Vocal Pattern Detection of Depression among Older Adults: A Pilot Project

Marianne Smith[1], Bryce J. Dietrich[2*], Er-wei Bai[3] and H. Jeremy Bockholt[4]

1 College of Nursing, University of Iowa, Iowa City, IA, USA
2 Iowa Informatics Initiative (UI3), University of Iowa, Iowa City, IA, USA
3 Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA, USA
4 Advanced Biomedical Informatics Group, Albuquerque, NM, USA

* Corresponding Author (bryce-dietrich@uiowa.edu)

Abstract

Background: Depression is a serious problem for many older adults but is too often undetected by the person, family or providers. Although vocal patterns have been successfully used to detect depression in adults aged 18 to 65 years, no studies to date have included older adults. The purpose of this pilot study is to determine if vocal patterns that have been associated with clinical depression in younger people also signify depression in older adults.

Methods: A total of 46 older adults living in selected senior living settings were recruited to participate in the study. Volunteer participants completed a semi-structured interview composed of a depression scale and selected speech measures that have been successfully used in earlier studies for younger adults. Interviews were audio recorded and recordings were analyzed by machine learning algorithms to evaluate if vocal patterns may predict presence of depression in older adults.

Results: Using the 9-item Patient Health Questionnaire (or PHQ-9) and a supervised machine learning algorithm, we were able to accurately predict high and low depression scores between 86.59 and 92.16 percent of the time. We were also able to predict changes in the raw PHQ-9 scores between interview cycles within 1.17 points while making few assumptions other than audio variables are indicative of depression.

Conclusion: Our results provide strong evidence that vocal patterns can be used effectively to assess clinical depression in adults who are 65 years and older.

Introduction

Late-life depression is a large public health problem that will escalate with population aging. Depression is the leading cause of disability worldwide, and the leading cause of disease burden in the United States [1]. The association between depression and medical illness, anxiety, pain, and functional decline contributes to a downward spiral of disability that threatens independence and quality of life, and is costly to the person and society as those with depression have higher health care costs overall [2]. Rates of depression are higher among those who require health-related assistance, including 38% of home health care recipients, 23% of older adults living in residential settings and 49% of nursing home care recipients [3]. Despite its frequency, depression symptoms are often not recognized as a clinical syndrome, but are accepted as a normal part of the aging process [4].

Interrupting this often devastating sequence will be best achieved using an objective biological measure of depressive symptoms, one that doesn't rely on the subjective assessments of older people, families, or their providers. New evidence indicates that vocal patterns can be used to identify and detect changes in clinical depression. Acoustic properties of speech have successfully detected depression in adults 18-65 years old [5–8]. However, no studies to date have included older adults, let alone focused on the vocal characteristics of depressed older people. This presents a critical gap in knowledge given that vocal cord function is affected by advancing age, sex, and medical problems that cluster in late life [9–12]. These aging vocal pattern changes, which range from changing speaking rates to pitch of speech, mean that previous approaches developed with younger depressed adults cannot be consistently applied to older people. Using vocal pattern data is a promising method of identifying late-life depression: it's objective, unobtrusive, and easy to capture. Novel in-home monitoring devices may eventually use vocal data to detect individual changes that signal worsening depression and risk of worsening health, which, in turn, may lead to earlier treatment that reduces healthcare costs and disease burden.

Detecting changes at the time of their occurrence using home monitoring that "alerts" the person and health provider that evaluation is needed will vastly shorten the time to treatment. The current approach of waiting until symptoms become severe enough that the individual seeks help, or the problem is identified at the next clinical visit, which may be months away, too often results in more severe depression that is harder to treat [13]. Identifying the onset of increasing depressive symptoms using passive monitoring offers exceptional opportunities to intervene early when depression is milder and responsive to non-drug interventions like enhanced physical, social, and meaningful activity engagement. Early intervention also reduces the risk of spiraling depression-related disability that is associated with a host of related problems: reduced quality of life and function, and increased healthcare costs and risk of needing more intensive care and support among many others. Passive monitoring of vocal patterns may also contribute to the treatment of late life depression by identifying changes that signal non-response to therapeutic agents, which, in turn, signal need for a quicker return to the treating health provider.

