Published in final edited form as:

Biol Psychiatry. 2012 October 1; 72(7): 580–587. doi:10.1016/j.biopsych.2012.03.015.

Vocal Acoustic Biomarkers of Depression Severity and Treatment Response

James C. Mundt^{1,2}, Adam P. Vogel³, Douglas E. Feltner^{4,5}, and William R. Lenderking^{4,6}

¹Center for Psychological Consultation (Madison, WI)

²University of Arkansas for Medical Sciences (Little Rock, AR)

³Speech Neuroscience Unit, Melbourne Medical School, The University of Melbourne (Melbourne, AU)

⁴formerly Pfizer Global Research & Development – Groton Laboratories (Groton, CT)

⁵Douglas E. Feltner, LLC (Novi, MI)

⁶United BioSource Corporation (Bethesda, MD)

Abstract

Background—Valid, reliable biomarkers of depression severity and treatment response would provide new targets for clinical research. Noticeable differences in speech production between depressed and nondepressed patients have been suggested as a potential biomarker.

Methods—One hundred and five adults with Major Depression were recruited into a four-week, randomized, double-blind, placebo-controlled research methodology study. An exploratory objective of the study was to evaluate the generalizability and repeatability of prior study results indicating vocal acoustic properties in speech may serve as biomarkers for depression severity and response to treatment. Speech samples, collected at baseline and study end-point using an automated telephone system, were analyzed as a function of clinician-rated and patient-reported measures of depression severity and treatment response.

Results—Regression models of speech pattern changes associated with clinical outcomes in the prior study were found to be reliable and significant predictors of outcome in the current study despite differences in the methodological design and implementation of the two studies. Results of

© 2012 Society of Biological Psychiatry. Published by Elsevier Inc. All rights reserved.

Corresponding author: James C. Mundt, Center for Psychological Consultation, 7601 Ganser Way, 2nd Floor, Madison, WI 53719, JMundt@centerforpsychconsulting.com, (608) 406-2625 (tel) (888) 705 3447 (fax).

Financial Disclosures

Dr. Mundt has received research support grants from GlaxoSmithKline, Pfizer, Eli Lilly, Eisai, the National Institutes of Health, and the US Department of Agriculture. Dr. Mundt provides consulting services to ERT, a company that provides services to the pharmaceutical industry. Dr. Mundt is a former employee of, and owns stock in Healthcare Technology Systems, a research and development company that develops computer-automated patient assessment programs. Dr. Vogel has received research support from the National Health and Medical Research Council (Australia, #10012302) and is a former employee of CogState Ltd, a company that provides services to the pharmaceutical industry, and owns CogState stock. Dr. Feltner is Chief Medical Officer for and has stock options in Embera Neurotherapeutics, a company focused on smoking cessation and other drug abuse treatments. Dr. Feltner is a former employee of Pfizer Inc, and owns Pfizer stock. Dr. Lenderking was the Principal Investigator of this study and was a full-time employee of Pfizer, the manufacturer of sertraline, during the time this study was conceived and implemented. He continues to be a Pfizer stockholder. He is currently an employee of United BioSource Corporation (UBC), a wholly owned subsidiary of Medco. UBC is a company that provides consulting services for hire to the pharmaceutical, biotech, and device industries.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the current study replicate and support findings from the prior study. Clinical changes in depressive symptoms among patients responding to the treatments provided also reflected significant differences in speech production patterns. Depressed patients who did not improve clinically showed smaller vocal acoustic changes and/or changes that were directionally opposite to treatment responders.

Conclusions—This study supports the feasibility and validity of obtaining clinically important, biologically-based vocal acoustic measures of depression severity and treatment response using an automated telephone system. National Institutes of Health Clinical Trials Registry: <http://clinicaltrials.gov> Identifier: NCT00406952.

Keywords

Depression assessment; methodology; speech; voice acoustics; telephone; interactive voice response (IVR)

Introduction

Depression and anxiety disorders are critical public health and safety concerns throughout the developed and developing worlds (1). Depression interferes with the lives of roughly 1 of every 6 adults in the United States and is associated with more than half of all suicides (2, 3). Development of evidenced-based treatments requires reliable and valid measures of clinical severity and patient response. Objective measures of illness severity in many medical fields reflect direct measurements of physical characteristics, such as tumor volume or blood chemistry tests. Treatment outcome measures of mental disorders typically rely upon subjective judgments provided by the patient and/or a trained clinician, such as the Hamilton Depression Rating Scale (HAM-D) (4, 5), Montgomery-Asberg Depression Rating Scale (6), Inventory of Depressive Symptomatology (7), or Quick Inventory of Depressive Symptomatology (QIDS) (8). Results of mood and anxiety clinical trials have been inconsistent in recent years (9). Methodological and study design features, such as patient enrollment pressures, high severity criteria for study inclusion, rater expectancies, baseline score inflation, and unblinding of raters may contribute to inaccurate clinical assessments and unreliable study findings (10-15).

Development of an objective, non-invasive, physiologically-based biomarker of depression severity that is sensitive to clinical change associated with treatment could provide new avenues for clinical research and development of treatments. Biomarkers for diagnosis and assessment of depression have been sought for years (16), including investigations of neuroendocrine response to dexamethasone and corticotrophin releasing hormone (17), electroencephalographic (EEG) measures during sleep (18), cytokine plasma concentrations (19, 20), monoamine levels in cerebrospinal fluids (21), and brain-derived neurotrophic factor (22). Validation and use of a practical, easy-to-use biomarker for diagnostic testing and treatment response for depression remains elusive however (23).

Qualitative differences in speech produced by depressed and nondepressed patients have been recognized for many years (24, 25), but quantitative validation of speech parameters as indicators of depression severity and treatment response were not demonstrated until the 1970's and '80's (26, 27). Change in speech pause times and other acoustic measures were identified and offered as non-invasive biomarkers for "reducing clinical subjectivity in monitoring treatment progress" (28, 29). Replication and generalizability of these findings to non-English speaking depressed patients has been demonstrated (30), and the effects of clinical depression on measures of speech pitch have also been established (31).

