Journal of the American Medical Informatics Association, 27(6), 2020, 929–933
doi: 10.1093/jamia/ocaa049
Advance Access Publication Date: 6 May 2020
Brief Communication

Brief Communication

Spoken words as biomarkers: using machine learning to gain insight into communication as a predictor of anxiety

George Demiris [1], **Kristin L. Corey Magan**,[1] **Debra Parker Oliver**,[2] **Karla T. Washington**,[2] **Chad Chadwick**,[3] **Jeffrey D. Voigt**,[3] **Sam Brotherton**,[3] **and Mary D. Naylor**[1]

[1]School of Nursing, University of Pennsylvania, Philadelphia, Pennsylvania, USA, [2]Family Medicine, School of Medicine, University of Missouri, Columbia, Missouri, USA, and [3]Live Circle Inc, Ridgewood, New Jersey, USA

Corresponding Author: George Demiris, PhD, FACMI, School of Nursing, University of Pennsylvania, 418 Curie Blvd, Rm 324, Philadelphia PA 19104, USA (gdemiris@upenn.edu)

Received 10 December 2019; Revised 3 March 2020; Editorial Decision 30 March 2020; Accepted 3 April 2020

ABSTRACT

Objective: The goal of this study was to explore whether features of recorded and transcribed audio communication data extracted by machine learning algorithms can be used to train a classifier for anxiety.
Materials and Methods: We used a secondary data set generated by a clinical trial examining problem-solving therapy for hospice caregivers consisting of 140 transcripts of multiple, sequential conversations between an interviewer and a family caregiver along with standardized assessments of anxiety prior to each session; 98 of these transcripts (70%) served as the training set, holding the remaining 30% of the data for evaluation.
Results: A classifier for anxiety was developed relying on language-based features. An 86% precision, 78% recall, 81% accuracy, and 84% specificity were achieved with the use of the trained classifiers. High anxiety inflections were found among recently bereaved caregivers and were usually connected to issues related to transitioning out of the caregiving role. This analysis highlighted the impact of lowering anxiety by increasing reciprocity between interviewers and caregivers.
Conclusion: Verbal communication can provide a platform for machine learning tools to highlight and predict behavioral health indicators and trends.
Key words: caregivers, anxiety, machine learning, communication, behavioral research

INTRODUCTION

Computational techniques have emerged to that enable the use of speech and language cues (a largely untapped data source) to gain insight into human behavior and health over time. Narayanan and Georgiou[1] use the term *behavioral signal processing* to refer to a suite of computational tools that enable automated measurement of communication behaviors (eg, recorded conversations) with the goal of extracting mathematical quantities from the continuous audio recording to measure and model specific behavioral markers. Baucom et al[2] highlighted the potential of recorded couple interactions to assess behavioral and cognitive markers of risk for suicide among US Army National Guard personnel. Using transcripts of recorded interviews, automatic speech analysis also has been utilized to establish predictors of psychotic disorders in young adults.[3] Konig and colleagues[4] demonstrated that automatic speech analysis is a viable approach to detecting the presence of early-stage dementia. Similarly, Nasir et al[5] conducted a series of experiments to demonstrate that predictions of relationship outcomes obtained directly from vocal acoustics based on recordings of conversations by couples in distressed relations were comparable to those obtained using human-related behavioral codes.

These studies, while still in an exploratory stage, demonstrate the potential of creating powerful tools that can provide interpretable communication feedback in real time, allowing the detection of communication behaviors of individuals at greater risk for adverse

outcomes. One such population includes family caregivers of persons with a serious illness, where the stress and demands on the caregiver over time have been found to affect caregivers' own mortality and morbidity.[6] Family caregivers represent a significant population of care providers; there are over 40 million family caregivers in the US providing 37 billion hours of care with close to $500 billion in unpaid contributions to care.[7] Healthcare costs for caregivers have been found to be 8%–10% higher than for noncaregivers.[8] Caregivers of patients at the end of life, more specifically, often report feeling isolated and rate their communication with healthcare providers as poor. This includes an inability to receive meaningful support which, along with the patient's health deterioration and anticipatory grief, increases their anxiety.[9]

The purpose of this study is to demonstrate a proof of concept as to whether features of recorded and transcribed audio communication data extracted by machine learning algorithms can be paired with standardized behavioral health assessment. The data set we used for our proof of concept constituted a series of transcribed conversations with family caregivers of hospice patients as they discuss challenges they face in their caregiving role over time, combined with standardized assessments of self-reported anxiety levels. We hypothesized that positive sentiment in these conversations might be correlated with lower anxiety scores and that, if the subject's language mirrored that of the interviewer's, this might suggest a deeper connection between the interviewer and subject which might lead to improved outcomes. This study is a first step in creating and optimizing a tool that can provide real-time diagnostic support and clinical feedback based on dyadic communication sessions.

