Evaluation of Depression Severity in Speech

Zhenyu Liu, Bin Hu(✉), Fei Liu, Huanyu Kang, Xiaoyu Li, Lihua Yan, and Tianyang Wang

Ubiquitous Awareness and Intelligent Solutions Lab, Lanzhou University, Lanzhou, China
{liuzhyl2, bh, fliul4, kanghyl5, yanlhl4, tywangl4}@lzu.edu.cn, li461547885@163.com

Abstract. Depression is a frequent affective disorder, leading to a high impact on patients, their families and society. Depression diagnosis is limited by assessment methods that rely on patient-reported or clinician judgments of symptom severity. Recently, many researches showed that voice is an objective indicator for depressive diagnosis. In this paper, we investigate a sample of 111 subjects (38 healthy controls, 36 mild depressed patients and 37 severe depressed patients) through comparative analysis to explore the correlation between acoustic features and depression severity. We extract features as many as possible according to previous researches to create a large voice feature set. Then we employ some feature selection methods to form compact subsets on different tasks. Finally, we evaluate depressive disorder severity by these acoustic feature subsets. Results show that interview is a better choice than reading and picture description for depression assessment. Meanwhile, speech signal correlate to depression severity in a medium-level with statistically significant ($p < 0.01$).

Keywords: Depression severity · Speech · Acoustic feature · Feature selection · PHQ-9

1 Introduction

The increase in the prevalence of clinical depression in human beings has been linked to a range of serious outcomes. It is a common mental disorder lasting for a long period and leads to a high impact on patients, their families and society. Depression is associated with half of all suicides and a significant economic burden [1]. The World Health Organization (WHO) estimated that about 350 million people of all ages suffer from this disease [2]. Moreover, depression is estimated to become the second greatest disease burden in the world by the year 2020.

However, current depression diagnosis methods almost rely on patient self-report and professional interview of symptom severity [3]. The patient self-report, like Self-rating Depression Scale (SDS) [4], risks a range of subjective biases. Similarly, professional interview varies depending on their clinical experience and the diagnostic methods used (e.g., Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [5]). So, an objective and convenient method for depression evaluation is necessary.

G.A. Ascoli et al. (Eds.): BIH 2016, LNAI 9919, pp. 312–321, 2016.
DOI: 10.1007/978-3-319-47103-7_31

Developments in affective sensing technology (e.g., facial expression, body gesture, speech, motion, eye movement, etc.) will potentially enable an objective depression evaluation method. Among these technologies, speech signals can be collected easily by non-invasive and portable instrument. Voice of depressed individuals reflect the perception of qualities such as monotony, slur, low intensity and less fluctuation [6]. Vocal characteristics have been verified to change with a speaker's mental condition and emotional state [7–9]. Such changes are complicated processes involving coordination of several brain areas and peripheral muscle controls [10]. And, researches support the feasibility and validity of vocal acoustic measures of depression severity [11, 12]. Therefore, we focus on depressed patients' speech analysis.

At early age, many researchers aimed at the correlation between depression and some particular speech features [13, 14]. A lot of experiments have been conducted to reveal relevance between depression and various acoustic features, like pitch, jitter, speaking rate, formants, Mel-Frequency Cepstral Coefficient (MFCC) and so on. Low et al. [15] and Mundt et al. [3] illustrated relation between depressive severity and some acoustic features. Lately, automatic detection approaches of depression have been investigated. Alghowinem et al. investigated and compared different features on depression classification. And, She figured out that spontaneous speech gives better results than reading [16]. Many researchers believe that feature combination optimization may lead to progress of recognition accuracy. Moore et al. proposed new feature sets with good performance on depression classification [7].

In this paper, we speculate speech signal correlate with severity of depression in a way. In order to validate our hypothesis, we take two steps: First, choose a feature set through comparing the classification accuracy in different tasks. Second, explore the correlation between the feature set and severity of depression.

