and the process of therapy. Many interventions, like Motivational Interviewing and Cognitive Behaviour Therapy, both prescribe the content of therapy (e.g., change-talk and cognitive challenging, respectively) and the process of therapy (e.g., both emphasise the importance of an empathic therapeutic alliance; Kazantzis, 2003; Madson et al., 2009). From a content perspective, it is common for therapists to drift away from the core, evidence-based foci of therapy (Bellg et al., 2004; Waller, 2009; Waller & Turner, 2016). They may fail to use interventions that faithfully incorporate the therapy (low adherence) or 'dabble' in interventions from other therapies (low differentiation). But fidelity can also refer to the non-judgemental, compassionate, empathic process that is central to many therapies. As such, quality interpersonal interactions are critical for competent treatment (Kornhaber et al., 2016). Psychologists that competently demonstrate evidence-based interpersonal skills are more effective at reducing maladaptive behaviours such as substance abuse and risky behaviours than clinicians with poorer skills (e.g., Parsons et al., 2005). Their clients are more likely to complete treatment and change behaviour too (Golin et al., 2002; Moyers, Miller, et al., 2005; Street et al., 2009).

As a result, researchers have developed a range of treatment integrity measures (Rodriguez-Quintana & Lewis, 2018), including many that assess the content of therapy (McGlinchey & Dobson, 2003) and the process of therapy (e.g., Motivational Interviewing Skill Code: Miller et al., 2003; Motivational Interviewing Treatment Integrity: Moyers, Martin, et al., 2005). There are even measures for assessing how well treatment fidelity is assessed (Perepletchikova et al., 2009). These measures improve the quality of research and the translation of evidence-based therapies into practice (Prowse & Nagel, 2015; Prowse et al., 2015). The most objective of these measures involve an observer rating the behaviours of the therapist at regular intervals or after having watched an entire session with a client. As a result, assessing fidelity requires significant resources (Fairburn & Cooper, 2011). Recently, researchers have begun applying machine learning models to automate this task. These models will not be useful if they fail to accurately assess fidelity, or if the methods used to create the models do not generalise to other samples. So, in this paper, we aimed to identify, synthesise, and critique the automated coding methods that have been applied to treatment fidelity.

What is Machine Learning?

Machine learning refers to any algorithm that learns patterns from data. A linear regression model, familiar to most readers, is a form of machine learning, where an algorithm discerns the strength of the linear relationship between variables. However, machine learning also includes a broad range of other, often more complex, algorithms. These algorithms can either learn the patterns automatically by themselves (i.e., unsupervised machine learning) by, for example, identifying how data points cluster together. Alternatively, they can be trained using labelled data (i.e., supervised machine learning), where, for example, thousands of sentences are labelled by humans as 'empathic' and the model identifies the words that might indicate empathy. The line between 'statistics' and 'machine learning' is imprecise. In common usage, 'statistics' refers to more interpretable models that allow for inferences that explain a phenomenon (Hastie et al., 2009; Shmueli, 2010). 'Machine learning' is a more encompassing, umbrella term that also includes less interpretable models that may predict but not explain (Hastie et al., 2009; Shmueli, 2010). So while traditional statistics aim to explain relationships between variables, machine learning also includes methods that focus on predictive accuracy over hypothesis-driven inference (Breiman 2001). With new computational capabilities, machine learning can use large, multidimensional data to construct complex, non-linear models (Breiman, 2001). Traditional statistical methods are more interpretable but those constraints mean they perform less well in these more complex problems (Bi et al., 2019). This is an important feature because predicting interpersonal interactions requires multidimensional models that account for the complexity of human language.

Concept of Accuracy in Machine Learning

In machine learning, accuracy evaluates how well the model identifies relationships and patterns between variables in a dataset. Several evaluation metrics and validation methods have been used to evaluate the prediction performance and generalization of machine learning methods. The commonly used metrics include accuracy, precision, F1, sensitivity, specificity, and area under the receiver operating characteristic (AUC ROC) curve (for a description of the performance metrics, see Supplementary file 1). There has been extensive debate on what metric is best for which task (Handelman et al., 2019). However, one way to choose the most appropriate metric is to consider the distribution of classes and the potential cost of misclassification (Hernandez-Orallo, 2012). For example, in psychotherapy, accuracy might be a good indication of a model's performance which shows the correct prediction of true positives out of all the observations. However, in detecting suicidality, the recall (or sensitivity) metric may be important as the correct identification of all high-risk cases may be crucial. So, considering the intended purpose of using machine learning models can be helpful to determine the most appropriate performance metric and threshold.

One of the important goals of developing machine learning models is to predict the outputs in the future unseen data. Validation techniques evaluate the generalizability of models to 'out of sample' data (i.e., data not used to train the model). After training a model, validation usually involves testing the model

on new data that was not used in training. This is different from the common practice of looking at, for example, *R*-squared from the output of a regression model. Here the prediction metric—*R*-squared—comes from the same data used to build the model. From the perspective of machine learning, only predictive accuracy from new data—that is data not used in building the model—is of interest. In machine learning, new data is referred to as unseen data because the model has not seen the data and thus does not have the option to update the model or its parameters in response to it. Several methods have been used to validate models such as cross validation and hold-out 'train and test'. Cross-validation (which is also called internal validation) is a commonly used method where a dataset is separated into a training subset and a testing subset. Then, the prediction metrics are calculated to assess the prediction accuracy on the testing subset. Some of the cross-validation methods include split-half (50% training, 50% test samples), imbalanced-split (i.e., 70:30), k-fold (split into k subsets, usually 5 or 10), leave-one-out (a single test case is held-out of the training sample), or bootstrapping methods (Delgadillo, 2021; Rodriguez et al. 2010). Another validation method, named hold-out 'train and test', better estimates the generalisability of models to future datasets. This process is called external validation, where the model is trained on some data (training dataset) and is tested on data from a different sample, study, or setting. This method is stronger than cross-validation because the validation set is more likely to be representative of future data and less likely to overlap with the training set.

Machine Learning May Improve Feedback for Therapists

Therapists vary greatly in their effectiveness, and with more experience they actually decrease their effectiveness (Goldberg et al., 2016). This decline in effectiveness may be partially explained by lapses in fidelity. For example, without feedback or coaching, fidelity to motivational interviewing substantially decreases within six months of training (Schwalbe et al., 2014). This is often described as 'therapist drift', where well-meaning therapists fail to adhere to the prescribed practice guidelines (Waller, 2009; Waller & Turner, 2016). Therapists are bad at identifying these problems themselves because they rely on unreliable signals of their own effectiveness (Tracey et al., 2014). However, it is possible to mitigate these problems through quality feedback, auditing, and supervision (Barwick et al., 2012; Ivers et al., 2012; Madson et al., 2009). Indeed, one of the core goals of training and clinical supervision is increasing treatment fidelity (Bellg et al., 2004; Reiser & Milne, 2014). Accurate and individualised feedback enables therapists to adopt effective strategies to enhance client outcomes (Ivers et al., 2012; Tracey et al., 2014). Research shows that feedback is most effective when it is distributed over a period of time on multiple occasions (Ivers et al., 2012). For example, three to four post-workshop feedback sessions prevent skill erosion among Motivational Interviewing trainees (Schwalbe et al., 2014). However, providing feedback using traditional methods is an expensive process for agencies and a time consuming job for supervisors. It can be even a more resource-intensive process when there are many therapists in a large scale training. New techniques, such as machine learning, are capable of quickly and cheaply analysing large-scale data, providing accurate individualised feedback.

Automated coding methods have been applied to large psychotherapy datasets up to 1,553 sessions (Xiao et al., 2016). Once these models are trained, they can be repeatedly applied at very low cost (Xiao et al., 2016). They can reduce the likelihood of implicit bias of human decision-making (Lum, 2017), where the look or the sound of the therapist may contribute to errors in judgments. While some may doubt whether therapists would accept the feedback from machine learning models, preliminary feedback has been promising. Hirsch et al. (2018) provided machine learning based-feedback for 21 counsellors and trainees. The results of their qualitative study showed that counsellors were receptive to a computerised assessment, and were less defensive toward critical feedback from a machine than a human. It has also been documented that therapists are quite open to receiving machine learning feedback (Imel et al. 2019). In sum, machine learning models can cheaply provide objective feedback to therapists in a way that they are likely to find valuable.

Verbal Behaviour May Be a Good Candidate for Machine Learning

Interpersonal interactions in a therapy process involves a range of behaviours such as verbal behaviours (i.e., what is said) and non-verbal behaviours (such as prosody, body movements, biological changes). However, verbal behaviours are the primary channel of transferring information in dyadic interactions (Miller et al., 2003). Systematic reviews have shown that therapists' verbal behaviours are associated with various client outcomes, such as patient satisfaction and adherence to treatment (Golin et al., 2002; Howard et al., 2009). Most existing measures for assessing treatment fidelity focus on the words used by the therapist, rather than their tone or non-verbal behaviour (McGlinchey & Dobson, 2003; Miller et al., 2003; Moyers, Martin, et al., 2005). Verbal behaviour is also easy to code automatically, where even simple 'word-counting' methods can reliably and validly predict many psychological constructs (Pennebaker et al., 2003). Further, methods for automatic assessment of verbal behaviour are different from those for non-verbal or para-verbal (e.g., signal-processing features like tone, pitch, and pacing) behaviours. Many such tools have allowed for automated assessment of patient characteristics, such as diagnoses (Low et al., 2020). Emerging technologies may be able to code some non-verbal behaviour like sign language, but those technologies are not sufficiently advanced that they can code the nuanced non-verbal cues involved in psychosocial interventions. So, while non-verbal and para-verbal modalities are critical components of therapy, we focused on verbal interactions as an important and tractable machine learning task.