In short, there are many advantages to using vocal patterns to detect clinically significant depression among older adults. However, the essential first step is to establish whether vocal biomarkers that have predicted depression in younger adults are also relevant in older people. The overall goal of the pilot study was to establish that vocal patterns in older adults can predict depression, and thus lay a foundation for a larger research to distinguish vocal biomarkers of depression that may be used in personalized monitoring devices.

Study Purpose and Research Questions

The purpose of this paper is to describe outcomes from a pilot study that was designed to evaluate vocal patterns among older adults related to their depressive symptoms using machine learning analytics. Two main research questions guide this pilot study:

1. Are vocal patterns that have been identified as salient for depression detection among younger adults also indicative of depression among older adults?

2. Are there additional vocal patterns that signify depression in older adults that are different from younger age groups?

In this paper, we describe the first set of data collected from older adults who reside in senior living settings. Recruitment from this setting was based on reports that 23-34% of residents experience depression [3, 14, 15].

Materials and methods

Our study used an observational repeated measures design in which volunteer older adults who reside in senior living settings completed semi-structured interviews conducted by trained research assistants. The study was approved by the University of Iowa Institutional Review Board in July 2017 and data described here was collected in the following 6 months. Segmenting and processing of the resulting audio recordings occurred between January and May 2018, and analyses reported below were completed June to August, 2018. The study was supported with funding from the University of Iowa Office of the Vice Presidents for Research and Economic Developments' Strategic Research Leadership program.

Setting and Sample

Three senior living communities that serve from 150 to 500 older adults agreed to assist with the study. Each setting had participated in earlier depression-focused research which contributed to ease of entrée and understanding of research processes. Senior living staff helped advertise the study and assisted the research team to identify and use quiet, private locations to conduct research-related activities and interviews.

Older adults were recruited using a combination of public advertising materials and personal invitations by senior living staff members. The study employed an "all comers" approach to promote a broad view of older people's vocal patterns/acoustic features, and how those related to levels of depression. To that end, we did not seek individuals who were depressed, but rather invited all willing older adults to participate. Inclusion/exclusion criteria included (a) being at least 65 years of age; (b) speaking English (the format of the interview); (c) ability to hear the interviewer's questions; (d) ability to read study-related information; and (e) ability to sign informed consent documents. The last aspect, signing informed consent, was a proxy for cognitive impairment that would otherwise be considered an exclusion criteria.

Measures and Scales

The semi-structured interview was composed of basic demographic information that included age, sex, race, Hispanic origin, and residence name; a widely used depression rating scale to quantify depression severity; and selected vocal measures that had been used successfully in earlier research with younger adults.

Depression Severity

The 9-item Patient Health Questionnaire, or PHQ-9 scale rates the nine criteria for clinical depression (Major Depressive Disorder) on a four-point scale where 0=not at all, 1=several days, 2=more than half the days, and 4=nearly every day. Total scores range from 0 to 27 and have established cut-points indicating depression severity: 0-4=None; 5-9=Mild; 10-14=Moderate; 15-19=Moderately severe; >20=Severe. Scores of 10 or greater indicate clinically significant depression; scores of >15 are associated with clinical depression (aka major depressive disorder) [16]. The PHQ-9 was selected based on its ease of administration by non-clinician personnel and its track record of successful use in clinical practice settings and research.

Reading Aloud

Two phonetically balanced passages that elicit a range of vocal characteristics were included. Both were used successfully in earlier research with younger adults. The Rainbow Passage [17] was used by Hashim and colleagues [18]. The somewhat shorter Grandfather Passage [19] was used by Mundt and colleagues [6, 20]. In the pilot study, participants were asked to read one of the two passages to allow comparisons of outcomes without increasing the length of the interview.

Free Speech

The second task was to collect speech that is spoken in manner that is typical (normal) for the individual [6, 20]. Previous research used depression-related questions to assess free speech. In contrast, our questions were devised to be "emotion neutral," meaning that we focus on topics that do not elicit either positive or negative thoughts or feelings. These questions include: (a) Tell me about your typical day; (b) What activities do you have planned for the rest of the day? (c) What are your favorite activities? (d) What type of music do you like? and (e) What type of food do you like? If needed, extemporary questions were added by interviewers to elicit a total of five minutes of free speech.