The rest of this paper is organized as follows: Sect. 2 is a presentation of the details of our method and experiment, consisting of seven parts: the participants and their basic information, the procedure of experiment, data collection, data preprocessing and feature extraction, feature selection, classifiers, and correlation analysis. In Sect. 3, we showed the results of our experiment. Following this, we presented a discussion in Sect. 4 and in Sect. 5 conclusions were draw.

2 Method

2.1 Participants

111 participants' (54 males, 57 females) data from an ongoing study in Beijing and Lanzhou, China, were used for the experimental validation. These participants, with the age range of 18-55, were selected by psychiatrists following Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). All participants were asked to sign informed consent, fill in basic information and a series of scales. These basic information of subjects are summarized in Table 1.

All the participants were interviewed by a psychiatrist to finish the Patient Health Questionnaire-9 (PHQ-9) [17]. They were divided into three groups according to the PHQ-9 scores: 38 healthy control subjects (PHQ-9 < 5), 36 mild depressive patients

Table 1. Basic information of subjects

Parameter	Male subjects	Female subjects
Number of subjects	54	57
Average age (years)	36.6 ± 10.3	40.5 ± 10.8
PHQ-9 score	10.8 ± 8.1	11.7 ± 8.9
Recordings	29	
Tasks	Interview, reading, picture description	
Native language	Chinese	

($5 \leq$ PHQ-9 < 17) and 37 severe depressive patients (PHQ-9 ≥ 17). This three groups division can describe the change trend of speech features and keep relative larger subjects of each groups. The results are showed in Table 2.

Table 2. Basic information of groups

Parameter	Male subjects			Female subjects		
	Healthy	Mild	Severe	Healthy	Mild	Severe
Number of subjects	19	17	18	19	19	19
Average age (years)	36 ± 9.6	37.5 ± 10.9	36.2 ± 11	40.3 ± 11.1	40.5 ± 11.1	40.3 ± 10.8
PHQ-9 score	1.9 ± 1.5	10.9 ± 3.7	20 ± 3.1	1.3 ± 1.5	11.9 ± 3.9	21.7 ± 3.2

2.2 The Procedure of Experiment

Our experiment comprises three parts: interview, reading and picture description. Each part can be divided into three groups in terms of its induced emotion: positive, neutral and negative. To counteract the sequence effect of evoked emotion, the emotion order of each participant is assigned randomly. Details of the experiment follows below.

Interview. The interview part consisted of 18 questions. These questions are divided into three groups according to emotion valence: 6 positive, 6 neutral and 6 negative. These topics came from DSM-IV and some depression scales which are often used in depressive disorder diagnosis. For examples: What is your favorite TV program? What is the best gift you have ever received [18]? Please describe one of your friends. How do you evaluate yourself? What makes you desperate?

Reading. This part consisted of a short story named "The North Wind and the Sun", which was often used in acoustic analysis in international, multilingual clinical research, and three groups words with positive (e.g., outstanding, happy), neutral (e.g., center, since) and negative (e.g., depression, wail) emotion valence. Positive and negative words were selected from affective ontology corpus created by Lin [19], and neutral ones were selected from Chinese affective words extremum table [20]. All of them are commonly used words in Chinese and have close stroke number.

Picture Description. This part comprises four pictures. Three of them, which express positive (happy), neutral and negative (sad) faces, were selected from Chinese Facial Affective Picture System (CFAPS) and the last one with a "crying woman" came from Thematic Apperception Test (TAT) [18]. Participants were asked to describe these four pictures freely.

2.3 Data Collection

We collected recording data in a clean, quiet and soundproof laboratory. The whole experiment lasted about 25 min for one participant. During the course of recording, the subject was asked not to touch any equipment and keep the distance between mouth and microphone about 20 cm. A NEUMANN TLM102 microphone and a RME FIREFACE UCX audio card with 44.1 kHz sampling rate and 24-bit sampling depth were used for collecting voice signals. All recording data were saved as uncompressed WAV format. During the whole experimental process, ambient noise was required under 60 dB to prevent interference with subject's audio signals.