METHODS

Sample

We used a secondary data set generated by a randomized clinical trial examining the role of problem-solving therapy for hospice caregivers. The intervention tested was called PISCES (Problem solving Intervention to Support Caregivers in End of life care Settings). Two large hospice agencies in the Pacific Northwest participated in the project. Inclusion criteria for caregivers were the following: a) a family/informal caregiver of a hospice patient of 1 of the participating hospice agencies; b) 18 years or older; c) access to a standard phone line or computer with Internet access; d) sufficient hearing for telephone or Internet conversations as assessed by the research staff; e) proficient in English with at least a 6th grade education. The 3-session protocol for PISCES involved 5 steps: adopting a positive attitude, defining the problem, creating alternatives, predicting consequences, and trying a solution. Each session lasted approximately 45 minutes. The main findings of this clinical trial and details of the intervention protocol are provided elsewhere.[10]

PISCES sessions were audio-recorded and transcribed. These transcripts of multiple, sequential conversations between an interviewer (interventionist) and the subject (family caregiver) provided the training data set for this study. For each subject, there were between 1 and 4 transcripts of sessions typically spaced apart by 1–3 weeks. Family caregivers' anxiety was measured using the Generalized Anxiety Disorder 7 (GAD-7) item scale. The GAD-7 is a brief, valid tool for screening for anxiety and assessing its severity.[11] The GAD-7 was administered twice (pre- and postintervention). The GAD-7 total score ranges from 0 to 21. For the purposes of training a binary classifier, the GAD instrument values were divided into "low anxiety" (0–9) and "moderate to high anxiety" (10+).

There were 514 family caregivers recruited for the study. Caregivers were predominantly female (75%); mean age was 60.3 years. A subset of all recorded sessions (N = 140) with 49 caregivers was professionally transcribed; 98 of these transcripts (70%) were selected to serve as the training set, with the remaining 30% of the data held for evaluation. For each caregiver there were up to 5 GAD-7 assessments (1 at baseline, 1 prior to each of the 3 sessions and 1 at follow-up). Since GAD-7 scores were linked to each of the conversations, we used these labels to train a classifier. The classifier was then applied to a rolling window across a given conversation to approximate GAD-7 score at any given time in the conversation.

The process of classification

The next step was to automate the GAD-7 scoring for the PISCES intervention, replacing hand-coding by trained humans with a machine learning-powered classifier. The classifier relied on language-based features to automate the prediction of pre- and postclinical trial anxiety levels, and to predict real-time anxiety levels during the trial, using the actual utterances of the family caregiver during a conversation. Before extracting features from the text, each line of the transcript was preprocessed. This process involved normalizing the line's text (converting to lowercase and removing extraneous whitespace) and tagging the line with its speaker ("interviewer" or "subject," which was denoted in a variety of ways by the transcribers). This speaker tag was then included in all features; for example, a feature X would be encoded as interviewer: X if it was extracted from an interviewer's line.

The Natural Language Toolkit (NLTK) word tokenize function was used to perform tokenization, namely the process of converting a string of text into words ("tokens"), as part of the classification process.[12] A basic tokenizer could consist of simply splitting a sentence on each space; the NLTK word tokenize function is punctuation-aware and tuned for the English language. Examples of token features include *interviewer: hospice* or *subject: brother*. Further, all tokens were dropped that occurred only once in the training set. All tokens that were stop words were also dropped (ie, very common words to be ignored by the computer such as "the," "a," etc).

Combinations of consecutive tokens were also retained in order to represent compound words and longer expressions.[13] For example, the entity "heart attack" is not expressible as a single token. To include phrases like this, features were added for all bigrams (2 consecutive words) and trigrams (3 consecutive words) that showed up at least twice in the training set and did not begin or end with a stop word. Examples included *subject: heart attack* or *interviewer: good thing*.

Sentiment analysis

A sentiment score of each line as a feature was included. The sentiment polarity was computed on a scale from −1.0 (extremely negative) to 1.0 (extremely positive) using the TextBlob library.[14] If the sentiment score of a line was over 0.3, the feature *subject: sentiment: pos* or *interviewer: sentiment: pos* was assigned to that line. If the sentiment score was below −0.3, the line was assigned the feature *subject: sentiment: neg* or *interviewer: sentiment: neg*.