To analyse verbal behaviour, human coders are trained to identify specific therapy behaviours. The reliability of human-to-human codes are evaluated via a process called interrater reliability. Just as therapists drift, coders do too, where interrater reliability can decrease with fatigue or without frequent re-calibration (Atkins et al., 2012; Haerens et al., 2013). Often when two humans code for fidelity using words therapists use, they are not perfectly aligned. Coders may overcome the 'coding drift' by meeting regularly to discuss their codes and instances of coder disagreement. However, human coding also faces other challenges such as being tedious, expensive, and time consuming (Moyers et al., 2005). This means that human coding is an imperfect reference point, but a useful one to compare machine learning models against.

Proof-of-concept comes from many other fields in which machine learning has been found to reliably automate laborious tasks (Russell & Norvig, 2002). Ryan et al. (2019) have argued that machine learning is already good enough to assess the content and delivery of healthcare by doctors. They have been applied to predict language disorders (Can et al., 2012), and addiction and suicide behaviour (Adamou et al., 2018). In psychotherapy, they have been used to predict counselling fidelity (Atkins et al., 2014), empathy (Xiao et al., 2015), and counsellor reflections (Can et al., 2016). A recent systematic review showed that 190 studies used machine learning methods to detect and diagnose mental disorders, indicating the applicability of machine learning in mental health research (Shatte et al., 2019). Similarly, Aafjes-van Doorn et al. (2021) did a scoping review of machine learning in the psychotherapy domain and showed that 51 studies applied machine

learning models to classify or predict labelled treatment process or outcome data, or to identify clusters in the unlabelled patient or treatment data (Aafjes-van Doorn et al., 2021). Machine learning methods have also been used in psychiatry to parse disease models in complex, multifactorial disease states (e.g., mental disorders; Tai et al., 2019). When taken together, there are a number of domains in which machine learning models have been helpful in coding verbal behaviours, indicating they may be a powerful tool for psychotherapists and other helping professions.

Review Aims

The primary goal of this review is to assess how well machine learning performs as a method for assessing treatment fidelity using verbal aspects of therapist language. By conducting a systematic review, we were able to assess how well those models applied across studies and contexts. Models may only work well under a narrow set of conditions, and systematic reviews are able to assess those conditions more robustly than a narrative review. There are also some well-established best-practices that influence whether a machine learning model will generalise to new data (Luo et al., 2016). By assessing adherence to these guidelines, our review was able to indicate how well these models may generalise. Finally, we included all interpersonal interactions from helping professionals, even those outside psychotherapy (e.g., medicine, education), in order to assess whether machine learning models to assess communication and fidelity have been successfully implemented in nearby fields. In doing so, we could see whether models applied to medicine or education might be useful to consider in future psychological research. In sum, we sought to answer the following research questions:

1. Which automated coding methods have been used to analyse interpersonal verbal behaviours of helping professionals (with specific focus on fidelity in psychotherapy)?
2. How accurate are machine learning methods?
3. To what extent have studies applying automated coding methods adhered to best-practice guidelines for machine learning?

Method

We report this systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (Moher et al., 2009).

Protocol and Registration

We prospectively registered the protocol in the Prospective Register of Systematic Reviews (PROSPERO registration number: CRD42019119883).

Eligibility Criteria

In this review, we included studies meeting the following criteria:

1. The participants or population studied were helping professionals. A helping professional engages in "a professional interaction with a client, started to nurture the growth of, or address the problems of, a person's physical, psychological, intellectual, or emotional constitution" (Graf et al., 2014, p. 1). Examples of helping professionals are psychotherapists, counsellors, doctors, nurses, teachers, and social workers.
2. They measured verbal interpersonal interactions between helping professionals and clients (e.g., clinician and client, or teacher and student).
3. They analysed the helping professionals' verbal behaviour (i.e., language) that occurred during interpersonal interactions.
4. They used an automated method for coding behaviour. Coding refers to the process of either rating or categorising an interpersonal interaction on at least one variable. Automated coding methods refer to the methods which code the input data without manual interference in the coding process. The input data for such systems could be transcripts, audio tracks, or video clips (with audio included). Codes are labels that are used to represent certain behaviours, and they may vary in their level of granularity or specificity and concreteness (ranging from physically to socially based codes; Bakeman & Quera, 2011).
5. Both peer reviewed and grey-literature (e.g., conference papers, theses) were eligible for inclusion.
6. Papers written in any language with title and abstract in English were included.
7. Any design, location or year were included.

Exclusion Criteria

We excluded studies if:

1. Participants were not helping professionals.
2. They analysed interprofessional interactions (e.g., doctors interacting with nurses).
3. They analysed interpersonal interactions using only aspects other than language (i.e., facial expressions, body posture and gestures).
4. They used semi-automated methods (where the final results still required some human coding) or manual methods (where a human is needed to code the behaviour).
5. They were published abstracts, without a full-length paper.

Search Strategy and Information Sources

To develop the search strategy, we created an initial pool of target papers that met the inclusion criteria. We conducted forward and backward citation searching on this initial pool (Hinde & Spackman, 2015) to identify six more papers meeting the eligibility criteria. We extracted potential search terms from these 11 papers by identifying key words from the title and abstract (Hausner et al., 2016). The final search strategy involved keywords and their MeSH terms or synonyms from four main groups including 'participants' (e.g., teacher or doctor), 'measurement' (such as assessment or coding), 'automated coding method' (e.g., Natural Language Processing or text mining), and 'type of behaviour' (e.g., fidelity or interaction). The search did not have any exclusion terms (see Supplementary file 2 for full search details and included papers).

We performed the search within PubMed, Scopus, PsycINFO, Education Source, ERIC, CINAHL Complete, Embase, SPORTDiscus, and Computers and Applied Sciences Complete databases. We performed the last search on the 21st of February 2021. To test the sensitivity of our strategy, we first confirmed that the identified records included 11 target papers described earlier. We then searched the first 200 results on Google Scholar to identify potentially relevant studies not indexed in electronic databases.

We conducted forward and backward citation searching on studies that passed full-text to identify related papers which did not appear in the systematic search (Greenhalgh & Peacock, 2005; Hinde & Spackman, 2015). We also emailed the first author of included papers and known experts in the automated coding of verbal behaviour to identify any unpublished manuscripts.

Study Selection

We imported search results into Covidence software (Babineau, 2014). We dealt with studies in two steps. First, we screened the titles and abstracts of the studies according to the pre-defined

Table 1. Quality Assessment (continued)

Item/Study	Does the paper clarify the clinical setting for the target predictive model?	Does the paper describe the modelling context in terms of facility type, size, volume, and duration of available data?	Does the paper define a measurement for the prediction goal (per patient or per hospitalization or per type of outcome)?	Does the paper define the success criteria for prediction (e.g., based on metrics in internal validation or external validation in the context of the clinical problem)?	Does the paper define the observational units on which the response variable and predictor variables are defined?	Does the paper describe the data pre-processing performed, including data cleaning and transformation?	Does the paper define model validation strategies?	Does the paper report the predictive performance of the final model in terms of the validation metrics specified in the methods section?	Does the paper report (if possible) what variables were shown to be predictive of the response variable?
Althoff et al., 2016	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Angus et al., 2012	Yes	Yes	Yes	No	Yes	Yes	No	No	No
Atkins et al., 2014	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Blanchard et al., 2016a	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes
Blanchard et al., 2016b	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Can et al., 2015	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
Can et al., 2012	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Can et al., 2016	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Cao et al., 2020	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Carcone et al., 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Chakravarthula et al., 2015	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Chen et al., 2019	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Donnelly et al., 2017	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Donelly et al., 2016a	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Donnely et al., 2016b	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Flemotomos et al., 2018	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Gallo et al., 2015	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No
Gaut et al., 2017	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gibson et al., 2019	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Gibson et al., 2017	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Gibson et al., 2016	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Goldberg et al., 2020	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Gupta et al., 2014	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Hasan et al., 2019	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes
Hasan et al., 2018	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Howes et al., 2013	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Imel et al., 2015	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Lacson and Barzilay, 2005	Yes	Yes	No	Yes	Yes	No	No	Yes	No
Malandrakis and Narayanan, 2015	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Mayfield et al., 2014	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Mieskes and Stiegelmayr, 2018	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Nitti et al., 2010	Yes	Yes	No	No	Yes	Yes	No	No	Yes
Park et al., 2021	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Park et al., 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Perez-Rosas et al., 2017	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Perez-Rosas et al., 2019	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Salvatore et al., 2012	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No
Samei et al., 2014	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Samei et al., 2015	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Sen et al., 2017	Yes	Yes	No	No	Yes	No	Yes	Yes	No
Singla et al., 2018	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
Song et al., 2020	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Suresh et al., 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Tanana et al., 2016	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Velasquez and Montiel, 2018	Yes	Yes	No	No	Yes	No	No	No	No

Table 1. Quality Assessment

Item/Study	Does the paper clarify the clinical setting for the target predictive model?	Does the paper describe the modelling context in terms of facility type, size, volume, and duration of available data?	Does the paper define a measurement for the prediction goal (per patient or per hospitalization or per type of outcome)?	Does the paper define the success criteria for prediction (e.g., based on metrics in internal validation or external validation in the context of the clinical problem)?	Does the paper define the observational units on which the response variable and predictor variables are defined?	Does the paper describe the data pre-processing performed, including data cleaning and transformation?	Does the paper define model validation strategies?	Does the paper report the predictive performance of the final model in terms of the validation metrics specified in the methods section?	Does the paper report (if possible) what variables were shown to be predictive of the response variable?
Wallace et al., 2014	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Wallace et al., 2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Wang et al., 2014	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes
Xiao et al., 2012	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Xiao, Can, et al., 2016	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Xiao, Huang, et al., 2016	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Xiao et al., 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sum of 'Yes' items	52	52	43	18	52	28	44	48	19

inclusion criteria. If the title or abstract did not provide enough information to decide, we moved the record to full-text screening. Second, we reviewed full texts of articles for final inclusion. At each stage, two reviewers (AA and MS, or AA and DA) independently made recommendations for inclusion or exclusion. We resolved any discrepancies in study selection at a meeting. Then, we resolved any conflicts by consulting with a third reviewer (MN). The PRISMA flow diagram (Figure 1) provides detailed information regarding the selection process.