Clinical improvement associated with antidepressant treatments was modeled using multivariate equations of voice acoustic parameters in the 1990's (32-34), and efforts to further refine and validate vocal acoustic measures as biomarkers of central nervous system functioning continue, particularly with respect to depression (35-39). Use of vocal acoustic measures of depression severity and treatment response in clinical trials has been limited, due to perceived needs for specialized recording equipment and environments, software, and the technical skills needed to analyze the speech samples (40). Automated collection and analysis of speech samples using inexpensive, widely available devices, public-domain software, and high capacity data servers now addresses these perceived barriers. Speech samples obtained over telephones are adequate for extracting useful clinical information (41), and analysis of these speech samples can be streamlined using automated batch processing (42, 43).

In 2007, an interactive voice response (IVR) system was used to collect speech samples and computer-automated HAM-D measures of depression severity (44-47) in a naturalistic observational study (48). Significant correlations were found between depression severity and several vocal acoustic measures; patients who responded clinically to their treatments showed greater change in several speech measures than the patients who did not respond to treatment. Additionally, patients responding to their treatments showed significant within-subject changes of vocal acoustic measures between the beginning and end of treatment, while treatment nonresponders did not. This suggests that the changes in vocal acoustic measures during speech production reflect underlying neurophysiologic changes associated with clinical response evident by changes in symptoms of depression over time. Speech data from the prior study were subsequently analyzed using different, independently developed methods; results were consistent with previous findings (49).

The present study was a Phase 4 randomized, double-blind, placebo-controlled methodology study. The main objective was to evaluate new methodology for detecting antidepressant treatment response. Another study objective was to replicate findings from the 2007 vocal acoustic study (48) and assess the generalizability of vocal acoustic measures as potential biomarkers of depression severity and treatment response in a randomized clinical trial. Important methodological similarities and differences between the previous and current study are identified in Table 1

Methods and Materials

Eleven investigational sites across the United States screened 183 participants and randomized 165 to study treatments between November 2006 and August 2007. The randomized sample included 61 males and 104 females. Mean participant age was 37.8 years (SD = 12.5); 125 were White, 26 were Black, 4 were Asian, 10 reported their race as 'Other.' Inclusion criteria were: 18 to 65 years old; not currently taking psychotropic medications; a 17-item clinician-administered HAM-D score of 22 or greater at baseline; primary diagnosis of major depressive disorder based on Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria, with symptoms present for at least one month and willingness to provide informed consent as required by federal and institutional guidelines. Exclusion criteria were: Prior failure to respond to sertraline treatment or two or more antidepressant treatments of adequate dose and duration in the past 5 years; current diagnosis of generalized anxiety, social anxiety, obsessive compulsive, panic, post-traumatic stress, anorexia, bulimia, or substance abuse disorder within the past 6 months; current or past diagnosis of delirium, schizophrenia, bipolar, psychotic, dementia, or amnesic cognitive disorder.

Participants attended four office visits during the 4-week study period. Screening procedures confirmed study eligibility and baseline data were collected at the same visit (Week 0). Clinical assessments of depression severity and treatment response were obtained at the end of Weeks 1, 2, and 4. Double-blind, randomized assignment to a forced-titration dosing regimen started all participants on 50 mg/day of sertraline or placebo at baseline. At the end of Weeks 1 and 2 dosages were either (a) maintained if treatment efficacy was evident with minimal side effects, (b) increased by 50 mg/day to improve inadequate therapeutic response, or (c) decreased to reduce adverse events.

Clinical assessments of depression severity and treatment response at weeks 1, 2, and 4 included clinician-rated QIDS (QIDS-C) and HAM-D, and a paper self-report form of the QIDS (QIDS-SR). The QIDS is a 16-item scale; total scores ranging from 0 – 27, based on severity scores of nine symptom domains used as DSM-IV (50) diagnosis criteria. 17-item HAM-D scores range from 0 – 52. Baseline HAM-D scores were used to determine study eligibility, but were not used as the primary clinical outcome measure. Use of different clinical measures for study eligibility and treatment response reduces confounding of results by inflated baseline scores (14). Treatment response, for evaluation of the vocal acoustic measures, was defined as a 50% or greater reduction in the total symptom scores between baseline and study end-point. The *a priori* statistical analysis plan for evaluating the vocal acoustic data identified the QIDS-C as the primary clinical outcome measure of depression severity and treatment response.

At baseline, and after weeks 1 and 4, speech samples were collected using a standardized, automated telephone interface developed previously (48, 51). Speech samples consisted of (a) “Free Speech” descriptions of recent physical and mental experiences and their effects on functioning over the prior week; (b) “Automated Speech” produced by reciting the alphabet and counting from 1 to 20; (c) “Reading” the phonetically-balanced, 169-syllable, 129-word *Grandfather Passage* commonly to assess communication disorders; and (d) “Sustained Vowels” of /a/, /i/, /u/, and /ae/ held for approximately 5 seconds. Participants were instructed to speak in a natural manner at their typical speaking rate. Their speech was recorded over standard telephone lines as single-channel, 8-bit .wav files sampled at 8 kHz.

Analysis of the speech samples focused on motor timing measures, such as the duration and proportion of silences and vocalizations within the samples, and frequency domain measures, such as the mean and variability of the fundamental frequency (F_0) and first and second formants (F_1 and F_2) produced by acoustic resonances within the human vocal tract. More complete description of the data processing and analytic methods have been published elsewhere (52) and are provided in Supplement 1. The acoustic measures extracted from each speech sample type are listed in Table 2.