Interview Approach

All participants were assigned a unique study identification number. Assessments were conducted in person and recorded using Sony digital recording devices. Time stamps to facilitate data slicing were recorded using a tablet. Interviews were conducted in locations preferred by individual participants, such as their own room or apartment or other private location. Interviewers were trained to minimize external noise once the recording starts, and not deviate from the interview script to reduce risks of influencing participant's speech qualities.

The number of interviews conducted with individual participants was guided by their baseline PHQ-9 score. All participants were asked to complete two audio recorded interviews. Participants with PHQ-9 scores of less than 10 will be asked to repeat the interview in 2 weeks. Those with PHQ-9 scores of 10 or greater were asked to repeat the interview at 2, 4, and 6 weeks post baseline. This approach allowed all participants to serve as their own control in analyses. The longer interval (4 and 6 weeks) for those whose score suggest clinically significant depression was intended to capture change in vocal characteristics associated with different levels of depression.

Hardware and Software

The PHQ-9 items and total score and background information were entered using REDCap (Research Electric Data Capture), a data management tool that provides a secure multi-user web-based interface for storing study information (https://its.uiowa.edu/redcap). Data were exported to a comma-separated values (CSV) spreadsheet. Descriptive summaries and all other analyses were conducted using the R statistical language, Version 3.4.3. All machine learning algorithms were estimated using the *caret* library of the R statistical language.

Audio recordings were transferred from the mobile recording devices used in the field to a secure research drive. Interviews were segmented into single audio files for each response using Audacity (`http://audacityteam.org/`). Features were then extracted using an open-source audio analysis and pattern recognition tool called openSMILE (`https://audeering.com/technology/opensmile/`). For this study we used the feature set from the 2010 InterSpeech challenge (`https://www.isca-speech.org/archive/interspeech_2010/index.html`) which can be found in the IS10_paraling openSMILE configuration file. The feature set includes 1,582 different permutations of 38 base-level features. These base-level features include: PCM loudness, MFCC, LPCC, fundamental frequency (F0), Jitter, and Shimmer. Each of these variables is explained in Table 1. All 38 measures were used for the purpose of this analysis in order to make few assumptions about which features best capture different states of depression.

Machine Learning

A Support Vector Machine (SVM) with a radial kernel was used to predict whether the participant was either moderately/severely depressed (1) or mildly/not depressed (0) using PHQ-9 categorical levels of depression [16]. This algorithm fits a multidimensional hyper-plane that optimally discriminates between high and low levels of depression. To fit the model, we used repeated 10-fold cross-validation. More specifically, we randomly assigned cases to 10 equally size folds and used 9 of those folds to train the SVM and the left-out fold for testing. This process was repeated for all 10 folds, meaning each fold was used exactly once in the testing data. We repeated this process using 3 different data partitions. To test the fitted model, we set aside 20 percent of our data. The other 80 percent of the data was used to fit the cross-validated SVM. For the raw PHQ-9 scores we used the same approach, except our dependent variable was continuous and ranged from 0-27. The unit of analysis is a given audio file clustered within each respondent meaning the same respondent can appear more than once.

To assess model performance, we used the measures suggested by Luo and colleagues [21]. More specifically, we report accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the Receiver Operating Characteristic curve (AUC) for the models predicting moderately/severely depressed (1) or mildly/not at all depressed (0). For the models predicting the raw PHQ-9 scores, we report the Mean Absolute Error (MAE) and the normalized Root Mean Squared Error (RMSE). All measures are described in Table 2. We focus our discussion of results on accuracy and the Mean Absolute Error based on their relative ease of interpretation even though all measures are reported.

Results

Baseline Characteristics

A total of 46 older adults were recruited over a six-month period, including 10 men and 36 women. All (100%) were white and non-Hispanic. Current age ranged from 66 to 93 years old, with a mean of 81 years. At baseline and two-week assessments, PHQ-9 scores ranged from 0 to 26, with a mean of 4.57 (SD= 5.04) and 4.56 (SD= 4.33) respectively. Nine participants had scores of 10 or greater. Scores at 4 weeks resulted in a range of 0 to 7 and mean of 3.50 (SD= 3.11), and at 6 weeks scores ranged from 1 to 5 with a mean of 2.0 (SD= 2.65).