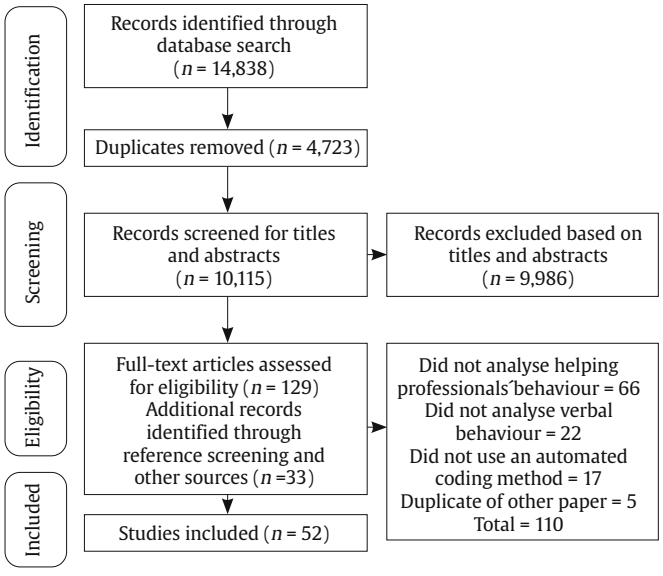

Figure 1. PRISMA Flow Diagram of the Study Selection Process.

Data Collection Process

We developed a data extraction form for this review to focus on the applied automated coding methods and their performance. We first tested the form by extracting data from four randomly selected papers. Two researchers (AA and MS or AA and DA) then independently extracted data from each study and organised it into tables to display themes within and across the included studies. Any discrepancies from the data extraction were discussed between the reviewers. In the case of unresolved disagreements, a third reviewer (MN) was consulted.

Adherence to Best-Practices in Machine Learning

We assessed study quality using a tool based on the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View" (Luo et al., 2016). This tool was used to judge the extent to which studies adhered to best-practice guidelines. The original checklist contained 51 items and investigated the quality of papers based on the information in each section of a paper. The checklist was used in two ways. One researcher (AA) assessed all 51 items. We refined this checklist by identifying the core items related to performance of automated coding methods. Of the 51 items, nine were related to the performance (see identified items in Table 1, and the complete checklist in Supplementary file 3); the others related to the reporting in the manuscript (e.g., three items are whether the abstract contains background, objectives, or data sources sections). The other researcher (MS/DA) assessed the core checklist. Specifically, the two researchers independently assigned the label "Yes" if the requisite information was described explicitly and "No" if the information was not adequately described. Rather than reporting a summary score (e.g., "high" or "low quality"), we followed Cochrane guidelines that recommend reporting quality scores for each item of the quality assessment checklist (Macaskill et al., 2010).

Results

Study Selection and Results of Individual Studies

Our systematic search resulted in 14,838 records. We removed 4,723 duplicates, with 9,986 papers remaining for title and abstract screening. Thirty-three further records were added by other methods (e.g., forward and backward searching). Fifty-two

Table 2. Context of Study

Context	Psychotherapy	Medical care	Education
Studies	Counselling, Motivational Interviewing (counsellors), (Atkins et al., 2014; Can et al., 2015; Can et al., 2012; Can et al., 2016; Cao et al., 2020; Carcone et al., 2019, Study 1; Chakravarthula et al., 2015; Chen et al., 2019; Gibson et al., 2019; Gibson et al., 2017; Gibson et al., 2016; Gupta et al., 2014; Hasan et al., 2019; Hasan et al., 2018; Imel et al., 2015; Perez-Rosas et al., 2017; Perez-Rosas et al., 2019; Singla et al., 2018; Tanana et al., 2016; Xiao et al., 2012; Xiao et al., 2015; Xiao, Can, et al., 2016; Xiao, Huang, et al., 2016)	Medical care, provider-patient clinical interactions (Carcone et al., 2019, Study 2; Park et al., 2019)	Education (teachers) (Blanchard et al., 2016a; Blanchard et al., 2016b; Donelly et al., 2017; Donnelly et al., 2016a; Donnelly et al., 2016b; Samei et al., 2014; Samei et al., 2015; Song et al., 2020; Suresh et al., 2019; Wang et al., 2014)
	Counselling, (counsellors), (Althoff et al., 2016; Flemotomos et al., 2018; Gallo et al., 2015; Gaut et al., 2017; Goldberg et al., 2020; Malandrakis & Narayanan, 2015; Mieskes & Stiegelmayr, 2018; Nitti et al., 2010; Salvatore et al., 2012; Velasquez & Montiel, 2018)	Medical care, (nurses) (Lacson & Barzilay, 2005)	
	Counselling (psychiatrists), Howes et al., 2013	Medical care (physicians, nurses, physician assistants) (Mayfield et al., 2014)	
		Medical care (oncologists) (Sen et al., 2017)	
		Medical care (physicians) (Angus et al., 2012; Park et al., 2021; Wallace et al., 2013, 2014)	
Total[1]	35 (64.8%)	9 (16.6%)	10 (18.5%)

Note. [1]One study was performed in two different contexts.

Table 3. Frequency of Behavioural Coding Measures Used in Included Studies

Behavioural Coding Measure	Frequency
Motivational Interviewing Skill Code	14
Motivational Interviewing Treatment Integrity	7
Nystrand et al.'s (2003) coding scheme	7
Minority Youth-Sequential Code for Observing Process Exchanges	3
Generalized Medical Interaction Analysis System	3
Diagnostic and Statistical Manual of Mental Disorders - 4th edition	2
A coding manual developed in a previous study (in Prado et al., 2006; Stigler et al., 2000)	2
Cognitive Therapy Rating System (CTRS)	2
Cognitive therapy scale for psychosis (in Lecomte et al., 2017)	1
Accountable Talk framework (Michaels et al., 2008)	1
Multi-Dimensional Interaction Analysis coding system	1
Did not apply a previously established behavioural coding system	12

Note. Some studies used more than one behavioural coding measure.

papers met the inclusion criteria and were included in this review (see Figure 1). All the included papers were written in English. Supplementary file 4 summarises the information from individual studies.

Synthesis of Results

Most of the studies were conducted in psychotherapy settings (k = 34, 65.3%) and involved counsellors, psychologists, or psychiatrists. Nine studies were conducted in a medical care setting (16.6%) and included physicians or nurses. Ten studies (18.5%) were conducted in education contexts and involved school teachers. Of the 53 studies, 23 (41.5%) examined Motivational Interviewing (Miller & Rollnick, 1991) with the rest of the studies scattered across different modalities (one paper included two studies, for details see Table 2).

Predicted outcomes. Studies in the psychotherapy context aimed to predict the fidelity to a prescribed therapeutic process (k = 28, 82.3% of psychotherapeutic studies). In medical care settings, the aim was to identify clients' symptoms (k = 1), topics discussed in conversations (k = 5), or conversational patterns (k = 5). In educational contexts, studies aimed to predict the number of teacher questions (k = 5) and the type of classroom activities (e.g., discussion, lecture, or group work, k = 5).

Behavioural coding measures and automated coding methods. Many studies used automated coding to implement pre-existing behavioural coding measures. Behavioural coding measures were usually designed to measure adherence to the practice guidelines or instructions. The majority of studies used a behavioural coding measure (for details, see Table 3). The most frequently applied coding measure was Motivational Interviewing Skills Code (k = 14; Miller et al., 2003), followed by the Motivational Interviewing Treatment Integrity measure (k = 7; Moyers, Martin, et al., 2005). Seven studies used a coding system to code whether teachers asked questions, provided instructions, or facilitated small-group activities (Nystrand et al., 2003).

In this context, the machine learning methods were designed to automatically assign codes from the behavioural coding measures to overt interactions recorded in the dataset (e.g., words/utteran-