Reciprocity score

To capture how the subject's language may mirror that of the interviewer, a "reciprocity score" was computed for each pair of consecutive utterances from different speakers. The reciprocity score is a

Table 1. Confusion matrix

	Predicted low anxiety	Predicted high anxiety
Actual low anxiety (GAD-7 score 0–9)	18 (TP)	5 (FN)
Actual High Anxiety (GAD-7 score 10 or higher)	3 (FP)	16 (TN)

Abbreviations: FN, False Negative; FP, False Positive; TN, True Negative; TP, True Positive.

measure of the similarity between words, sounds, or phrases used by 2 different speakers in sequential turns of a conversation. We used GloVe word vectors (mappings from each English word to a 300-dimensional vector in "semantic space," mined from co-occurrence data on the Internet) to quantify the semantic similarity between 2 words or phrases. First, stop words were removed from each utterance, and then the word vector for each remaining word was retrieved.[15] All the individual word vectors were then averaged to create a vector representation of each utterance. The cosine distance between the 2 vectors was computed, scaled between 0 and 1, and designated the reciprocity score value. The validity of using cosine distance in an embedding space to measure text similarity depends largely on how well the embedding space represents the semantic concepts present in the text. In our case, the word embeddings were GloVe vectors trained on Common Crawl. Cosine distance is generally considered a better metric for semantic similarity than Euclidean distance[16,17] since it is robust to differing vector magnitudes. The values we chose as cutoffs for low, medium, and high reciprocity score were the 33rd and 66th percentile of average reciprocity score across a transcript.

Finally, a subject's utterance with the feature *subject: reciprocity: high* was tagged if the reciprocity score between it and the previous interviewer's line was greater than 0.8, and *subject: reciprocity: low* if the score was lower than 0.5. Likewise, interviewers' lines with the features *interviewer: reciprocity: high* or *interviewer: reciprocity: low* were tagged based on the reciprocity score between a given utterance and the previous subject's utterance.

A binary logistic regression classifier was then trained on the features and labels described above, choosing a limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) solver.[18] This is an extension of the BFGS algorithm for solving nonlinear optimization problems commonly used when training machine learning models. In addition to evaluating the classifier on the test set, the classifier also was evaluated in a rolling window of each transcript. The goal of this exercise was to identify the exact points within each conversation where language was used that was most correlated with high or low GAD-7 scores.

RESULTS

Classifier results

When evaluated on the test set, 86% precision, 78% recall (or "sensitivity"), 81% accuracy, 84% specificity, 86% positive predictive value, and 76% negative predictive values were achieved with the use of the trained classifiers. Sensitivity is defined as the performance of an identified classifier in correctly assessing low anxiety in a test data set (as validated by the GAD-7 score); specificity is defined as the performance of an identified classifier in correctly assessing high anxiety in a test data set. A true positive is defined as the case where the classifier predicted high anxiety based on the transcript of a session when the subject's GAD-7 score was rated as high (10 or higher) prior to that session. Similarly, a true negative is defined when the classifier predicted low anxiety when the subject's GAD-7 score was low (0–9).

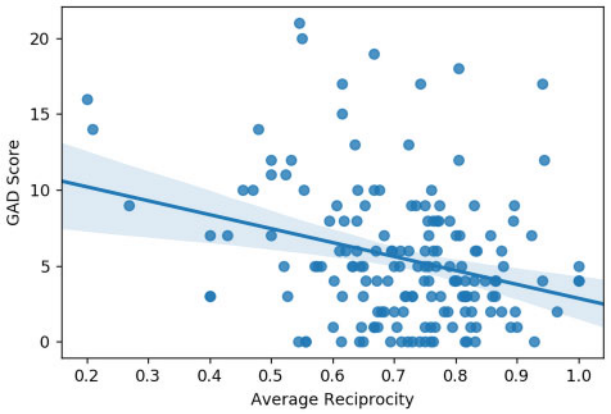

Figure 1. Correlation between GAD-7 score and average reciprocity score.

The confusion matrix, which describes the performance of a classification model (or "classifier") on a set of test data for which the true values are known, is presented in Table 1.

Overall findings

The findings of the machine learning exercise confirmed those of a previously published study on the clinical trial,[10] namely that overall GAD-7 classifier scores decreased following completion of the intervention, suggesting that the intervention had a positive impact on anxiety among caregivers. High anxiety inflections were found among caregivers of patients who had recently died and were usually connected to issues related to transitioning out of the caregiving role. High anxiety inflections encompassed issues related to caregiver stress and burden, negative emotions (eg, anger, guilt), poor self-care, and death of care recipient/bereavement.