Reported results were specified in a statistical analysis plan created prior to study completion and data lock. Statistical Package for the Social Sciences (SPSS Version 16) was used to conduct the analyses and an alpha value of $p < .05$ was specified for the planned comparisons. Between-group differences (e.g., responders versus nonresponders; sertraline versus placebo) were tested using two-group t-tests of means (53), and change over time was tested using paired mean t-tests (54). Tests of proportional equivalence in cross-tabulated frequencies used Fisher’s exact test (55), and classification agreement was tested using Cohen’s kappa (56), a measure of inter-rater agreement for categorical items that takes into account the degree of agreement that would occur due to chance alone.

Prior to the current study, logistic regression analyses of data obtained from the previous study (48) were conducted. Additional details regarding these analyses, and the calculation of response probabilities and likelihoods using data from the current study, are provided in

Supplement 1. In the first analysis of the data from the prior study, baseline to Week 6 *change* (Δ) in vocal acoustic measures were computed. These change measures were entered as independent variables to model the differences between patients who responded to treatment (Responders; N = 13) and those who did not (Nonresponders; N = 19). Equation 1 provides the multivariate equation that maximized classification accuracy of the patients (91%) in the prior study data, $\chi^2(7) = 25.4, p < .001$.

$$\begin{aligned} \text{Log Odds Response} = & 4.71 - 5.68 (\Delta \text{ Speaking rate}) - 214.1 (\Delta F_0 \text{ COV}) + 0.28 \\ & (\Delta \text{ Speech/pause ratio}) - 0.30 (\Delta \text{ Total recording time}) + 76.3 (\Delta \text{ Percent pause time}) \\ & + 0.42 (\Delta \text{ Total vocalization time}) - 18.2 (\Delta \text{ Pause variability}) \end{aligned} \quad \text{EQ. 1}$$

A second analysis entered the vocal acoustic measures observed at baseline (only) as potential *predictors* of response available *prior* to receiving treatment. Equation 2 shows the best logistic regression model that significantly ‘predicted’ treatment response status at Week 6 in the prior study, $\chi^2(2) = 6.9, p = .032$.

$$\text{Log Odds Response} = 2.45 + 0.06 (\text{Baseline Total recording length}) - 62.93 (\text{Baseline } F_0 \text{ COV}) \quad \text{EQ. 2}$$

Logistic regression equations can be used to estimate the *probability* that a particular case belongs to one binary class or another (e.g., treatment responder or nonresponder) by applying Equation 3 (see Supplement 1).

$$\text{Prob(R)} = (e^{\text{Log Odds Response}} / (1 + e^{\text{Log Odds Response}})). \quad \text{EQ. 3}$$

Generalizability of Equation 1 was evaluated by entering baseline to Week 4 vocal acoustic measure change – from the present study – and classifying participants with probability estimates of .5 or greater as ‘Vocal Acoustic Responders’. Equation 2 generalizability was evaluated by entering the baseline speech data – from the present study – and classifying participants with probability estimates of .5 or greater as ‘*Likely* Vocal Acoustic Responders’. These *a priori* classifications were compared with the observed clinical measures of treatment response or nonresponse from the present study (see Tables 5 and 6).

Results

Of the 165 participants randomized to treatment, 39 discontinued prior to study completion, 20 were missing speech samples at baseline or end-point prohibiting meaningful analysis, and 1 subject was missing baseline QIDS-C data. Data from 105 evaluable participants were analyzed. The QIDS-C total depression scores were significantly correlated with both the HAM-D and QIDS-SR scores at baseline ($r = .43$ and $.53$; $r^2 = .18$ and $.28$) and at Week 4 ($r = .87$ and $.81$; $r^2 = .76$ and $.66$, respectively).

Fifty-five participants were randomized to sertraline treatment and fifty received placebo treatment. Means and standard deviations of the clinical depression measures for each treatment group at baseline and Week 4 are shown in Table 3. Significant clinical improvement from baseline to end-point is evident in both treatment groups across all three clinical outcome measures. Neither the QIDS-C nor the QIDS-SR found statistically significant mean score differences between the treatment groups, $t(103) = 0.31$ and 0.14 (*ns*), respectively. The mean improvement of 13.4 HAM-D points in the sertraline group is significantly greater than the 10.4 point improvement seen in the placebo-treated participants, $t(103) = -2.21, p < .05$.

The proportion of treatment responders and nonresponders within each treatment group was not statistically different by either the QIDS-C or QIDS-SR measure, $\chi^2 = 0.12$ and 0.08 (*ns*), respectively. With the HAM-D measure, the proportion of treatment responders in the sertraline group (60%) is greater than the proportion of placebo responders (40%), but was not statistically significant, $\chi^2 = 4.19$, $p = .051$.

Statistical inferences regarding sertraline/placebo differences are consistent across the clinical outcome measures. All three measures are consistent in classifying individual study participants as treatment Responders or Nonresponders based on a 50% or greater improvement from baseline. The QIDS-C and QIDS-SR agreed in 84 of the 105 cases ($\kappa = .60$; $p < .001$), the QIDS-C and HAM-D agreed in 87 of the 105 cases ($\kappa = .66$; $p < .001$), and the QIDS-SR and the HAM-D agreed in 84 of the 107 cases ($\kappa = .60$; $p < .001$).

Seven vocal acoustic measures were significantly correlated with depression severity in the prior study: F₂ Coefficient of Variation (COV) ($r = -.17$; $r^2 = .03$); total recording time ($r = .20$; $r^2 = .04$); total pause time ($r = .29$; $r^2 = .08$); pause variability ($r = .38$; $r^2 = .14$); percent pause time ($r = .18$; $r^2 = .03$); speech/pause ratio ($r = -.22$; $r^2 = .05$); and speaking rate ($r = -.23$; $r^2 = .05$). Six of these seven measures, obtained from the automated/reading speech samples in this study, correlated significantly with QIDS-C total scores. These were total recording time ($r = .25$; $r^2 = .06$), total pause time ($r = .25$; $r^2 = .06$), pause variability ($r = .27$; $r^2 = .07$), percent pause time ($r = .20$; $r^2 = .04$), speech/pause ratio ($r = -.14$; $r^2 = .02$), and speaking rate ($r = -.18$; $r^2 = .03$). Two of the seven measures obtained from the Free speech samples were significantly correlated with the QIDS-C: pause variability ($r = .26$; $r^2 = .07$) and percent pause time ($r = .11$; $r^2 = .01$). Two speech measures that were not significantly correlated with depression severity in the prior study were significantly correlated in the present study. These were total vocalization time ($r = .13$; $r^2 = .02$) and the number of pauses ($r = .16$; $r^2 = .03$) during the automated/reading tasks.