Table 4. Automated Coding Methods

Automated Coding Method	Frequency[1]	Citations
Support Vector Machine	8	Carcone et al., 2019 (Study 1 and 2); Howes et al., 2013; Perez-Rosas et al., 2017; Perez-Rosas et al., 2019; Xiao et al., 2015; Park et al., 2019; Flemotomos et al., 2018.
Random Forest	7	Carcone et al., 2019; Imel et al., 2015; Mieskes and Stiegelmayr, 2018; Blanchard et al., 2016a; Blanchard et al., 2016b; Donnelly et al., 2017; Wang et al., 2014.
Logistic Regression	7	Park et al., 2019; Sen et al., 2017; Donnelly et al., 2017; Blanchard et al., 2016a; Blanchard et al., 2016b; Park et al., 2021; Mayfield et al., 2014.
J48 (Decision Tree)	6	Carcone et al., 2019; Howes et al., 2013; Blanchard et al., 2016a; Blanchard et al., 2016b; Donnelly et al., 2017; Samei et al., 2014.
Maximum Entropy Markov	5	Can et al., 2012; Can et al., 2016; Gupta et al., 2014; Xiao, Can, et al., 2016; Xiao., Huang et al., 2016.
Naive Bayes	5	Carcone et al., 2019; Blanchard et al., 2016a; Donnelly et al., 2016a; Donnely et al., 2016b; Donnelly et al., 2017.
Recurrent Neural Networks	5	Hasan et al., 2018; Singla et al., 2018; Blanchard et al., 2016a; Park et al., 2021; Gibson et al., 2017.
Hidden Markov Model	4	Althoff et al., 2016; Can et al., 2012; Hasan et al., 2019; Hasan et al., 2018
K-Nearest Neighbours	4	Blanchard et al., 2016a; Sen et al., 2017; Blanchard et al., 2016b; Donnelly et al., 2017
Conditional Random Field	4	Can et al., 2015; Carcone et al., 2019; Wallace et al., 2014; Park et al., 2019
Bi-directional Long Short Term Memory (Bi-LSTM)	3	Chen et al., 2019; Gibson et al., 2019; Suresh et al., 2019
Labelled Topic Model	2	Atkins et al., 2014; Imel et al., 2015
Bayesian Network	2	Blanchard et al., 2016a; Blanchard et al., 2016b
Gared Recurrent Unit (GRU)	2	Cao et al., 2019; Park et al., 2019
30 models were used once each[2]	1 each (30 in total)	Angus et al., 2012; Carcone et al., 2019; Chakravarthula et al., 2015; Gallo et al., 2015; Gaut et al., 2017; Gibson et al., 2016; Gibson et al., 2017; Hasan et al., 2018; Howes et al., 2013; Imel et al., 2015; Lacson et al., 2005; Malandrakis and Narayanan, 2015; Nitti et al., 2010; Salvatore et al., 2012; Tanana et al., 2016; Velasquez & Montiel, 2018; Wallace et al., 2013; Xiao et al., 2012; Xiao, Huang et al., 2016; Xiao, Can, et al., 2016; Samei et al., 2015; Park et al., 2019; Song et al., 2020; Goldberg et al., 2020.
Unique Models = 41	All the used models = 94	

Note. [1]Some studies applied more than one coding method. We reported all the specific models that were applied in the studies. Some models might be variations of another model.
[2]The models were: Activation-based Dynamic Behaviour Model (ADBM) using Hidden Markov Model, AdaBoost, Automated Co-occurrence Analysis for Semantic Mapping (ACASM), Boostexter tool, Deep Neural Networks, DiscLDA, Discourse Flow Analysis (DFA), Discrete Sentence Features using Multinomial Logistic Regression, Discursis software, Fidelity Automatic RatEr (FARE system), Joint Additive Sequential (JAS) model using Log-linear classifier, Labeled Latent Dirichlet Allocation, Lasso Logistic Regression (LLR), Latent Dirichlet Allocation, Likelihood-based Dynamic Behaviour Model (LDBM) using Hidden Markov Model, Linear Regression, Markov Chain, Markov-Multinomial, Maximum Likelihood Classifier with Universal Background Model (UBM) and Kneser-Ney algorithm, Maximum Likelihood Model with Kneser-Ney algorithm, Naive Bayes-Multinomial, RapidMiner, Recurrent Neural Networks with Gated Recurrent Unit (GRU), Recursive Neural Network (RNN), Ridge Regression model, Static Behaviour Model (SBM) using Universal Background Model, Hidden Markov Model Logistic Regression (HMM-LR), Hidden Markov Model-Support Vector Machine (HMM-SVM), Hidden Markov Model-Gated Recurrent Unit (HMM-GRU), Convolutional Neural Network - Bidirectional Long Short Term Memory (CNN-BiLSTM) model.

ces). Most studies assessed more than one machine learning method; the most frequently applied were Support Vector Machine (k = 8), Random Forests (k = 7), Logistic Regression (k = 7), J48 classifiers (a type of decision tree, k = 6), Maximum Entropy Markov models (k = 5), and Naive Bayes (k = 5; for details, see Table 4).

Which methods performed best? In Supplementary file 5, we report the predictive performance of each method (e.g., F1-score measure for the Support Vector Machine in Xiao et al., 2015 is .89). We also reported a brief description of each coding method and accuracy measures in the Supplementary file 1. Methods generally performed well in terms of their agreement with human coders. Overall, kappa ranged from .24 to .66, with all but one study (Samei et al., 2014) falling between .38 and .66. These results suggested fair to excellent levels of agreement, compared with established thresholds for kappa used for human-to-human agreement (Landis & Koch, 1977). Accuracy—meaning the ratio of correctly predicted codes to the total number of predictions—was greater than 50% in all studies and sometimes higher than 80% (e.g., Chakravarthula et al., 2015; Wang et al., 2014; Xiao et al., 2016).

Support Vector Machine methods generally performed well. For example, Xiao et al. (2015) found that the Support Vector Machines methods performed almost as well as trained coders. Similar results were reported in other studies (e.g., Flemotomos et al., 2018; Pérez-Rosas et al., 2019; Pérez-Rosas et al., 2017). Most studies only examined one type of method's performance. In one study that directly compared different methods on the same dataset, Support Vector Machines outperformed seven alternative method strategies in terms of agreement with human coders and accuracy (Carcone et al., 2019).

Because few studies examined the performance of methods when transferred to other similar settings—for example, with similar predictors and outcomes but different participants—we are unable to ascertain whether any particular method predicted new data better than others. There were three studies that compared the performance of methods but did not report the predictive performance of all the tested methods and only chose the best performing method (Blanchard et al., 2016a, 2016b; Donnelly et al., 2017). Only one study developed a Support Vector Machine method in psychotherapy and applied it on new data from another context (i.e., medicine; Carcone et al., 2019). The method performed well, achieving a substantial level of agreement with human coding.

Larger datasets lead to more accurate performance. Dataset sizes ranged from 13 sessions (Wang et al., 2014) to 1,235 sessions (Goldberg et al., 2020). When the dataset size was larger, methods performed more accurately. For example, Imel et al. (2015) analysed more than 9 million words and the method achieved an accuracy of 87% (using a Random Forest). Similar results were reported in other studies with large datasets (e.g., Gaut et al., 2017; Xiao et al., 2016; Xiao et al., 2015). Pérez-Rosas et al., 2019 showed that as they increased the amount of data in their training set they observed significant improvement in prediction accuracy. Aligned with this finding, frequently observed codes (i.e., categories) in a dataset

were predicted more accurately, while low base rate codes were predicted less accurately (e.g., Can et al., 2015; Cao et al., 2019; Carcone et al., 2019; Gibson et al., 2017; Tanana et al., 2016; Wallace et al., 2014). An example of frequently observed code is 'open questions' and an example for low base rate codes is 'confrontational statements'.

The fewer the codes the more accurate the performance. Methods classified data into codes, with the number of codes ranging from two (Blanchard et al., 2016a; Xiao et al., 2015) to 89 (Gaut et al., 2017). When the number of codes decreased, performance of the method increased, and vice versa. Carcone et al. (2019) showed that the methods performed better in 17-code prediction than 20-code prediction, and 20-code prediction was superior to 41-code prediction. Similar results were reported in other studies that directly compared coding frameworks of differing complexity (e.g., Gallo et al., 2015; Gibson et al., 2016). When methods were simpler (i.e., two codes), accuracy was greater than 80% (e.g., Blanchard et al., 2016a; Chakravarthula et al., 2015; Gallo et al., 2015; Pérez-Rosas et al., 2019; Xiao et al., 2016). When the number of codes was higher, prediction was less accurate (i.e., accuracy = 54% with 41 codes in Carcone et al., 2019; accuracy = 66% with 20 codes in Howes et al., 2013).

More concrete and less abstract codes lead to better performance. The conceptual meaning of the codes affects the predictive performance of methods. Methods accurately predicted some types of codes. For example, questions (e.g., a counsellor or teacher asking questions to gather information, such as "How do you feel about that?") and facilitation (i.e., simple utterances that function as acknowledgements and a cue to continue speaking, such as "hmm-mm") seem to be conceptually concrete. These codes were predicted more accurately than conceptual abstract codes, such as empathy (Atkins et al., 2014), confrontation, and advising (Imel et al., 2015; Tanana et al., 2016).

Session-level prediction is more accurate than utterance-level prediction. Utterance-level prediction refers to the prediction of a small unit of spoken words that have a specific meaning (i.e., complete thoughts). For instance, "You feel overwhelmed" is an utterance that may signal reflective listening. Session-level prediction refers to the prediction of a behaviour or skill over a session. For example, in Motivational Interviewing Treatment Integrity coding measure, the empathic quality of the provider is rated on a 1-5 Likert scale, taking the entire session or segment of the session into account (Moyers et al., 2016). Session-level prediction may also code whether the therapist implemented a specific behaviour (e.g., reflective listening) *frequently* (e.g., 10/10) or *rarely* (0/10). Compared with utterance-level prediction, Tanana et al. (2016) showed that the session-level prediction results had stronger concordance with human-based coding. Atkins et al. (2014), and Park et al. (2019) reported similar results, where the session-level prediction was generally closer to human coding rather than utterance-level prediction.

Quality of Reporting Within Studies

Results of our study quality assessment can be found in Table 1. Inter-rater reliability analysis of the quality assessment among this systematic review team showed agreement on 89% of the instances assessed by two independent reviewers. We resolved discrepancies by discussion between the two researchers (AA and MS or AA and DA) and consultation with a third reviewer (MN).

We report quality assessment results for each item of the core checklist (nine items). All the papers reported the clinical setting, dataset details, and observational units. Forty-five papers (86.5% of studies) coded behaviours using a behavioural coding measure. These types of concrete guidelines facilitate utterance level comparison. Twenty-eight papers reported data pre-processing (53.8% of the studies), which improves performance of a method by removing outliers or poor quality data (e.g., removing very low quality voice recordings; García et al., 2014). Thirty-four papers (64.1% of studies) validated the methods using some form of cross-validation (where a method is trained on a dataset and tested on a unseen set of observations; Browne, 2000). Yet, only eleven papers (21.1% of studies) applied a hold-out 'train and test' method. Studies that do not test the accuracy on unseen data can overfit the data to the training set, and give misleading estimates of how accurately the method can predict new data (Yarkoni & Westfall, 2017). Eighteen papers (34.6% of studies) reported success criteria (e.g., mean-squared error), which help to interpret performance of a method. Relative importance of predictor variables (e.g., which feature is most important in predicting the outcome variable) were reported in 19 papers (36.5% of studies). For full details about each quality indicator, see Supplementary file 3.