Patterns of GAD-7 classifier scores during the intervention varied among individuals, suggesting that responses to the intervention are individualized. Other findings included a potential impact of reciprocity between interventionist and caregiver, on anxiety. The correlation between reciprocity and anxiety was a negative relationship. As average reciprocity increases, average GAD-7 score decreases, with an r-squared value of 0.08. Figure 1 showcases the correlation between reciprocity and GAD-7 score.

The following case presents insights from the classifiers at the individual level, for high-anxiety inflections at the caregiver level.

Case: high-anxiety exemplar

This was a spousal caregiver whose husband had died and who was experiencing multiple stressors including returning to work after being on medical leave, relocating, and acute grief.

An excerpt from this inflection point included: "... it's a really stressful time. It was like boom, as soon as he died, came planning

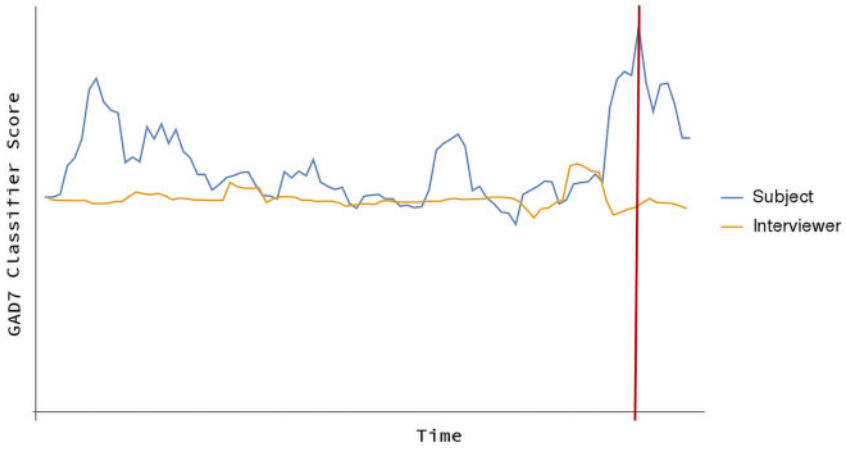

Figure 2. High-anxiety exemplar: trajectory and key inflection point. The high-anxiety inflection point reflects the stressors associated with transitioning out of the caregiver role following the care recipient's death.

Figure 3. World Cloud (from inflection).

for this huge and amazing memorial for him which, you know, I wanted to do it that way. And I channeled everything into that, so that was really stressful. And then finding a place to move and moving in—that was really stressful. I don't have the best relationship with his family, so that's been really stressful. My stepdaughter has just lost her mind with her grief and I'm really worried about her. So, you know it's just really stressful."

DISCUSSION

In this study, we used GAD-7 anxiety assessment scores, linked to conversations between caregivers and interviewers, to train a classifier. The classifier was applied to approximate GAD-7 scores at any given time in the conversation, allowing us to identify these points within a conversation where language was used that most correlated with predicted high or low anxiety.

The high-anxiety inflections found in this machine learning analysis are in alignment with prior published qualitative data that identified caregiver burden, negative emotions, poor self-care, and healthcare system issues as prominent challenges faced by caregivers.[19] This analysis also highlighted anxiety in relation to reciprocity between interventionists and caregivers. Reciprocity occurs when 1 conversation partner repeats a statement made by the other. Reciprocity among caregiver and the therapist responses validates that the caregiver likely understands the information provided by the interventionist. Caregivers who repeat statements made by an interventionist have higher reciprocity (as measured herein), indicating likely increased understanding of the problem-solving process and, therefore, may be adopting problem-solving strategies and thus coping more effectively. This may also be reflective of caregivers actively collaborating with the interventionist which in prior analyses was central to the experience of problem-solving.[20] When therapists validate caregivers' feelings, they also increase reciprocity which may lower anxiety, an observation that would quantify the effect of validation on emotion regulation. Sentiment scores were also valuable measures when considering the impact of the problem-solving intervention. Caregivers who express more positive sentiments likely have a positive attitude toward their life circumstances and may problem-solve and cope more effectively than those with negative sentiment. Adopting a positive, optimistic attitude to problem-solving is 1 of the tenets of the intervention used in this analysis.[21]

Classifiers trained on clinical trial data are important for several reasons. Since the classifier automates the use of the instruments it is based on, these instruments may not have to be administered as separate exercises before or after interventions to deliver clinically rigorous assessments. This automation may mean lower research costs and a reduction of study participant burden. Further, the opportunity for machine learning to incorporate problem-solving techniques and provide 24/7 conversational and ambient artificial intelligence to caregivers could be a helpful coping tool for caregivers. The advantage of systems using machine learning is that they continue to "learn" and improve upon what is known to be effective—which is reflected in the above findings. Further work is being undertaken to develop an artificial intelligence problem-solving coach using the above classifiers.