Table 3 shows 51 treatment Responders (27 sertraline, 24 placebo) and 54 Nonresponders (28 sertraline, 26 placebo) based on the primary outcome (QIDS-C). Table 4 compares ten vocal acoustic measures that found significantly different baseline to end-point changes separating the treatment Responder and Nonresponder groups. The statistical significance of the within-group changes over time are also shown. In all eight timing measures (recording lengths, vocalization durations, pause durations and variability, speaking rate), the within-subject change was significant in the treatment Responders. Among Nonresponders, only four showed significant change over time; for the measures that did show significant change, the changes over time are of smaller magnitudes and/or in opposite directions than those of the treatment Responders.

Equations 1 and 3, applied to individual patient's baseline to Week 4 *change* in the vocal acoustic measures, were used to compute an *expected* probability that each participant would or would not respond to treatment. Patients with probability estimates of 0.5 or greater were classified as 'Predicted Responders' and the other patients were classified as 'Predicted Nonresponders'. Similarly, Equations 2 and 3 were applied to the baseline speech measures for each participant to compute an *a priori propensity* for response. The probability estimates above and below 0.5 were used to classify each participant as a 'Likely Responder' or 'Likely Nonresponder.' While only nine participants had a predicted response probability of .5 or greater (See Table 5), the 51 treatment Responders did have significantly higher mean response probability estimates (Mean = 0.137; SD = 0.326) than the Nonresponders (Mean = .025; SD = 0.111), $t(102) = -2.36$, $p = .020$. Similarly, response *propensity* estimates of the true Responders (Mean = 0.639; SD = 0.356) were significantly

higher than the mean propensity estimate of the Nonresponders (Mean = 0.467; SD = 0.374), $t(103) = -2.41$, $p = .018$.

Chi-square analysis of the cross-tabulated “Predicted Responder/Nonresponder” and “Likely Responder/Nonresponder” classifications with the “True Responder/Nonresponder” categories are shown in Tables 5 and 6. Both *a priori* logistic regression predictions of (a) treatment response *probability* based vocal acoustic changes using Equation 1, and (b) treatment response *propensity* using baseline speech measures using Equation 2, were statistically significant when QIDS-C criterion to define treatment response was examined. The statistical inferences that would have been drawn had the secondary outcomes measures been used to define treatment response are also presented.

Equation 4 provides the results of logistic regression modeling using baseline to Week 4 acoustic changes (Δ) as comparing the treatment Responders and Nonresponders observed in the present study. The model produces a test sensitivity estimate of 70.6% and specificity estimate of 79.2%, $\chi^2(5) = 41.3$, $p < .001$. Equation 5 models Responder/Nonresponder differences of the baseline acoustic measures, and suggests a potential test sensitivity and specificity of 62.7% and 66.7% respectively, $\chi^2(5) = 17.0$, $p = .004$. Since these models are maximized to “fit” the speech data obtained in this study, the repeatability/generalizability of these models requires independent evaluation in future research (like the tests of Equations 1 and 2 described above).

$$\begin{aligned} \text{Log Odds Response} = & 0.385 - 0.014 (\Delta \text{ Free speech total recording time}) + \\ & 0.832 (\Delta \text{ Free speech speech/pause ratio}) + 0.277 (\Delta \text{ Auto speech vocalization} \\ & \text{time}) + 0.079 (\Delta \text{ Mean } F_0 \text{ frequency}) + 0.013 (\Delta \text{ Mean } F_1 \text{ frequency}) \end{aligned} \quad \text{EQ. 4}$$

$$\begin{aligned} \text{Log Odds response} = & 3.854 + 0.014 (\text{Baseline Free speech number of pauses}) - \\ & 0.048 (\text{Baseline Free speech total pause length}) + 0.033 (\text{Baseline Free speech} \\ & \text{total vocalization time}) + 9.761 (\text{Baseline } F_2 \text{ COV}) \end{aligned} \quad \text{EQ. 5}$$

Discussion

This is the first study to investigate vocal acoustic speech measures as biomarkers of depression severity and treatment response in a multisite double-blind, randomized clinical trial. The results are generally consistent with prior research and describe reliable methods for collecting and analyzing clinically meaningful speech data that are readily available to clinicians and researchers. This study replicates prior studies that have shown depressed patients produce longer speech pause times, which shorten with clinical improvement following treatment (26, 27, 57, 58). Other studies have also found significant reductions in total recording and vocalization times due to increased speaking rates following antidepressant treatment of depressed patients with psychomotor retardation (35, 48).

An important aspect of this study is the replication and support of findings from a much smaller pilot study (48), strengthening confidence in the results of both studies. This replication was found despite important methodological differences that could have, but did not, influence study outcomes. The two studies used different depression severity measures and different methods for administering the assessments. The prior study was an open-label, six-week observational design, whereas the current study was a double-blind, four-week, placebo-controlled clinical trial. Despite these important differences in study design and implementation, the results obtained were very consistent across both studies. Replication of results relating vocal acoustic measures with depression severity and response to treatment using different clinical scales, different modes of administering the assessments, different

treatment durations, and different methods of blinding patients and providers provides strong evidence for generalizability of the results from both studies.