Table 1. Description of audio variables.

Variable	Description
PCM Loudness	Given that a particular change in amplitude is not perceived as a proportional change in loudness, the amplitude of a signal must be standardized. In Pulse Code Modulation (PCM) systems, perceived loudness is calculated using the ratio of the maximum amplitude and the inherent noise in the system.
MFCC	Speech is analyzed at the frame- and trend-level. The Mel Frequency Cepstral Coefficients (MFCC) is an example of the latter since they provide a summary of the energy distribution at specific frequencies for the entire speech signal. Ultimately, they are returned in the Mel scale which relates the perceived frequency (or pitch) to the actual measured frequency. Since the vocal tract is manipulated to change the perceived frequency (or pitch), the MFCC essentially captures the shape of the vocal tract.
LPCC	The Linear Predictive Coding Coefficients (LPCC) is another type of frame-level measure which provides a summary of the entire speech signal. However, instead of using a quasi-logarithmic scale similar to the MFCC, the LPCC uses the past values in a speech signal to predict the current values using a linear function. Unlike the MFCC, the LPCC is mostly used to model how a speech signal is produced rather than how it is perceived. With that said, both are trend-level measures and capture the energy distribution at specific frequencies for the entire speech signal.
Fundamental Frequency (F0)	The Fundamental Frequency (F0) is perceived by the human ear as pitch. It represents the frequency at which the vocal folds are opening and closing. This serves as the basis for all human speech since this quasi-periodic function resonates throughout the vocal tract ultimately producing the sound we hear. It is "fundamental" since it is often associated with the source of a speech signal (i.e., emotional activation) as compared to the vocal tract which filters the source to create specific sounds (i.e., words and phrases).
Jitter	The number of cycles the vocal folds make in a second is the fundamental frequency. These cycles are primarily determined by the degree of longitudinal stress placed on the vocal folds and the dimensions of the vocal folds themselves. Jitter is the variability of the fundamental frequency which ultimately captures the degree to which an individual has control over the vocal fold vibration. Rough (or hoarse) voices tend to have high jitter.
Shimmer	Shimmer is very similar to jitter, but instead of capturing the variability of the fundamental frequency shimmer captures the variability of amplitude. Unlike the fundamental frequency, the amplitude of a speech signal does not measure the rate of vocal fold vibration. Instead, it measures the size of the oscillations with greater amplitude implying the speech signal has more energy and will ultimately be perceived as being louder. Rough (or hoarse) voices also tend to have high shimmer which is why jitter and shimmer are often used together in the same model.

Note: This table provides a description of the variables we used in the Support Vector Machine (SVM) described on page 5.

Binary Outcomes

Results of analyses using binary outcomes of moderately/severely depressed (1) and mildly/not depressed (0) are presented in Fig 1. Here, all the assessment measures are

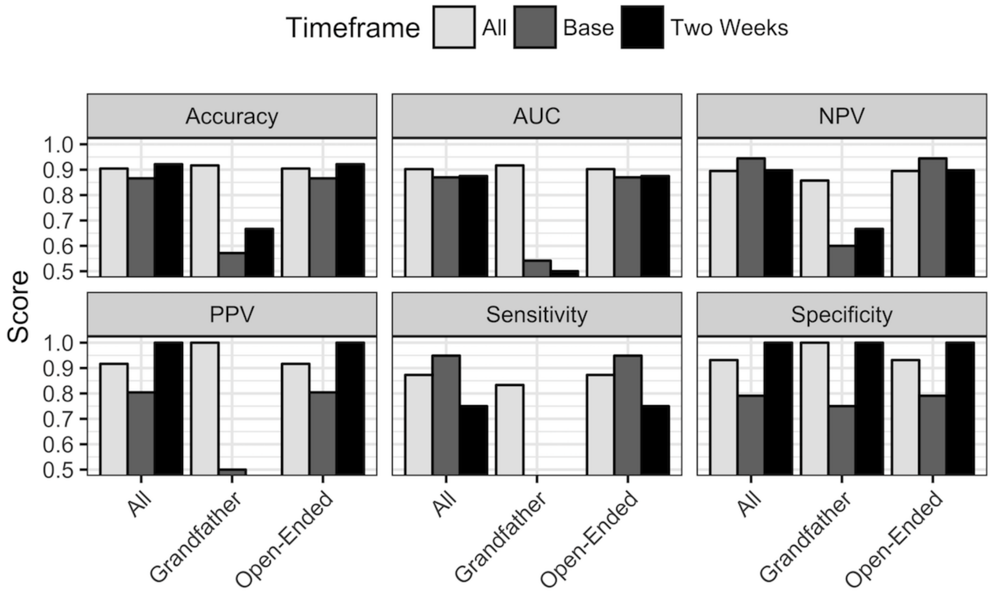
Table 2. Description of assessment measures.