Discussion

Our systematic review found that several automated coding methods have been applied to assess fidelity in psychological interventions. We also identified many methods used to analyse verbal interactions of other helping professionals, not just therapists. These methods generally demonstrated promising results with accuracy comparable to human coding. Methods performed better on large datasets, coding frameworks with fewer behaviours, and verbal behaviours that represent concrete (rather than abstract) codes. However, studies rarely reported adherence to best-practice machine learning guidelines, meaning that the machine learning models may not generalise well to new interactions with new clients, reflecting a deficit in the field.

Methods showed promising performance in automatic annotation of therapists' verbal behaviour, including treatment fidelity to a number of models (most frequently Motivational Interviewing). This result suggests machine learning could reduce financial costs of traditional methods. Doing so would improve the scalability and efficiency of behavioural coding for assessment and feedback on treatment fidelity. When directly compared with other methods, the Support Vector Machines method showed superior performance and appeared to be an appropriate method for generalisability purposes (Carcone et al., 2019). The higher performance of the Support Vector Machines method was also reported in other studies in the similar applications (Hasan et al., 2016; Kotov et al., 2014). This method might have potential in less-explored contexts such as fidelity for cognitive behaviour therapy or acceptance and commitment therapy, because the machine learning models efficiently process sparse, high-dimensional data and non-linearities with few interactions.

Having said that, the field of machine learning is advancing quickly and the methods reported here may not reflect the current state-of-the-art. For example, Kaggle's machine learning competitions have recently been dominated by Extreme Gradient Boosting or Neural Network methods (Abou Omar, 2018). New, powerful, natural language models contain up to 175 billion parameters and require only a few pieces of training data (Brown et al., 2020). We expect that automated coding methods will become even more powerful, and better-able to manage ambiguity, once researchers start implementing these cutting-edge methods. Our findings were restricted by the small number of studies that directly compared different machine learning methods; therefore, caution should be taken when generalising the predictive performance of these methods to other cases. Researchers in this area could help accelerate the field by transparently reporting which models were tested and discarded, and why. It is common practice in machine learning to test a number of models using cross-validation on the training set

(Cawley & Talbot, 2010); we were therefore surprised to see so few head-to-head comparisons reported. It is possible that researchers only reported the performance of a model that performed best with their data. This is concerning because few studies reported how well the models predicted unseen data on a hold-out, 'test set' and thus the risk of over-fitting was potentially high.

There were rare cases where automated coding methods did not perform well (Gallo et al., 2015; Samei et al., 2014). While the method itself can be an important factor in prediction accuracy, there are important conditional factors, such as dataset size, that affect a method's accuracy (Yarkoni & Westfall, 2017). Considering these conditions, it was not easy to provide a fair comparison between statistical models because the choice of model was often confounded by differences in samples and prediction objectives. In the following section, we present a cautious overview of the factors that influence the methods' predictive performance and provide suggestions for future research and practice.

While determining the appropriate size of a dataset remains a matter of debate, large datasets support training, testing, and generalization of predictions in new datasets (Yarkoni & Westfall, 2017). Future studies could identify whether or not more data are needed by looking at the learning curves, which show whether the method can be adequately fit with the data available (Perlich, 2009). In general, our results showed that larger datasets lead to better performance. This finding is in line with previous studies where machine learning algorithms generally performed better on larger datasets (Domingos, 2012). It is important to note, however, that additional data have diminishing returns. As such, it is important for analysts to monitor method performance as sample sizes increase in order to maintain reasonable cost-benefit ratios (Ng, 2019).

Another factor influencing methods' performance is the number of codes a method is built to predict. Methods generally performed worse when the number of codes increased (e.g., Gallo et al., 2015; Hasan et al., 2016). As such, we recommend analysts carefully consider which codes are most critical as a means of increasing method performance. When learning curves indicate that data is under-fit, then authors could consider using fewer codes (e.g., by collapsing conceptually similar codes) to allow for more reliable methods.

Codes with simple conceptual meaning were predicted more accurately (e.g., open-ended questions), while complicated codes were predicted weakly (e.g., informing with permission from the client vs. informing without permission). Researchers might consider the trade-off between the lower prediction accuracy for complicated codes and the higher costs of coding them using alternative methods (e.g., manual coding). Similarly, codes that can be objectively identified in a transcript (e.g., questions, affirmations, and facilitations) are likely to be more easily coded than those that require inference and subject-matter expertise.

Many accurate methods in this review were applied in the Motivational Interviewing context. The behavioural coding systems for Motivational Interviewing are well defined and more reliably coded than many other therapeutic approaches (Miller & Rollnick, 1991). This may be because Motivational Interviewing explicitly prescribes a number of conversational devices (e.g., reflections, affirmations, open questions) to be used in session, where other practices are less prescriptive regarding the conversation process and more focused on the content of discussion (e.g., a client's idiosyncratic negative automatic thoughts). Similarly, the techniques prescribed by motivational interviewing may occur hundreds of times a session (e.g., reflective listening). Core techniques from other treatment approaches may only happen once per session (e.g., checking homework). As a result, machine learning methods may be less reliable where behavioural codes are less clear, like in other psychological treatment approaches (e.g., cognitive-behaviour therapy).

Finally, methods tend to perform poorly when codes are constructed at the utterance-level; the overall prediction of a code was more reliable over a session. Part of the reason for this arises from the difficulty of utterance-level coding tasks—even for human coders—if they do not rely on the prior or subsequent utterances (Tanana et al., 2016). Without context, it is difficult to know whether "your drinking is a problem" is an empathic response to a client's self-awareness or a controlling, unsolicited prescription. As a result, it is more reasonable to rely on the overall prediction results over a session rather than each individual utterance. Recently, Cao et al. (2019) investigated the prediction of therapist utterance labels by taking the context of the utterance into consideration. They found that by increasing the history window size (i.e., by accounting for the last 8 utterances), categorization accuracy improved (Cao et al., 2019). This indicates that providing machine learning with more context may improve the accuracy of models. The other reason for poor performance at utterance-level prediction compared to session-level prediction may be that, across a session, the machine-learning task is closer to a regression problem than a classification problem. That is, it may be hard to classify a moment as 'empathic' from a set of words, but it may be easier to correlate ratings of empathy with the frequency of specific words across an entire session (e.g., "you feel…", "it sounds like…").

Atkins et al. (2014) presented the potential factors impacting the accuracy of Topic Models in predicting client and therapist codes in the Motivational Interviewing Skill Code. Like our review, they argued that models worked less accurately at utterance (i.e., talk-turn) level than at session level. They also stated that more abstract codes were weakly predicted than more concrete ones. However, their findings only focused on one of the many psychosocial interventions (motivational interviewing), and our systematic review identified other factors which are likely to influence the performance of machine learning methods. Particularly, this systematic review showed that larger datasets and more frequently observed codes lead to better prediction accuracy. Also, fewer target behaviours leads to higher accuracy. Further, other factors impact the predictive power of a model, such as the machine learning model selection process, pre-processing, and validation method.

Potential Applications

Specific and immediate feedback is essential to the development of skills across domains (Kahneman & Klein, 2009). Feedback works best when it is provided several times, spaced over a period of time (Ivers et al., 2012). However, providing individualised, distributed, and prompt feedback multiple times for a big group of therapists can be prohibitively expensive. Automated coding methods showed promising results in analysing helping professionals' language, so they can be used to provide feedback and improve practitioners' skills. Our systematic review shows that automated coding methods provided accurate estimation of treatment fidelity, including all three components (adherence, differentiation, and competence; Rodriguez-Quintana & Lewis, 2018). In motivational interviewing, for example, automated methods were able to code adherence to therapeutic strategies (e.g., affirming change), differentiation of proscribed strategies (e.g., use of closed questions; Tanana et al., 2016), and competence in delivery (e.g., session-level empathy ratings; Gibson et al., 2016). Specific, prompt feedback on all three of these may be useful for therapists. In the medical care setting, automated coding methods identified conversation patterns and discussed symptoms. In the education context, automated coding methods successfully predicted the number of questions teachers asked and the types of class activity they set. These automated methods are well tolerated (Skipp & Tanner, 2015). Imel et al. (2019) used automated coding methods to provide prompt feedback on therapists' performance in a laboratory setting. Therapists found the provided feedback representative of their performance and easy to

be understood. Psychologists were shown to be more receptive to computerised feedback than from a supervisor (Hirsch et al., 2018; Imel et al., 2019). We are aware of only a few commercially available tools for assessing the fidelity of psychosocial interventions. For example, Atkins and colleagues deployed models (Imel et al., 2015; Tanana et al., 2016; Xiao et al., 2015) for automatic coding of therapy sessions including CBT and motivational interviewing (Tanana, 2021). However, the dearth of publicly available tools reveals an opportunity for better collaboration between research and industry and improved knowledge translation.

From a research perspective, machine learning may allow for more affordable, reliable, scalable assessments of treatment fidelity. There is a substantial outlay in the initial annotation of therapy transcripts, but once this annotation is complete for a large trial, the data can be easily used to assess fidelity in other trials. The heterogeneity in fidelity assessment tools does add another level of difficulty for many modalities, like cognitive behavioural therapy, acceptance and commitment therapy, or interpersonal psychotherapy. If studies continue to use different assessments of treatment fidelity, then the generalisability of the machine learning models will be small. If the research community for each of these therapies agreed upon a set of core principles of change that were observable in therapy, then more annotated data would be available to train automated fidelity assessments for these therapies. In health, a number of Delphi studies have been conducted that allowed experts to reach consensus on both a-theoretical and theory-driven strategies (Michie et al., 2013; Teixeira et al., 2020). Using these taxonomies, or more consistent use of a smaller number of fidelity assessment (e.g., Motivational Interviewing Skill Code; Miller et al., 2003; Motivational Interviewing Treatment Integrity; Moyers, Martin, et al., 2005), does lay the platform for machine learning methods of automated coding.