Six of seven correlations between depression severity and specific vocal acoustic measures found statistically significant in the prior study were also significant in the current study. More severe depression produces longer recordings with more pause time, more variable pause lengths, a greater percentage of pause time, smaller speech/pause ratios and slower speaking rates. This is consistent with clinical observations that depressed patients often present with psychomotor retardation, report difficulties with thinking and concentration, and have problems making decisions – such as choosing words (50). These symptoms manifest behaviorally during speech production (particularly during automated speech tasks) as objectively measurable biomarkers of depression. Treatment responders decrease total recording times during automated tasks primarily by increasing speaking rate and producing less overall vocalization. They also reduce the overall length of recordings and total vocalization, as well as the number, total time, and variability of pausing during free speech generation. The extent to which the behavioral manifestations of speech production evident in these vocal acoustic measures primarily reflects changes in patients' interests or motivations to engage externally or reflect internal changes associated with concentration or information processing capacity, or both, remains to be determined.

This study found the mean fundamental and first formant frequencies increased or remained unchanged among treatment responders, and decreased or remained unchanged in nonresponders. This pattern was not found in the prior study suggesting replication is needed to assure that the findings are not spurious. Prior research in healthy adults, however, using the same measurement extraction methods have found these frequencies to remain stable over 4-week periods (59).

It is important that the change defining an individual's *clinical* response to treatment is reflected in significant within-patient changes of the speech measures over time. The consistency of this relationship with treatment effectiveness, regardless of the nature of the treatment received (i.e., placebo or active drug), is also important. It suggests common neurophysiologic mechanisms influence symptoms of depression and the production of speech. All eight motor-based vocal acoustic measures distinguishing observed Responders from Nonresponders showed significant within-group change from baseline to end-point among treatment Responders, while only half did so in the Nonresponders. Of the speech measures that did show significant within-group change in the Nonresponders, the changes were of smaller magnitudes and/or directionally opposite to those seen in the Responders. The same pattern was seen in the prior study.

The statistical significance of the categorical predictions based on logistic models from just 32 participants of the prior study (13 Responders, 19 Nonresponders) applied to the present data from 105 randomized clinical trial participants is quite remarkable. The predictive validity of the models and consistent pattern of significant relationships between the clinical and acoustic measures suggests that the bio-behavioral linkage between speech production and depressive symptoms is robust. Although application of Equation 1 substantially underestimated the number of actual responders (see Table 5), the positive predictive value of the model was quite high (89%). Suboptimal coefficient estimations in the applied regression models could reflect the limited sample size of the prior study, modeling of 6-week (rather than 4-week) change in speech measures, or both. Ongoing studies continue to collect clinical and vocal acoustic data that will permit continued evaluation of the repeatability, validity, and extensions of the models applied to, and derived from, the present study. The present results are encouraging. Whether or not vocal acoustic measures of speech production might someday serve as objective biomarkers of depression severity for

patient screening or monitoring of clinical progress during treatment, as suggested by others (60), or could be used to identify patients likely to be treatment responsive (or resistant), as suggested by the present study, remains to be demonstrated by further research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The support, encouragement, and efforts of many people is gratefully acknowledged during the planning and conduct of this study, and during the analysis, interpretation, and presentation of these results, including Drs. John Greist, Michael Treglia, Charles Petrie, Joe Cappelleri, Alex Collie, Peter Snyder, Paul Maruff, Tracy Reyes, and Ben Barth. Funding support was provided by Pfizer, Inc. and through National Institute of Mental Health Small Business Innovation Research Grant NIMH (R44 MH068950).

References

1. Nock MK, Hwang I, Sampson N, Kessler RC, Angermeyer M, Beautrais A, et al. Cross-national analysis of the associations among mental disorders and suicidal behavior: findings from the WHO World Mental Health Surveys. *PLoS Med.* 2009; 6:e1000123. [PubMed: 19668361]
2. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Jama.* 2003; 289:3095–3105. [PubMed: 12813115]
3. Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry.* 1994; 51:8–19. [PubMed: 8279933]
4. Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry.* 1960; 23:56–62.
5. Williams JB. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry.* 1988; 45:742–747. [PubMed: 3395203]
6. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry.* 1979; 134:382–389. [PubMed: 444788]
7. Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, et al. The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol Med.* 2004; 34:73–82. [PubMed: 14971628]
8. Rush AJ, Bernstein IH, Trivedi MH, Carmody TJ, Wisniewski S, Mundt JC, et al. An evaluation of the quick inventory of depressive symptomatology and the Hamilton rating scale for depression: a sequenced treatment alternatives to relieve depression trial report. *Biological Psychiatry.* 2006; 59:493–501. [PubMed: 16199008]
9. Yang H, Cusin C, Fava M. Is there a placebo problem in antidepressant trials? *Curr Top Med Chem.* 2005; 5:1077–1086. [PubMed: 16181132]
10. Demitrack MA, Faries D, Herrera JM, DeBroda D, Potter WZ. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull.* 1998; 34:19–24. [PubMed: 9564194]
11. Greenberg RP, Bornstein RF, Greenberg MD, Fisher S. A meta-analysis of antidepressant outcome under “blinder” conditions. *J Consult Clin Psychol.* 1992; 60:664–669. discussion 670–667. [PubMed: 1401382]
12. Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: An analysis of the Food and Drug Administration database. *Journal of Clinical Psychopharmacology.* 2002; 22:40–45. [PubMed: 11799341]
13. Marcus SM, Gorman JM, Tu X, Gibbons RD, Barlow DH, Woods SW, et al. Rater bias in a blinded randomized placebo-controlled psychiatry trial. *Stat Med.* 2006; 25:2762–2770. [PubMed: 16345016]