Measure	Description
Accuracy	Accuracy is simply the number of cases correctly predicted divided by the total number of cases.
True Positive Rate	See Sensitivity
False Positive Rate	The false positive rate is the probability that a respondent who is mildly/not depressed will be said to be moderately/severely depressed by our model.
AUC	The Receiver Operating Characteristic (ROC) curve plots the False Positive Rate by the True Positive Rate. AUC (or Area Under the ROC Curve) is the total area under the ROC Curve. This essentially equates to the probability our model returns a 1 for a random person with moderate/severe depression as compared to a random person with mild/no depression. A model who does not classify any of the respondents correctly has an AUC of 0, whereas a model who classifies all respondents correctly would have an AUC of 1.
Negative Predictive Value (NPV)	Negative predictive value is the probability a respondent who is said to be mildly/not depressed by our model is actually mildly/not depressed.
Positive Predictive Value (PPV)	Positive predictive value is the probability a respondent who is said to be moderately/severely depressed by our model is actually moderately/severely depressed.
Sensitivity	A test with 100% sensitivity correctly identifies all respondents with the condition. A test with 75% sensitivity identifies 75% of respondents with the condition (true positives) but 25% with the condition go undetected (false negatives). Thus, sensitivity is simply the probability a respondent with moderate/severe depression will be identified as moderately/severely depressed by our model. Sensitivity is also equal to the true positive rate.
Specificity	A test with 100% specificity correctly identifies all respondents without the condition. A test with 75% specificity correctly identifies 75% of respondents without the condition as negative (true negatives) but 25% patients without the condition are incorrectly identified as positive (false positives). Thus, specificity is simply the probability a respondent with mildly/no depression will be identified as moderately/severely depressed by our model. Specificity is also 1 minus the false positive rate.
Mean Absolute Error (MAE)	The Mean Absolute Error (MAE) is the average absolute difference between the predicted and actual score.
Root Mean Squared Error (RMSE)	The Root Mean Squared Error (RMSE) is the average square root of the squared difference between the actual and predicted score.

Note: This table provides a description of the measures we used to assess the performance of the Support Vector Machine (SVM) described on page 5.

reported in each of the six panels. On the x-axis we report whether we used data from all the audio files (see “All”), only the reading of the Grandfather or Rainbow passage (see "Grandfather"), or only the open-ended questions (see "Open-Ended"). The light grey, dark grey, and black bars indicate whether we used data from all interviews (see light grey), the baseline interview (see dark grey), or the two-week interview (see black bar). All of the assessment measures range from 0 to 1 and are shown on the y-axis.

Fig 1. Dependent variable is moderately/severely depressed (1) vs. mildly/not depressed (0).



Please refer to Table 2 for descriptions of each measure. The range of the y-axis is restricted to .50 to 1 which affects the reporting of the sensitivity scores for the Grandfather Passage during the baseline and two-week interviews. These scores were 0.33 and 0.00 and were not captured by the truncated scale. Similarly, the positive predictive value (PPV) for the Grandfather Passage during the two-week interview was 0.00 so was omitted from Fig 1. In the light gray bars, all available interview data was used. In the dark gray and black bars, we only used data from the baseline and two-week interviews, respectively.