This research, however, needs to be careful to build models that perform well on future data, not just the data included in the original study. Assessing model fit on new data is a primary difference between predictive methods (i.e., machine learning) and more traditional explanatory modelling in research contexts (Breiman, 2001). Decision-rules that work in one dataset may not work with future data. For example, Google Flu Trends was able to predict historial flu rates from their search data, but it failed to accurately predict future data because methods became too sensitive to noise in the historical data (Lazer et al., 2014). To avoid these traps, machine learning experts identified a set of best-practice guidelines (Luo et al., 2016), which we used to evaluate studies. Our review found that few studies met these criteria. For example, guidelines recommend using a section of available data to refine the method (e.g., 70% of participants), but new data (e.g., 30% of participants), not used to refine the method, should be used for testing the final method (Luo et al., 2016; Yarkoni & Westfall, 2017). Only 21.1% of studies tested their methods on hold-out data. This is despite testing methods on novel data being an essential measure of method performance in machine learning. Six studies (11.1%) did not report how they refined their method at all (i.e., the validation process). Without transparently reporting these processes, readers cannot assume that machine learning methods will work on future data. Similarly, 46.2% of studies did not report if or how they undertook pre-processing of data. Pre-processing involves the cleaning and rescaling of data which usually occurs before training the method (García et al., 2014). Without these details, methods are not reproducible. While the general conditions of the studies were reported (e.g., where authors got the data and how much data they had), future predictive methods will be more useful, accurate, and generalisable if studies adhere to best-practice guidelines.

Limitations

The studies in this review used a wide variety of accuracy measures, behavioural coding measures, and outcomes which made it difficult to compare the methods. We could have calculated a common metric with a confusion matrix. Confusion matrices represent the predictive results of each code in utterance level (i.e., how many utterances predicted correctly or incorrectly), but only nine studies (three studies in psychotherapy and six studies in education) reported such a matrix. Another limitation was that treatment is a collaborative dialogue, but we only analysed the helping professionals' language. Some studies analysed both helping professionals' and clients' language, and methods that predicted both may be useful for clinicians and researchers to assess fidelity (e.g., did the technique produce the desired outcome). Also, predictive performance of a method might be different when analysing the clients' language, so future reviews could assess the methods used to automatically annotate client/patient language. Similarly, we excluded studies that only focused on signal-processing models of para-verbal behaviour, or object-classification models of non-verbal behaviour from video. Both non-verbal and para-verbal behaviour are important components of therapy, particularly with respect to common factors like therapeutic alliance. Future reviews may want to assess whether models involving those features perform well in therapeutic environments. We also excluded studies that exclusively coded patient behaviour, though many patient behaviours (e.g., change-talk in motivational interviewing; Tanana et al., 2016) are indicators of therapist fidelity. Reviews that focus on patient indicators of quality therapy may be helpful complements to our review here. We included a broad range of helping professions to try and promote knowledge crossover between related fields; however, doing so may mean approaches described here do not generalise. The models that have been used in education or medicine might not perform equally well in other settings and vice versa. Even within the field of psychotherapy, models that work well on one therapeutic intervention (e.g., motivational interviewing) may not perform well for other interventions (e.g., cognitive-behaviour therapy).

Finally, our search may have missed some grey literature or publications in other languages. While we searched our chosen databases for grey literature, we did not systematically search other websites for potential papers to include. Similarly, while we did not exclude any full-texts on the basis of language, our search terms were in English, meaning we may have missed important contributions that were indexed in other languages. The authorship team of this systematic review are fluent in the other languages (e.g., German, Mandarin) and when automated translation tools (e.g., Google Translate) did not suffice, those authors helped with full-text screening. In the cases where our authorship team was not able to read the full-text, we got help from other members of our institute who were fluent in that language. However, we used comprehensive search terms and MeSH headings, ran the search in the major databases, did forward and backward searching, and sent enquiry emails to related researchers. Still, the techniques encompassing 'machine learning' with researchers around the world are often shared without peer review, so it is possible we missed some papers that may have been eligible.

Conclusions

The results of this systematic review have implications for both research and practice. While more work is needed to reveal what methods work best in which circumstances, our systematic review showed that machine learning is a promising tool for assessing treatment fidelity, promoting best-practice in psychological interventions (Bellg et al., 2004). Therefore, organisations and

agencies may be able to use these methods to provide prompt feedback, conduct research, and scale up training to improve therapists' work. We have also shown that automated methods are most likely to be accurate on session level prediction with larger datasets, fewer number of codes and conceptually concrete codes. Finally, we provided recommendations for a minimal list of considerations when developing generalisable machine learning models for treatment fidelity. In sum, machine learning shows promise as a way of decreasing barriers to assessment and feedback for treatment fidelity. Doing so can improve scientific progress by improving the consistency of interventions being studied, but also improve service delivery, ensuring clients receive effective treatments that have been validated through rigorous research.

Supplementary data

Supplementary data are available at https://doi.org/10.5093/pi2021a4

Conflict of Interest

The authors of this article declare no conflict of interest.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research: Journal of the Society for Psychotherapy Research, 31*(1), 92-116. https://doi.org/10.1080/10503307.2020.1808729

Abou Omar, K. B. (2018). *XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. Preprint Semester Project.* https://pub.tik.ee.ethz.ch/students/2017-HS/SA-2017-98.pdf

Adamou, M., Antoniou, G., Greasidou, E., Lagani, V., Charonyktakis, P., & Tsamardinos, I. (2018). Mining free-text medical notes for suicide risk assessment. *Proceedings of the 10th Hellenic Conference on Artificial Intelligence, e47*, 1-8. https://doi.org/10.1145/3200947.3201020

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 for Computational Linguistics, 4,* 463-476. https://doi.org/10.1162/tacl_a_00111

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PloS One, 7*(6). https://doi.org/10.1371/journal.pone.0038014

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology: JFP: Journal of the Division of Family Psychology of the American Psychological Association, 26*(5), 816-827. https://doi.org/10.1037/a0029607

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: IS, 9*(1), 49. https://doi.org/10.1186/1748-5908-9-49

Babineau, J. (2014). Product review: Covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association des Bibliothèques de la Santé du Canada, 35*(2), 68-71. https://journals.library.ualberta.ca/jchla/index.php/jchla/article/view/22892/17064

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences.* Cambridge University Press. https://play.google.com/store/books/details?id=_yOZAKivQfIC

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., Jüni, P., & Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis. *PLoS Medicine, 10*(5), e1001454. https://doi.org/10.1371/journal.pmed.1001454

Barwick, M. A., Bennett, L. M., Johnson, S. N., McGowan, J., & Moore, J. E. (2012). Training health and mental health professionals in motivational interviewing: A systematic review. *Children and Youth Services Review, 34*(9), 1786-1795. https://doi.org/10.1016/j.childyouth.2012.05.012

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D., Czajkowski, S., & Treatment Fidelity Workgroup of the NIH Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association, 23*(5), 443-451. https://doi.org/10.1037/0278-6133.23.5.443

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology, 188*(12), 2222-2239. https://doi.org/10.1093/aje/kwz189

Blanchard, N., Donnelly, P., Olney, A., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016a). Identifying teacher questions using automatic speech recognition in classrooms. *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 191-201. https://www.aclweb.org/anthology/W16-3623.pdf

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016b). Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. *Proceedings of the 9th International Conference on Educational Data Mining.* https://eric.ed.gov/?id=ED592742

Blanck, P., Perleth, S., Heidenreich, T., Kröger, P., Ditzen, B., Bents, H., & Mander, J. (2018). Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: Systematic review and meta-analysis. *Behaviour Research and Therapy, 102*, 25-35. https://doi.org/10.1016/j.brat.2017.12.002

Borrelli, B. (2011). The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. *Journal of Public Health Dentistry, 71*(s1), S52-S63. https://doi.org/10.1111/j.1752-7325.2011.00233.x

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 16*(3), 199-231. https://doi.org/10.1214/ss/1009213726

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners.* ArXiv:2005.14165v.4 [cs.CL]. http://arxiv.org/abs/2005.14165

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*(1), 108-132. https://doi.org/10.1006/jmps.1999.1279

Can, D., Atkins, D. C., & Narayanan, S. S. (2015). *A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations.* Sixteenth Annual Conference of the International Speech Communication Association. https://188.166.204.102/archive/interspeech_2015/papers/i15_0339.pdf

Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2012). A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. *INTERSPEECH-2012*, 2251-2254. https://www.isca-speech.org/archive/interspeech_2012/i12_2254.html

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology, 63*(3), 343-350. https://doi.org/10.1037/cou0000111

Cao, J., Tanana, M., Imel, Z., Poitras, E., Atkins, D., & Srikumar, V. (2019). Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5599-5611. https://doi.org/10.18653/v1/P19-1563

Carcone, A. I., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K., & Kotov, A. (2019). Developing machine learning models for behavioral coding. *Journal of Pediatric Psychology, 44*(3), 289-299. https://doi.org/10.1093/jpepsy/jsy113

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research: JMLR, 11*, 2079–2107. http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Chakravarthula, S. N., Gupta, R., Baucom, B., & Georgiou, P. (2015). A language-based generative model framework for behavioral analysis of couples' therapy. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2090-2094. https://doi.org/10.1109/ICASSP.2015.7178339

Chen, Z., Singla, K., Gibson, J., Can, D., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. (2019). Improving the prediction of therapist behaviors in addiction counseling by exploiting class confusions. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6605 -6609. https://doi.org/10.1109/ICASSP.2019.8682885

Delgadillo, J. (2012). Machine learning: A primer for psychotherapy researchers. *Psychotherapy Research, 31*(1), 1-4.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87. https://doi.org/10.1145/2347736.2347755

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference,* 218-227. https://doi.org/10.1145/3027385.3027417

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy, 49*(6-7), 373-378. https://doi.org/10.1016/j.brat.2011.03.005

Fairburn, C. G., & Patel, V. (2017). The impact of digital technology on psychological treatments and their dissemination. *Behaviour Research and Therapy, 88*, 19-25. https://doi.org/10.1016/j.brat.2016.08.012

Flemotomos, N., Martinez, V. R., Gibson, J., Atkins, D. C., Creed, T., & Narayanan, S. S. (2018). Language features for automated evaluation of cognitive behavior psychotherapy sessions. *INTERSPEECH*, 1908-1912.