Study limitations

This study was conceptualized as a proof of concept and an exploratory approach to examining the potential of machine learning. Our sample size, while substantial for behavioral intervention research, provided a limited training data set; additional work with larger data sets is needed. Additionally, all participants were located in a

single region of the United States; a geographically more diverse sample would provide greater insight into the development of classifier approaches.

FUNDING

This study was supported in part by the National Institutes of Health, National Institute for Nursing Research (Grant Nr. R01NR012213; principal investigator: G. Demiris).

AUTHOR CONTRIBUTIONS

GD: conception and design of the work, data collection, data analysis and interpretation, and drafting the article; KC: data collection, data analysis and interpretation, and drafting the article; DO: data collection, data analysis and interpretation, and critical revision of the article; KW: data collection, data analysis and interpretation, and critical revision of the article; CC, JV, SB, and MN: data analysis and interpretation and critical revision of the article. All authors approved this version of the manuscript.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Narayanan S, Georgiou PG. Behavioral signal processing: deriving human behavioral informatics from speech and language. *Proc IEEE* 2013; 101 (5): 1203–31. PP(99)
2. Baucom BR, Georgiou P, Bryan CJ, *et al*. The promise and the challenge of technology-facilitated methods for assessing behavioral and cognitive markers of risk for suicide among U.S. Army National Guard Personnel. *Int J Environ Res Public Health* 2017; 14 (4): 361.
3. Bedi G, Carrillo F, Cecchi GA, *et al*. Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr* 2015; 1 (1): 15030.
4. Konig A, Satt A, Sorin A, *et al*. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement* 2015; 1 (1): 112–24.
5. Nasir M, Baucom BR, Georgiou P, *et al*. Predicting couple therapy outcomes based on speech acoustic features. *PLoS ONE* 2017; 12 (9): e0185123.
6. Schulz R, Beach SR. Caregiving as a risk factor for mortality: the caregiver health effects study. *JAMA* 1999; 282 (23): 2215–9.
7. Reinhard S, Feinberg LF, Choula R, *et al*. Valuing the invaluable 2015 update: undeniable progress, but big gaps remain. https://www.aarp.org/ppi/info-2015/valuing-the-invaluable-2015-update.html Accessed September 30, 2019
8. Feinberg LF, Choula R. Understanding the impact of family caregiving on work. https://www.aarp.org/content/dam/aarp/research/public_policy_institute/ltc/2012/understanding-impact-family-caregiving-work-AARP-ppi-ltc.pdf Accessed September 30, 2019.
9. Washington KT, Parker Oliver D, Smith JB, *et al*. Sleep problems, anxiety, and global self-rated health among hospice family caregivers. *Am J Hosp Palliat Care* 2018; 35 (2): 244–9.
10. Demiris G, Oliver DP, Washington K, *et al*. A problem-solving intervention for hospice family caregivers: a randomized clinical trial. *J Am Geriatr Soc* 2019; 67 (7): 1345–52.
11. Spitzer RL, Kroenke K, Williams JB, *et al*. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006; 166 (10): 1092–7.
12. Tokenizing words and sentences with NLTK. https://pythonspot.com/tokenizing-words-and-sentences-with-nltk/ Accessed September 30, 2019
13. Tokenizing words and sentences with NLTK. https://pythonspot.com/category/nltk/ Accessed September 30, 2019
14. Loria S. TextBlob: simplified text processing. https://textblob.readthedocs.io/en/dev/ Accessed September 30, 2019
15. Pennington J, Socher R. Manning CD. GloVe: global vectors for word representation. https://nlp.stanford.edu/projects/glove Accessed September 30, 2019
16. Tan L, Zhang H, Clarke C, Smucker M. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In: *proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*; July, 2015; Beijing, China.
17. Rothe S, Schütze H. (2016, August). Word embedding calculus in meaningful ultradense subspaces. In: proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; August 7–12, 2016; Berlin, Germany.
18. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Math Program* 1989; 45 (1-3): 503–28.
19. Oliver DP, Demiris G, Washington KT, *et al*. Challenges and strategies for hospice caregivers: a qualitative analysis. *Gerontologist* 2017; 57 (4): 648–56.
20. Washington KT, Demiris G, Oliver DP, *et al*. Qualitative evaluation of a problem-solving intervention for informal hospice caregivers. *Palliat Med* 2012; 26 (8): 1018–24.
21. D'Zurilla TJ, Nezu AM. *Problem-Solving Therapy: A Positive Approach to Clinical Intervention*. 3rd ed. New York: Springer; 2007.