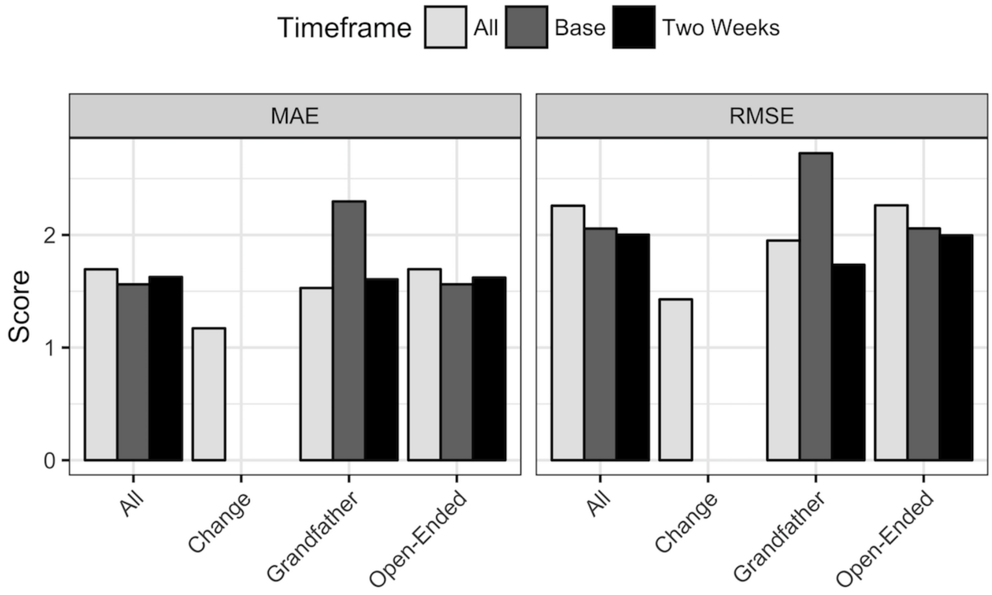
When the model was fit using all available audio files our overall accuracy ranged from 86.59 to 92.16 percent. Prediction of respondents who are moderately/severely depressed, positive predictive value (PPV), ranged between 80.43 and 100 percent. Similar results are found for the negative predictive value (NPV), as the model accurately predicted respondents who were mildly/not depressed from 89.47 to 94.44 percent. Thus, the model effectively discriminated between older adults who have moderate/severe levels of depression versus those who do not.

Some of the models which only used the Grandfather and Rainbow Passages noticeably underperformed compared to using all the audio data. Beginning with accuracy, when only the audio from the baseline interview (see dark gray bar in Fig 1) is utilized we predict whether the individual is moderately/severely depressed 57.14 percent of the time. When we use audio variables from the two-week interview (see black bar in Fig 1), the model accuracy slightly improves to 66.67 percent. Finally, using all available interview data (see light gray bar in Fig 1) we predict 91.67 percent of cases correctly. This improvement is partially attributed to respondents reading only one of the passages in the baseline interview and the other passage at the two-week interview, meaning the “All” column is the only column which uses audio from both the Grandfather and Rainbow Passage.

Raw PHQ-6 Score Outcomes

In Figure 2 we report outcomes for raw PHQ-9 scores. Again, we report both assessment measures in each panel. On the x-axis we report whether we used data from all the audio files (see “All”), only the reading of the Grandfather or Rainbow passage (see “Grandfather”), or only the open-ended questions (see “Open-Ended”). Similarly, the light grey, dark grey, and black bars indicate whether we used data from all interviews (see light grey), the baseline interview (see dark grey), or the two-week interview (see black bar). All of the assessment measures can range from zero to infinity and are shown on the y-axis.

Fig 2. Dependent variable is the raw PHQ-9 scores which can range from 0 to 27.



Please refer to Table 2 for descriptions of each measure. For all models we used the raw PHQ-9 scores, but the “Change” models are difference-in-differences. All available interview data was used in the light gray bars. In the dark gray and black bars, we only used data from the baseline and two-week interviews, respectively.

The main difference between Figures 1 and 2 can be found in the column labeled “Change” which reports the results from a series of difference-in-differences models. Here, the dependent variable is the PHQ-9 score from the two-week interview minus the PHQ-9 score from the baseline interview positive values indicating the respondent became more depressed between the two interviews. The same differencing was used for each audio feature with positive values implying the audio feature increased from the baseline to two-week interview. We then used the same repeated cross-validated SVM to determine whether changes in the audio variables from the baseline to two-week interviews were predictive of the corresponding changes in the PHQ-9 scores.