Gallo, C., Pantin, H., Villamar, J., Prado, G., Tapia, M., Ogihara, M., Cruden, G., & Brown, C. H. (2015). Blending qualitative and computational linguistics methods for fidelity assessment: Experience with the Familias Unidas Preventive Intervention. *Administration and Policy in Mental Health, 42*(5), 574-585. https://doi.org/10.1007/s10488-014-0538-4

García, S., Luengo, J., & Herrera, F. (2014). *Data preprocessing in data mining.* Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., & Smyth, P. (2017). Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics, 21*(2), 476-487. https://doi.org/10.1109/JBHI.2015.2503985

Gibson, J., Can, D., Georgiou, P., Atkins, D., C., & Narayanan, S. S. (2017). Attention networks for modeling behaviors in addiction counseling. *Interspeech-2017,* 3251-3255. https://doi.org/10.21437/Interspeech.2017-218

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. S. (2016). A deep learning approach to modeling empathy in addiction counseling. *Interspeech-2016, 2016*, 1447-1451. https://doi.org/10.21437/Interspeech.2016-554

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., Imel, Z. E., Narayanan, S. S., & Atkins, D. C. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology, 67*(4), 438-448. https://doi.org/10.1037/cou0000382

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63*(1), 1-11. https://doi.org/10.1037/cou0000131

Golin, C. E., Liu, H., Hays, R. D., Miller, L. G., Beck, C. K., Ickovics, J., Kaplan, A. H., & Wenger, N. S. (2002). A prospective study of predictors of adherence to combination antiretroviral medication. *Journal of General Internal Medicine, 17*(10), 756-765. https://doi.org/10.1046/j.1525-1497.2002.11214.x

Graf, E.-M., Sator, M., & Spranz-Fogasy, T. (2014). *Discourses of helping professions.* John Benjamins Publishing Company. https://play.google.com/store/books/details?id=hKK2BQAAQBAJ

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ, 331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Gupta, R., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2014). Predicting client's inclination towards target behavior change in motivational interviewing and investigating the role of laughter. *Fifteenth Annual Conference of the International Speech Communication Association.* https://www.isca-speech.org/archive/interspeech_2014/i14_0208.html

Haerens, L., Aelterman, N., Van den Berghe, L., De Meyer, J., Soenens, B., & Vansteenkiste, M. (2013). Observing physical education teachers' need-supportive interactions in classroom settings. *Journal of Sport & Exercise Psychology, 35*(1), 3-17. https://doi.org/10.1123/jsep.35.1.3

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *AJR. American Journal of Roentgenology, 212*(1), 38-43. https://doi.org/10.2214/AJR.18.20224

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer. http://thuvien.thanglong.edu.vn:8081/dspace/handle/DHTL_123456789/4053

Hausner, E., Guddat, C., Hermanns, T., Lampert, U., & Waffenschmidt, S. (2016). Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *Journal of Clinical Epidemiology, 77,* 118-124. https://doi.org/10.1016/j.jclinepi.2016.05.002

Hernandez-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research, 13,* 2813-2869. https://www.jmlr.org/papers/volume13/hernandez-orallo12a/hernandez-orallo12a.pdf

Hinde, S., & Spackman, E. (2015). Bidirectional citation searching to completion: An exploration of literature searching methods. *PharmacoEconomics, 33*(1), 5-11. https://doi.org/10.1007/s40273-014-0205-3

Hirsch, T., Soma, C., Merced, K., Kuo, P., Dembe, A., Caperton, D. D., Atkins, D. C., & Imel, Z. E. (2018). " It's hard to argue with a computer": Investigating psychotherapists' attitudes towards automated evaluation. *Proceedings of the 2018 Designing Interactive Systems Conference*, 559-571. https://dl.acm.org/doi/abs/10.1145/3196709.3196776

Howard, M., Agarwal, G., & Hilts, L. (2009). Patient satisfaction with access in two interprofessional academic family medicine clinics. *Family Practice, 26*(5), 407-412. https://doi.org/10.1093/fampra/cmp049

Howes, C., Purver, M., & McCabe, R. (2013). Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical Informatics Insights, 6*(Suppl 1), 39–50. https://doi.org/10.4137/BII.S11661

Imel, Z. E., Pace, B. T., Soma, C. S., Tanana, M., Hirsch, T., Gibson, J., Georgiou, P., Narayanan, S., & Atkins, D. C. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy, 56*(2), 318-328. https://doi.org/10.1037/pst0000221

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy, 52*(1), 19-30. https://doi.org/10.1037/a0036841

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., O'Brien, M. A., Johansen, M., Grimshaw, J., & Oxman, A. D. (2012). Audit and feedback: Effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews, 6*, CD000259. https://doi.org/10.1002/14651858.CD000259.pub3

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *The American Psychologist, 64*(6), 515-526. https://doi.org/10.1037/a0016755

Kazantzis, N. (2003). Therapist competence in cognitive-behavioural therapies: Review of the contemporary empirical evidence. *Behaviour Change, 20*(1), 1-12. https://doi.org/10.1375/bech.20.1.1.24845

Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy, 88,* 7-18. https://doi.org/10.1016/j.brat.2016.06.004

Kornhaber, R., Walsh, K., Duff, J., & Walker, K. (2016). Enhancing adult therapeutic interpersonal relationships in the acute health care setting: an integrative review. *Journal of Multidisciplinary Healthcare, 9,* 537-546. https://doi.org/10.2147/JMDH.S116957

Kotov, A., Idalski Carcone, A., Dong, M., Naar-King, S., & Brogan, K. E. (2014). Towards automatic coding of interview transcripts for public health research. *Proceedings of the Big Data Analytic Technology for Bioinformatics and Heath Informatics Workshop (KDD-BHI) in Conjunction with ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* New York, NY.

Lacson, R., & Barzilay, R. (2005). Automatic processing of spoken dialogue in the home hemodialysis domain. *AMIA Annual Symposium Proceedings*, 420-424. https://www.scopus.com/inward/record.uri?eid=2-s2.0-39049191868&partnerID=40&md5=c99769d457f6ce7d55d48949fa865b04

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363-374. https://www.ncbi.nlm.nih.gov/pubmed/884196

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203-1205. https://doi.org/10.1126/science.1248506

Lecomte, T., Kingdon, D., and Munro-Clark, D. (2017). *Cognitive therapy scale for psychosis (revised version).* Manuscript in preparation.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*(1), 96–116. https://doi.org/10.1002/lio2.354

Lum, K. (2017). *Limitations of mitigating judicial bias with machine learning machine-learning algorithms trained with data that encode human bias will reproduce, not eliminate, the bias, says Kristian Lum.* Nature Publishing Group.

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T. B., Venkatesh, S., & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research, 18*(12), e323. https://doi.org/10.2196/jmir.5870

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., & Takwoingi, Y. (2010). *Cochrane handbook for systematic reviews of diagnostic test accuracy* (Version 0. 9. 0). The Cochrane Collaboration. http://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter%2010%20-%20Version%201.0.pdf

Madson, M. B., Loignon, A. C., & Lane, C. (2009). Training in motivational interviewing: A systematic review. *Journal of Substance Abuse Treatment, 36*(1), 101-109. https://doi.org/10.1016/j.jsat.2008.05.005

Malandrakis, N., & Narayanan, S. S. (2015). Therapy language analysis using automatically generated psycholinguistic norms. *Sixteenth Annual Conference of the International Speech Communication Association.* https://www.scopus.com/inward/record.uri?eid=2-s2.0-84959123480&partnerID=40&md5=48c4f18021e9a3d03bf0493b5f6ee649

Mayfield, E., Laws, M. B., Wilson, I. B., & Penstein Rosé, C. (2014). Automating annotation of information-giving for analysis of clinical conversation.