In the experiment, 29 recordings for every single participant were stored and named as 1 to 29 in a determined sequence. The details were as follows: The positive, neutral and negative interview recordings are named as 1–6, 7–12 and 13–18 separately. The record of the short story is name as 19. The readings of six word groups are named as 20–21, 22–23 and 24–25 in accordance with the sequence of positive, neutral and negative emotion. 26–28 were the picture description with the same order to reading part. The record of TAT was numbered as 29.

2.4 Data Preprocessing and Feature Extraction

All recordings are segmented and labeled manually. Only subjects' voice signal are reserved for analysis. Preprocessing mainly includes of filtering (a band-pass filter with 60–4500 Hz), framing, windowing and sometimes endpoint detection for some particular feature extraction. Each frame is 25 ms length with 50 % overlap. Voice characteristics can be divided into two categories: acoustic and linguistic features [21, 22]. The latter will not be analyzed since we are aiming at general characteristics for depressed speech regardless of the language used. Several software tools are employed for extracting sound features. We used the open-source software 'openSMILE' [23], VOICEBOX [24] and Praat [25] to extract 1753-dimension features. These features will be used in the following feature selection, classification and correlation analysis.

There are two steps to get the final acoustic feature subset of the speech signal: First, the signals of story (19) and TAT (29) are excluded in this paper. So, only 27 recordings for every subject were analysed. Second, compute the average value of every feature in the same part and induced emotion for one participant. For example, the speech 1–6 are for interview in positive emotion and we stored the mean values of all features as the Data 1 (in Table 3). The details are presented in Table 3.

Table 3. Names of nine data sets

Task	Positive	Neutral	Negative
Interview	Data_1	Data_2	Data_3
Reading	Data_4	Data_5	Data_6
Picture description	Data_7	Data_8	Data_9

2.5 Feature Selection

Feature selection refers to selecting effective features for classification in universal feature set. It is a critical problem in preprocess of data mining to cope with the curse of dimensionality [26]. In our experiment, we utilize a two-stage feature selection method by combining a filter and a wrapper method to reduce the feature dimension. Filter approach only utilizes data to decide which features should be kept. In general, filter approach has an efficient searching strategy with a result tradeoff. With "wrapping" accuracy of classifier, wrapper method may lead to a better performance compared to filter. Combining both we may have a high efficient method.

Here are the details about our two-stage feature selection method. We combine the minimal-redundancy-maximal-relevance (mRMR) criterion [27] as the filter approach and the Sequential Forward Floating Selection (SFFS) algorithm [28] as the search strategy of the wrapper approach. On the first stage, a candidate subset is selected from the universal feature set by mRMR. On the second stage, final subset is obtained from the candidate subset by SFFS. The final feature subset is used for the following discussion. In this process, the Support Vector Machine (SVM) [29] and Leave-One-Out Cross-Validation (LOOCV) scheme are be employed for evaluating and testing. This feature selection scheme is carried out on the nine data sets separately, which means nine feature subsets will be gained. We named these nine feature subsets as fs_1, fs_2, … fs_9, etc.

2.6 Classifier

We intend to evaluate feature subset in a specific situation to measure the severity of depression by pattern classification approach. Three widely used classifiers were employed in this paper: SVM, Naïve Bayes (NB) [29] and Random Forest (RF) [30]. The Radial Basis Function (RBF) kernel function was chosen in LIBSVM package [31]. Compared with the dimensionality of feature set, the sample size is often so small that we use the LOOCV scheme in testing. LOOCV is a special cross-validation. More specifically, one sample is for testing and the others are for training within a process. This repeated for all the samples and the result is the average accuracy of all repeats.