We begin with the Mean Absolute Error (MAE) which is the average absolute difference between the actual and predicted PHQ-9 score. Here, when all the audio files are used we are able to predict the raw PHQ-9 score for each respondent in our training set within 1.69 units (scale points) in either direction. We found the audio variables extracted from the reading of either the Grandfather or Rainbow Passages performed slightly better than the audio extracted from the responses to the open-ended questions. When all the interviews were used our MAE was 1.70 for the open-ended questions, whereas the MAE decreased to 1.53 when only the audio from the reading of the Grandfather/Rainbow Passages was used. Given that we found the opposite result for the models predicting the binary outcome variable, these results suggest audio from structured and unstructured responses can be used to classify depression in older adults.

Using the Root Means Squared Error (RMSE), which is the average square root of the squared difference between the actual and predicted PHQ-9 score, we find very similar results. Not only is the RMSE slightly smaller for the models which only use audio from the reading of the Grandfather and Rainbow passages (1.95) as compared to the models which only use the audio from the responses to the open-ended questions (2.26), but we find the models that perform the best are the difference-in-differences models predicting the change in the PHQ-9 scores from the baseline to two-week interviews.

More specifically, when we used all the available audio files to predict the change in each respondent's PHQ-9 score from the baseline to two-week interview we were able to predict the change within 1.17 units. We find the same results for the RMSE. For example, when only the audio from the open-ended questions reading in the first two interviews (baseline and two-week) is used, the RMSE is 2.00 as compared to the change in the PHQ-9 scores which has an RMSE of 1.43 for the same time period. Again, the models reported in the "Change" column are difference-in-differences, so they cannot be directly compared to the models which predict the raw PHQ-9 scores. With that said, our results provide preliminary evidence audio data can be used to not only predict the current level of depression in older adults, but also changes in depression from one assessment period to the next.

Discussion

The vast majority of our models performed extremely well given that our data set provided some inherent limitations related to the level of depression among our participants. We used senior living settings for recruitment based on reports that a quarter or more of older people in this setting have depression [3]. Our sample had somewhat lower levels of depression as mean scores (4.6) hovered in the range of no to mild depression, and 20% (9 of 46) had scores of 10 or greater that indicate clinically meaningful depression. In turn, models to predict depression relied on a limited range of depression scores that were skewed toward having no depression. Even so, the machine learning methods consistently supported that vocal data from older adults can be successfully used to predict both binary outcomes and change in change in raw PHQ-9 scale scores.

We were also able to accurately predict the binary outcome of moderately/severely depressed vs. mildly/not depressed. Not only were outcomes well above chance, but they equaled or exceeded the performance of other machine learning models used to predict levels of depression in younger adults. We also learned that performance of models using the structured Grandfather and Rainbow Passages is largely a function of the amount of audio data used to train the models. Unlike the models which used the audio from both the open-ended questions and the reading of the Grandfather/Rainbow Passage, the models which only used the latter have on average two audio files per respondent whereas the full models have on average 22 audio files. Consequently, the models which use all the data perform better because they have over ten times more audio information. When both passages and both baseline and two-week data were included, the model predicted 91.67% of the cases. In short, when enough audio is available from reading the structured Grandfather and Rainbow passages then the model performs equally well as using all data.

Conclusion

Findings from our pilot project provide substantial support that vocal pattern characteristics can predict depression among older adults. Even though only 20% of the sample experienced clinically meaningful levels of depression, we demonstrate machine learning can be used to effectively discriminate between those with and without depression. Our findings support the need for further research that will establish a large data base on which late life depression algorithms may be refined, and then later deployed in devices that will detect early changes in mood so that earlier referral and treatments may be provided.

Acknowledgments

Conflicts. Drs. Smith, Dietrich, Bai and Mr. Bockholt have no financial conflicts, no personal conflicts, and no conflicts of interest to declare.

Author Contributions. Drs. Smith, Dietrich Bai, and Mr. Bockholt have each made substantial contributions to: the conception and design of the study, or acquisition of the data, or analysis and interpretation of the data; and, drafting the article, or revising it critically for important intellectual consent. Moreover, they have each granted final approval of the content version for submission to be published.