14. Mundt JC, Greist JH, Jefferson JW, Katzelnick DJ, DeBrotta DJ, Chappell PB, et al. Is it easier to find what you are looking for if you think you know what it looks like? *J Clin Psychopharmacol.* 2007; 27:121–125. [PubMed: 17414233]
15. Psaty BM, Prentice RL. Minimizing bias in randomized trials: the importance of blinding. *JAMA.* 2010; 304:793–794. [PubMed: 20716744]
16. Blumer D, Zorick F, Heilbronn M, Roth T. Biological markers for depression in chronic pain. *J Nerv Ment Dis.* 1982; 170:425–428. [PubMed: 7086402]
17. Ising M, Horstmann S, Kloiber S, Lucae S, Binder EB, Kern N, et al. Combined Dexamethasone/Corticotropin Releasing Hormone Test Predicts Treatment Response in Major Depression-A potential Biomarker? *Biol Psychiatry.* 2006
18. Stefos G, Staner L, Kerkhofs M, Hubain P, Mendlewicz J, Linkowski P. Shortened REM latency as a psychobiological marker for psychotic depression? An age-, gender-, and polarity-controlled study. *Biol Psychiatry.* 1998; 44:1314–1320. [PubMed: 9861474]
19. O'Brien SM, Scott LV, Dinan TG. Cytokines: abnormalities in major depression and implications for pharmacological treatment. *Hum Psychopharmacol.* 2004; 19:397–403. [PubMed: 15303243]
20. Tsao CW, Lin YS, Chen CC, Bai CH, Wu SR. Cytokines and serotonin transporter in patients with major depression. *Prog Neuropsychopharmacol Biol Psychiatry.* 2006; 30:899–905. [PubMed: 16616982]
21. Placidi GP, Oquendo MA, Malone KM, Huang YY, Ellis SP, Mann JJ. Aggressivity, suicide attempts, and depression: relationship to cerebrospinal fluid monoamine metabolite levels. *Biol Psychiatry.* 2001; 50:783–791. [PubMed: 11720697]
22. Lee BH, Kim H, Park SH, Kim YK. Decreased plasma BDNF level in depressive patients. *J Affect Disord.* 2006
23. Insel TR, Charney DS. Research on major depression: strategies and priorities. *JAMA.* 2003; 289:3167–3168. [PubMed: 12813123]
24. Moses, JP. *The Voice of Neurosis.* Grune and Stratton; 1954.
25. Darby JK, Hollien H. Vocal and speech patterns of depressive patients. *Folia phoniat.* 1977; 29:279–291. [PubMed: 604242]
26. Szabadi E, Bradshaw CM, Besson JA. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. *Br J Psychiatry.* 1976; 129:592–597. [PubMed: 1000144]
27. Greden JF, Albala AA, Smokler IA, Gardner R, Carroll BJ. Speech pause time: a marker of psychomotor retardation among endogenous depressives. *Biol Psychiatry.* 1981; 16:851–859. [PubMed: 7295844]
28. Greden JF. Biological markers of melancholia and reclassification of depressive disorders. *Encephale.* 1982; 8:193–202. [PubMed: 6809444]
29. Greden JF, Carroll BJ. Decrease in speech pause times with treatment of endogenous depression. *Biol Psychiatry.* 1980; 15:575–587. [PubMed: 7397288]
30. Hardy P, Jouvent R, Widlocher D. Speech pause time and the retardation rating scale for depression (ERD). Towards a reciprocal validation. *J Affect Disord.* 1984; 6:123–127. [PubMed: 6231326]
31. Nilsson A. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatr Scand.* 1987; 76:235–245. [PubMed: 3673650]
32. Sobin C, Alpert M. Emotion in speech: the acoustic attributes of fear, anger, sadness, and joy. *J Psycholinguist Res.* 1999; 28:347–365. [PubMed: 10380660]
33. Stassen HH, Kuny S, Hell D. The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology.* 1998; 8:303–310. [PubMed: 9928921]
34. Wuyts F, De Bodt Ms, Molenberghs G, Remacle M, Heylen L, Millet B, et al. The dysphonia severity index: An objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research.* 2000; 43:796–809.
35. Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. *J Affect Disord.* 2001; 66:59–69. [PubMed: 11532533]