2.7 Correlation Analysis

Our main target is to explore the correlation between vocal features and depression severity. The PHQ-9 is a brief depression assessment instrument with severity categories. It is the depression module of the Primary Care Evaluation of Mental Disorder

[32, 33] that was designed to be used in primary care [34] and provides scores on each of the nine DSM-IV criteria using a severity scale from "0" (not all) to "3" (nearly every day). In our research, one or more feature subsets selected from the nine sets based on classification accuracy are used to explore the relation between voice and depression severity. Principal Component Analysis (PCA) will be applied on the normalized data of these feature subsets. We observe the Pearson's Correlation Coefficient (r) and the corresponding significance level (ρ) between the first principal component (FPC) and the PHQ-9 score with significance being tested with a T-test.

3 Result

Table 4 shows the average classification accuracy of three groups with three classifiers on nine feature subsets respectively. Although the accuracy of interview on positive and negative is inferior to reading or picture description for male, interview is with the best performance on average accuracy. And, it has a minimal standard deviation. For both male and female interview is the best choice of three for speech signal collection.

Table 4. Classification accuracy using data on nine feature subsets separately

Gender	Task	Positive	Neutral	Negative	AVG	STDEV
Male	Interview	0.648	0.630	0.605	0.628	0.022
	Reading	0.537	0.525	0.605	0.556	0.043
	Picture description	0.506	0.598	0.475	0.526	0.064
Female	Interview	0.544	0.526	0.579	0.550	0.027
	Reading	0.444	0.432	0.608	0.495	0.098
	Picture description	0.608	0.491	0.462	0.520	0.077

Table 5 presents the Pearson correlation coefficients between FPCs from fs_1, fs_2 and fs_3 on interview and PHQ-9 scores separately with significance levels. The values of r and ρ show that these FPCs are related to depression severity at a moderate level and statistically significant for both male and female.

Table 5. Pearson's correlation coefficient (r) and corresponding significant level (ρ) between FPC from the data of fs_1, fs_2 and fs_3 in interview and PHQ-9 score separately

Gender	Parameter	Positive	Neutral	Negative
Female	r	−0.400	0.501	0.543
	ρ	0.002	0.000	0.000
Male	r	0.499	−0.481	−0.544
	ρ	0.000	0.000	0.000

Figures 1, 2 and 3 show the scatter diagram of FPCs from Data_1, Data_2 and Data_3 and PHQ-9 scores in order to observe the linear correlation directly with