Journal of the American Medical Informatics Association: JAMIA, 21(e1), e122-e128. https://doi.org/10.1136/amiajnl-2013-001898
McGlinchey, J. B., & Dobson, K. S. (2003). Treatment integrity concerns in cognitive therapy for depression. *Journal of Cognitive Psychotherapy, 17*(4), 299-318. https://doi.org/10.1891/jcop.17.4.299.52543
Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283-297. https://doi.org/10.1007/s11217-007-9071-1
Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine, 46*(1), 81-95. https://doi.org/10.1007/s12160-013-9486-6
Mieskes, M., & Stiegelmayr, A. (2019). Preparing data from psychotherapy for natural language processing. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2896-2902. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059902190&partnerID=40&md5=f539dde0dce50ce71931c25747add595
Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2003). *Manual for the motivational interviewing skill code (MISC).* Unpublished manuscript. Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico, Albuquerque. https://casaa.unm.edu/download/misc.pdf
Miller, W., & Rollnick, S. (1991). *Motivational interviewing: Preparing people to change addictive behaviour.* Guilford Press.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097
Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment, 28*(1), 19-26. https://doi.org/10.1016/j.jsat.2004.11.001
Moyers, T. B., Miller, W. R., & Hendrickson, S. M. L. (2005). How does motivational interviewing work? Therapist interpersonal skill predicts client involvement within motivational interviewing sessions. *Journal of Consulting and Clinical Psychology, 73*(4), 590-598. https://doi.org/10.1037/0022-006X.73.4.590
Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment, 65*, 36-42. https://doi.org/10.1016/j.jsat.2016.01.001
Ng, A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. *Retrieved Online at* https://www.mlyearning.org
Nitti, M., Ciavolino, E., Salvatore, S., & Gennaro, A. (2010). Analyzing psychotherapy process as intersubjective sensemaking: An approach based on discourse analysis and neural networks. *Psychotherapy Research*: *Journal of the Society for Psychotherapy Research, 20*(5), 546-563. https://doi.org/10.1080/10503301003641886
Nystrand, M., Wu, L. L., Gamoran, A., Zeiser, S., & Long, D. A. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes, 35*(2), 135-198. https://doi.org/10.1207/S15326950DP3502_3
Öst, L.-G., & Ollendick, T. H. (2017). Brief, intensive and concentrated cognitive behavioral treatments for anxiety disorders in children: A systematic review and meta-analysis. *Behaviour Research and Therapy, 97*, 134-145. https://doi.org/10.1016/j.brat.2017.07.008
Park, J., Kotzias, D., Kuo, P., Logan, R. L., Iv, Merced, K., Singh, S., Tanana, M., Karra Taniskidou, E., Lafata, J. E., Atkins, D. C., Tai-Seale, M., Imel, Z. E., & Smyth, P. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association: JAMIA, 26*(12), 1493-1504. https://doi.org/10.1093/jamia/ocz140
Parsons, J. T., Rosof, E., Punzalan, J. C., & Di Maria, L. (2005). Integration of motivational interviewing and cognitive behavioral therapy to improve HIV medication adherence and reduce substance use among HIV-positive men and women: results of a pilot project. *AIDS Patient Care and STDs, 19*(1), 31-39. https://doi.org/10.1089/apc.2005.19.31
Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language. use: our words, our selves. *Annual Review of Psychology, 54*, 547-577. https://doi.org/10.1146/annurev.psych.54.101601.145041
Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology, 77*(2), 212-218. https://doi.org/10.1037/a0015232
Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*(4), 365-383. https://doi.org/10.1093/clipsy.bpi045
Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829-841. https://doi.org/10.1037/0022-006X.75.6.829
Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., An, L., Goggin, K. J., & Catley, D. (2017). Predicting counselor behaviors in motivational interviewing encounters. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1128-1137. https://www.aclweb.org/anthology/E17-1106
Pérez-Rosas, V., Sun, X., Li, C., Wang, Y., Resnicow, K., & Mihalcea, R. (2019). Analyzing the quality of counseling conversations: The tell-tale signs of high-quality counselling. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059886164&partnerID=40&md5=dac1c5b64d2f82e022ecb542d991b9db
Perlich, C. (2009). *Learning curves in machine learning* (No. RC24756). IBM Research Division. https://domino.research.ibm.com/library/cyberdig.nsf/papers/491B767CE4518A4585257576006BCD2D/$File/rc24756.pdf
Prado, G., Pantin, H., Schwartz, S. J., Schwartz, S. J., Lupei, N. S., & Szapocznik, J. (2006). Predictors of engagement and retention into a parent-centered, ecodevelopmental HIV preventive intervention for Hispanic adolescents and their families. *Journal of Pediatric Psychology, 31*(9), 874–890. https://doi.org/10.1093/jpepsy/jsj046
Prowse, P.-T. D., & Nagel, T. (2015). A meta-evaluation: The role of treatment fidelity within psychosocial interventions during the last decade. *Journal of Psychiatry, 18*(2). https://doi.org/10.4172/psychiatry.1000251
Prowse, P.-T. D., Nagel, T., Meadows, G. N., & Enticott, J. C. (2015). Treatment fidelity over the last decade in psychosocial clinical trials outcome studies: A systematic review. *Journal of Psychiatry, 18*(2). https://doi.org/10.4172/psychiatry.1000258
Reiser, R. P., & Milne, D. L. (2014). A systematic review and reformulation of outcome evaluation in clinical supervision: Applying the fidelity framework. *Training and Education in Professional Psychology, 8*(3), 149-157. https://doi.org/10.1037/tep0000031
Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(3), 569 -575. https://doi.org/10.1109/TPAMI.2009.187
Rodriguez-Quintana, N., & Lewis, C. C. (2018). Observational coding training methods for CBT treatment fidelity: A systematic review. *Cognitive Therapy and Research, 42*(4), 358-368. https://doi.org/10.1007/s10608-018-9898-5
Russell, S., & Norvig, P. (2002). *Artificial intelligence: A modern approach.* https://research.google/pubs/pub27702.pdf
Ryan, P., Luz, S., Albert, P., Vogel, C., Normand, C., & Elwyn, G. (2019). Using artificial intelligence to assess clinicians' communication skills. *BMJ, 364*, l161. https://doi.org/10.1136/bmj.l161
Salvatore, S., Gennaro, A., Auletta, A. F., Tonti, M., & Nitti, M. (2012). Automated method of content analysis: A device for psychotherapy process research. *Psychotherapy Research: Journal of the Society for Psychotherapy Research, 22*(3), 256-273. https://doi.org/10.1080/10503307.2011.647930
Samei, B., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. (2014, July). *Domain independent assessment of dialogic properties of classroom discourse*. Seventh International Conference on Educational Data Mining. London. https://eric.ed.gov/?id=ED566380
Schwalbe, C. S., Oh, H. Y., & Zweben, A. (2014). Sustaining motivational interviewing: A meta-analysis of training studies. *Addiction, 109*(8), 1287-1294. https://doi.org/10.1111/add.12558
Sen, T., Ali, M. R., Hoque, M. E., Epstein, R., & Duberstein, P. (2017). Modeling doctor-patient communication with affective text analysis. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).* https://doi.org/10.1109/acii.2017.8273596
Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(9), 1426-1448. https://doi.org/10.1017/S0033291719000151
Shmueli, G. (2010). To explain or to predict? *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 25*(3), 289-310. https://doi.org/10.1214/10-STS330
Singla, K., Chen, Z., Flemotomos, N., Gibson, J., Can, D., Atkins, D., & Narayanan, S. (2018). Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. *Interspeech 2018*, 3413-3417. https://doi.org/10.21437/Interspeech.2018-2551
Skipp, A., & Tanner, E. (2015). *The visible classroom: Evaluation report and executive summary.* Education Endowment Foundation. https://eric.ed.gov/?id=ED581106
Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87e100. https://doi.org/10.1207/S15326985EP3502_3
Song, Y., Lei, S., Hao, T., Lan, Z., & Ding, Y. (2020). Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research*, 0735633120968554. https://doi.org/10.1177/0735633120968554
Street, R. L., Jr, Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician-patient communica-

tion to health outcomes. *Patient Education and Counseling, 74*(3), 295-301. https://doi.org/10.1016/j.pec.2008.11.015

Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers' classroom discourse. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 9721-9728. https://doi.org/10.1609/aaai.v33i01.33019721

Tai, A. M. Y., Albuquerque, A., Carmona, N. E., Subramanieapillai, M., Cha, D. S., Sheko, M., Lee, Y., Mansur, R., & McIntyre, R. S. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artificial Intelligence in Medicine, 99,* 101704. https://doi.org/10.1016/j.artmed.2019.101704

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. https://doi.org/10.1016/j.jsat.2016.01.006

Tanana, M. (2021, November 3). *Predicting CBT fidelity like a human.* Lyssn | Intelligent Counselling Recording Platform. https://www.lyssn.io/blog/predicting-cbt-fidelity-like-a-human

Teixeira, P. J., Marques, M. M., Silva, M. N., Brunet, J., Duda, J., Haerens, L., La Guardia, J., Lindwall, M., Londsdale, C., Markland, D., Michie, S., Moller, A. C., Ntoumanis, N., Patrick, H., Reeve, J., Ryan, R. M., Sebire, S., Standage, M., Vansteenkinste, M., … Hagger, M. S. (2020). Classification of techniques used in self-determination theory-based interventions in health contexts: An expert consensus study. *Motivation Science, 6*(4), 438-455. https://doi.org/10.1037/mot0000172

Tracey, T. J. G., Wampold, B. E., Lichtenberg, J. W., & Goodyear, R. K. (2014). Expertise in psychotherapy: An elusive goal? *The American Psychologist, 69*(3), 218-229. https://doi.org/10.1037/a0035099

Velasquez, P. A. E., & Montiel, C. J. (2018). Reapproaching Rogers: A discursive examination of client-centered therapy. *Person-Centered and Experiential Psychotherapies, 17*(3), 253-269. https://doi.org/10.1080/14779757.2018.1527243

Wallace, B. C., Laws, M. B., Small, K., Wilson, I. B., & Trikalinos, T. A. (2014). Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Medical Decision Making: An International Journal of the Society for Medical Decision Making, 34*(4), 503-512. https://doi.org/10.1177/0272989X13514777

Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy, 47*(2), 119-127. https://doi.org/10.1016/j.brat.2008.10.018

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy, 77*, 129-137. https://doi.org/10.1016/j.brat.2015.12.005

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. *Computers & Education, 78*, 115-123. https://doi.org/10.1016/j.compedu.2014.05.010

Xiao, B., Huang, C. Imel, Z. C., 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*(4). https://doi.org/10.7717/peerj-cs.59

Xiao, B., Imel, Z. C., Georgiou, P., 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.https://doi.org/10.1371/journal.pone.0143055

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393