36. Alpert M, Shaw RJ, Pouget ER, Lim KO. A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia. *J Psychiatr Res.* 2002; 36:347–353. [PubMed: 12127603]
37. Cannizzaro MS, Cohen H, Rappard F, Snyder PJ. Bradyphrenia and bradykinesia both contribute to altered speech in schizophrenia: a quantitative acoustic study. *Cogn Behav Neurol.* 2005; 18:206–210. [PubMed: 16340393]
38. Cannizzaro MS, Harel B, Reilly N, Chappell P, Snyder PJ. Voice acoustical measurement of the severity of major depression. *Brain Cogn.* 2004; 56:30–35. [PubMed: 15380873]
39. 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; 47:829–837. [PubMed: 10916253]
40. Vogel AP, Morgan AT. Factors affecting the quality of sound recording for speech and voice analysis. *Int J Speech Lang Pathol.* 2009; 11:431–437. [PubMed: 21271920]
41. Cannizzaro MS, Reilly N, Mundt JC, Snyder PJ. Remote capture of human voice acoustical data by telephone: a methods study. *Clin Linguist Phon.* 2005; 19:649–658. [PubMed: 16147408]
42. Vogel AP, Maruff P, Snyder PJ, Mundt JC. Standardization of pitch-range settings in voice acoustic analysis. *Behavior Research Methods.* 2009; 41:318–324. [PubMed: 19363172]
43. Rosen KM, Murdoch BE, Folker JE, Vogel AP, Cahill L, Delatycki M, et al. Automatic method of pause measurement for normal and dysarthric speech. *Clinical Linguistics & Phonetics.* 2010; 24:141–154. [PubMed: 20100043]
44. DeBrotta, D.; Demitrack, M.; Landin, R.; Kobak, K.; Greist, J.; Potter, W. A comparison between interactive voice response system-administered HAM-D and clinician-administered HAM-D in patients with major depressive disorder. New Clinical Drug Evaluation Unit, 39th Annual Meeting; Boca Raton, FL. 1999.
45. Kobak K, Greist J, Jefferson J, Mundt J, Katzelnick D. Computerized assessment of depression and anxiety over the telephone using interactive voice response. *MD Computing.* 1999; 16:64–68. [PubMed: 10439605]
46. Mundt JC, Kobak KA, Taylor LVH, Mantle JM, Jefferson JW, Katzelnick DJ, et al. Administration of the hamilton depression rating scale using interactive voice response technology. *MD Computing.* 1998; 15:31–39. [PubMed: 9458661]
47. Moore HK, Mundt JC, Modell JG, Rodrigues HE, DeBrotta DJ, Jefferson JJ, et al. An Examination of 26,168 Hamilton Depression Rating Scale Scores Administered via Interactive Voice Response Across 17 Randomized Clinical Trials. *J Clin Psychopharmacol.* 2006; 26:321–324. [PubMed: 16702899]
48. 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; 20:50–64. [PubMed: 21253440]
49. Trevino A, Quatieri T, Malyska N. Phonologically-Based Biomarkers for Major Depressive Disorder. *EURASIP Journal on Advances in Signal Processing.* 2011; 2011:15.
50. APA. Diagnostic and Statistical Manual of Mental Disorders. 4. Washington, DC: American Psychiatric Association; 2000.
51. Mundt JC, DeBrotta DJ, Greist JH. Anchoring perceptions of clinical change on accurate recollection of the past: Memory Enhanced Retrospective Evaluation of Treatment (MERET®). *Psychiatry* 2007. 2007; 4:39–44.
52. Vogel AP, Fletcher J, Maruff P. Acoustic analysis of the effects of sustained wakefulness on speech. *J Acoust Soc Am.* 2010; 128:3747–3756. [PubMed: 21218906]
53. Dixon, WJ.; Massey, FJ. Introduction to Statistical Analysis. 4. McGraw-Hill; 1983.
54. O'Brien, RG.; Muller, KE. Applied Analysis of Variance in Behavioral Sciences. New York: Marcel Dekker; 1993. Chapter 8; p. 297-344.
55. Fleiss, JL. Statistical methods for rates and proportions. 2. New York: John Wiley & Sons; 1981.
56. Carletta J. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics.* 1996; 22:1–6.
57. Greden JF, Carroll BJ. Psychomotor function in affective disorders: an overview of new monitoring techniques. *Am J Psychiatry.* 1981; 138:1441–1448. [PubMed: 7294212]

58. Szabadi E, Bradshaw CM. Speech pause time: behavioral correlate of mood. *Am J Psychiatry*. 1983; 140:265. [PubMed: 6849453]
59. Vogel AP, Fletcher J, Snyder PJ, Fredrickson A, Maruff P. Reliability, stability, and sensitivity to change and impairment in acoustic measures of timing and frequency. *J Voice*. 2011; 25:137–149. [PubMed: 20171828]
60. Greden JF. Psychomotor monitoring: A promise being fulfilled? (Expert Commentary). *J Psychiat Res*. 1993; 27:285–287. [PubMed: 8295160]

Methodological design comparison between previous and current study to evaluate repeatability and generalizability of vocal acoustic speech measures as biomarkers of depression severity and treatment response.

Table 1

Study Design	Prior Study (Mundt et al 2007) (48)	Current Study (NCT00406952)
Treatment Indication	Major Depressive Disorder	Major Depressive Disorder
Study Design	Open-label treatment at physician's discretion, naturalistic observation of treatment outcomes	Placebo-controlled, double-blind, randomized clinical trial. Sertraline vs. placebo, forced-titration dosing regimen.
Inclusion/Exclusion	Any patient beginning treatment for new major depressive episode.	Screening/baseline clinician-rated HAM-D 22, extensive list of co-morbid exclusion criteria.
Treatment Duration	Six weeks	Four weeks
Primary Outcome Measure	Computer-administered HAM-D	Clinician-rated QIDS-C
Patient Recruitment (N)	Single US site (N = 33)	Eleven US sites (N = 105)

Table 2

Vocal acoustic characteristics measured for each type of speech sample obtained.

VA Measures	Types of Speech Samples Collected Via IVR telephone system		
	Free Speech	Automated Speech	Held Vowels
Total recording time			
Total vocalization time			
Total pause time			
Number of pauses			
Mean pause length			
Pause variability (SD)			
Percent pause time			
Speech/pause ratio			
Speaking rate (syllables/sec)			
Mean, SD, COV of F ₀			
Mean, SD, COV of F ₁			
Mean, SD, COV of F ₂			

SD = Standard Deviation; COV = Coefficient of variation (ratio of SD of frequencies divided by the mean frequency)
F₀ = Fundamental frequency; F₁ = First Formant frequency; F₂ = Second Formant frequency;

Table 3

Means ($\pm$ Standard deviations) of primary and secondary clinical measures of depression severity by treatment assignment.

Treatment	N	Depression Measure	Mean Score ($\pm$ Standard Deviation)		Number of Participants			
			Baseline	Week 4	Change	Responders ^a	Nonresponders	
Sertraline	55	QIDS-C	16.4 ($\pm$ 3.3)	8.5 ($\pm$ 4.3)	-7.9 ($\pm$ 4.9)	***	27	28
		HDRS-17	24.9 ($\pm$ 2.8)	11.5 ($\pm$ 5.8)	-13.4 ($\pm$ 5.7)	***	33	22
		QIDS-SR	15.8 ($\pm$ 3.7)	8.1 ($\pm$ 4.7)	-7.7 ($\pm$ 5.5)	***	29	26
Placebo	50	QIDS-C	17.4 ($\pm$ 2.8)	9.2 ($\pm$ 4.6)	-8.2 ($\pm$ 4.7)	***	24	26
		HDRS-17	24.6 ($\pm$ 2.5)	13.9 ($\pm$ 6.4)	-10.7 ($\pm$ 6.6)	***	20	30
		QIDS-SR	16.6 ($\pm$ 3.6)	8.8 ($\pm$ 4.6)	-7.9 ($\pm$ 4.3)	***	25	25

p < .001

^aResponder = 50% or greater reduction of depression severity from baseline to Week 4

Table 4

Means ($\pm$ Standard Deviations) for significant vocal acoustic measure differences between treatment responders^a and nonresponders defined by primary clinical outcome measure (QIDS-C).