Funding. This pilot study was funded by the University of Iowa Office of the Vice President for Research and Economic Development (OVPRED) under the Strategic Research Leadership Program that supported work of the Iowa Collaboratory for Healthy and Successful Aging through Innovative Technologies that led to the study. The authors wish to acknowledge and thank: Dr. Richard Hichwa, Senior Associate Vice President for Research and Economic Development, Deepak Gaddipatti, VirtuSense Technologies, and Collaboratory members James Schappet, and Drs. Fredrik Wolinsky, Gary Christensen, and Jean K. Gordon for assistance in the conceptualization of vocal pattern detection in older adults; research assistants Jill Colbert, Megan Dotson and Christine Bricker for help with data collection and management; and leaders and staff at Western Home Communities, Cedar Falls, Meth-Wick Community, Cedar Rapids, and Melrose Meadows Retirement Community, Iowa City, for entrée to their senior living communities and assistance with identifying participants.

References

1. Depression: Fact Sheet. 2018 March 22 [cited 28 November 2018]. In: World Health Organization [Internet]. Geneva, SUI: 2018. [about 5 screens]. Available from: http://www.who.int/news-room/fact-sheets/detail/depression.

2. Katon WJ, Lin E, Russo J, Unutzer J. Increased medical costs of a population-based sample of depressed elderly patients. Arch Gen Psychiatry. 2003; 60(9):897-903.

3. Centers for Disease Control and Prevention. Long-Term Care Providers and Services Users in the United States: Data from the National Study of Long-Term Care Providers, 2013-14. Hyattsville, MD: DHHS publication No. 2016-1422; 2016.

4. Older Adults: Depression and Suicide Facts. 2009 May 18 [cited 1 May 2013]. In: National Institute of Mental Health [Internet]. Bethesda, MD: 2018. [about 2 screens]. Available from: http://www.nimh.nih.gov/health/publications/older-adults-depression-and-suicide-facts-fact-sheet/index.shtml.

5. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng. 2000; 47(7):829-837.

6. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry. 2012; 72(7):580-587.

7. Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ. Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. J Voice. 2017; 31(2):256 e251-256 e256.

8. Yang Y, Fairbairn C, Cohn JF. Detecting Depression Severity from Vocal Prosody. IEEE transactions on affective computing. 2013; 4(2):142-150

9. Gorham-Rowan MM, Laures-Gore J. Acoustic-perceptual correlates of voice quality in elderly men and women. J Commun Disord. 2006; 39(3):171-184.

10. Mezzedimi C, Di Francesco M, Livi W, Spinosi MC, De Felice C. Objective Evaluation of Presbyphonia: Spectroacoustic Study on 142 Patients with Praat. J Voice. 2017; 31(2):257.e225-257.e232.

11. Torre P, 3rd, Barlow JA. Age-related changes in acoustic characteristics of adult speech. J Commun Disord. 2009; 42(5):324-333.

12. Winkler R, Sendlmeier W. EGG open quotient in aging voices–changes with increasing chronological age and its perception. Logopedics, phoniatrics, vocology. 2006; 31(2):51-56.

13. Grabovich A, Lu N, Tang W, Tu X, Lyness JM. Outcomes of subsyndromal depression in older primary care patients. Am J Geriatr Psychiatry. 2010; 18(3):227-235.

14. Gimm GW, Kitsantas P. Falls, Depression, and Other Hospitalization Risk Factors for Adults in Residential Care Facilities. Int J Aging Hum Dev. 2016; 83(1):44-62.

15. Watson LC, Lehmann S, Mayer L, et al. Depression in assisted living is common and related to physical burden. Am J Geriatr Psychiatry. 2006; 14(10):876-883.

16. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16(9):606-613.

17. Fairbanks G. Voice and Articulation Drillbook, 2nd Edition. New York: Harper & Row; 1960.

18. Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ. Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. J Voice. 2017; 31(2):256.e251-256.e256.

19. Reilly J, Fisher J. Sherlock Homes and the strange case of the missing attribution: A historical note on "The Grandfather Passage". J Speech Lang Hear Res. 2012; 55(February):84-88.

20. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics. 2007; 20(1):50-64.

21. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016; 18(12):e323.