Response Category	N	VA Measure	Mean ($\pm$ Standard Deviation)		Responder - Nonresponder				
			Baseline	Week 4	Change	Group <i>t</i> -test	<i>df</i>	<i>p</i> .	
Responders ^a									
	51								
		Free Speech total recording time	120.4 ($\pm$ 67.0)	94.4 ($\pm$ 37.5)	-26.0 ($\pm$ 50.0)	***			
		Free Speech total vocalization time	68.5 ($\pm$ 39.6)	58.0 ($\pm$ 23.5)	-10.5 ($\pm$ 31.3)	*			
		Free Speech number of pauses	118.6 ($\pm$ 73.5)	98.5 ($\pm$ 53.3)	-20.1 ($\pm$ 59.4)	*			
		Free Speech total pause time	51.9 ($\pm$ 31.5)	36.4 ($\pm$ 19.4)	-15.5 ($\pm$ 25.9)	***			
		Free Speech pause variability	0.69 ($\pm$ 0.25)	0.51 ($\pm$ 0.15)	-0.18 ($\pm$ 0.20)	***			
		Automatic Speech total recording time	30.2 ($\pm$ 9.28)	23.2 ($\pm$ 7.03)	-7.00 ($\pm$ 9.45)	***			
		Automatic Speech total vocalization time	15.6 ($\pm$ 3.87)	14.3 ($\pm$ 3.28)	-1.26 ($\pm$ 3.88)	*			
		Automatic Speech speaking rate	2.27 ($\pm$ 0.70)	2.89 ($\pm$ 0.85)	0.62 ($\pm$ 0.82)	*			
		Mean F ₀ frequency	151.8 ($\pm$ 36.5)	153.3 ($\pm$ 35.7)	1.45 ($\pm$ 7.25)	NS			
		Mean F ₁ frequency	546.8 ($\pm$ 67.1)	558.2 ($\pm$ 51.8)	11.4 ($\pm$ 48.9)	NS			
Nonresponders									
	54								
		Free Speech total recording time	98.0 ($\pm$ 57.9)	107.0 ($\pm$ 67.7)	8.93 ($\pm$ 49.9)	NS	3.58	103	.001
		Free Speech total vocalization time	51.3 ($\pm$ 27.5)	61.6 ($\pm$ 36.8)	10.2 ($\pm$ 29.3)	*	3.50	103	.001
		Free Speech number of pauses	97.1 ($\pm$ 58.7)	108.2 ($\pm$ 79.9)	10.3 ($\pm$ 54.6)	NS	2.72	102	.008
		Free Speech total pause time	46.7 ($\pm$ 34.3)	46.3 ($\pm$ 34.6)	-1.00 ($\pm$ 27.7)	NS	2.75	102	.007
		Free Speech pause variability	0.70 ($\pm$ 0.26)	0.61 ($\pm$ 0.20)	-0.09 ($\pm$ 0.20)	**	2.26	102	.026
		Automatic Speech total recording time	28.9 ($\pm$ 7.87)	25.2 ($\pm$ 7.00)	-3.68 ($\pm$ 6.90)	***	2.06	103	.042
		Automatic Speech total vocalization time	14.6 ($\pm$ 3.30)	15.9 ($\pm$ 3.56)	1.24 ($\pm$ 4.38)	*	3.10	103	.003
		Automatic Speech speaking rate	2.46 ($\pm$ 1.02)	2.70 ($\pm$ 0.88)	0.24 ($\pm$ 0.95)	NS	-2.19	103	.031
		Mean F ₀ frequency	157.3 ($\pm$ 33.6)	155.7 ($\pm$ 33.5)	-1.65 ($\pm$ 7.95)	NS	-2.09	103	.040
		Mean F ₁ frequency	570.5 ($\pm$ 62.1)	556.6 ($\pm$ 59.9)	-14.3 ($\pm$ 46.3)	*	-2.76	103	.007

p. .001

Biol Psychiatry. Author manuscript; available in PMC 2013 October 01.

Repeatability and generalizability of treatment response^a *probability* in current study using baseline to end-point changes in vocal acoustic measures based on logistic regression model derived from prior study.

Table 5

Clinical Treatment Outcome Measure (QIDS-C) ^b	Predicted by Logistic Regression Model (EQ. 1)	
	Responder	Nonresponder
Responder	8	43
Nonresponder (HAM-D-17) ^c	1	52
Responder	8	45
Nonresponder (QIDS-SR) ^d	1	50
Responder	8	47
Nonresponder	1	49

^aResponder = 50% or greater reduction of depression severity from baseline to Week 4

^bTwo-sided Fisher exact test, $p = .015$

^cTwo-sided Fisher exact test, $p = .031$

^dTwo-sided Fisher exact test, $p = .033$

Repeatability and generalizability of treatment response^a *propensity* in current study using baseline (only) vocal acoustic measures based on logistic regression model derived from prior study.

Table 6

Clinical Treatment Outcome Measure (QIDS-C) ^b	Predicted by Logistic Regression Model (EQ. 2)	
	Responder	Nonresponder
Responder	34	17
Nonresponder	25	29
(HAM-D-17) ^c		
Responder	34	19
Nonresponder	25	27
(QIDS-SR) ^d		
Responder	36	19
Nonresponder	23	28

^aResponder = 50% or greater reduction of depression severity from baseline to Week 4

^bTwo-sided Fisher exact test, $p = .049$

^cTwo-sided Fisher exact test, $p = .117$

^dTwo-sided Fisher exact test, $p = .050$