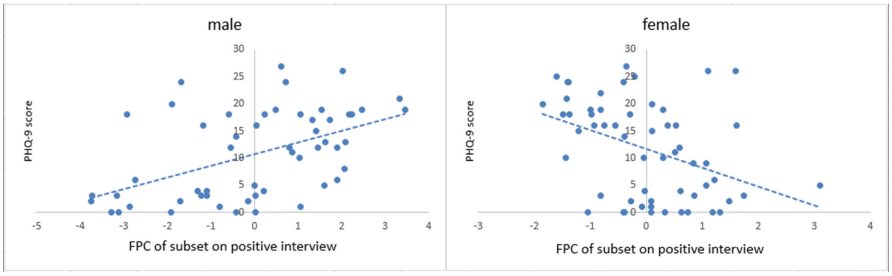

Fig. 1. Scatter diagram of FPC and PHQ-9 score on the data of positive interview

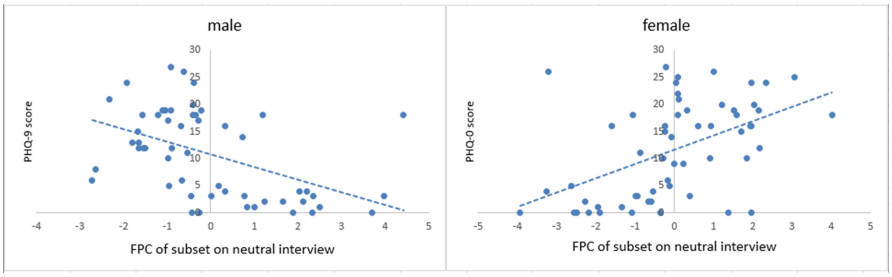

Fig. 2. Scatter diagram of FPC and PHQ-9 score on the data of neutral interview

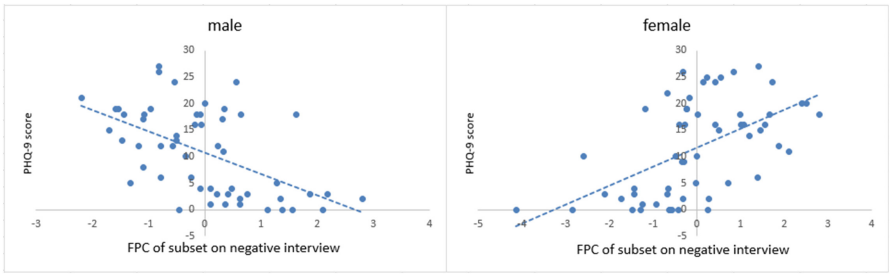

Fig. 3. Scatter diagram of FPC and PHQ-9 score on the data of negative interview

different emotion respectively. We can find that the negative questions perform better than positive and neutral on both male and female. All the correlation coefficients have opposite signs between genders.

4 Discussion

Our research aims at exploring the correlation between acoustic features and depression to evaluate depression severity. From results above, we get three points: First, the average classification accuracies (male: 0.57, female: 0.52) are probably limited for a

real system, nonetheless, they are much higher than chance level. Second, for both male and female, interview is the best pattern among these three ways to pick up the speech signals to evaluate severity of depression. Third, the correlation between the feature subsets from interview and PHQ-9 manifest that depressive severity is related to speech at a moderate level and the correlation is statistically significant.

Recording patterns may influence the classification performance. In our experiment, interview performs better than reading and picture description, which is consistent with the conclusion of Alghowinem et al. [23]. She pointed out that spontaneous speech gives a better results than reading. Both interview and picture description can be considered as spontaneous speech. However, picture description is worse than interview, we speculate that most of interview questions refer to the subject himself so that they are easy to get into emotional state.

In our further study, we intend to seek a more stable feature subset for depression assessment on a larger size of participants. And, we will combine speech features with other physiological feature (e.g. facial expression, gait, head movement etc.) to improve the classification accuracy.

5 Conclusion

Our work aims at an objective diagnostic aid supporting clinicians in evaluating severity of depression. The results confirmed our hypothesis by examining subjects' acoustic features on interview, reading and picture description patterns. Speech may be considered as a biomarker on depressive severity. Interview is a proper way to gain effective speech signal for depression assessment. The correlation between the FPC of speech feature subset and PHQ-9 score with statistical significance indicate that there may exist some features sets can be used to evaluate depression severity.

Acknowledgment. This work was supported by the National Basic Research Program of China (973 Program) (No. 2014CB744600), the Program of International S&T Cooperation of MOST (No. 2013DFA11140), the National Natural Science Foundation of China (grant No. 61210010, No. 61300231). Grateful acknowledgement is made to my classmates: Xiang Gao, Jinning Zhao, Xin Guo, Fei Heng and Lele He. They gave us considerable help by means of data collection, comments and criticism.

References

1. Lecrubier, Y.: Depressive illness and disability. Eur. Neuropsychopharmacol. **10**, S439–S443 (2000)
2. World Health Organization. http://www.who.int/mediacentre/factsheets/fs396/en/
3. Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S.: Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguist. **20**, 50–64 (2007)
4. Zung, W.W., Richards, C.B., Short, M.J.: Self-rating depression scale in an outpatient clinic: further validation of the SDS. Arch. Gen. Psychiatry **13**, 508–515 (1965)

5. American Psychiatric Association: DSM-III-R: Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, Arlington (1980)
6. Horwitz, R., Quatieri, T.F., Helfer, B.S., Yu, B., Williamson, J.R., Mundt, J.: On the relative importance of vocal source, system, and prosody in human depression. In: 2013 IEEE International Conference on Body Sensor Networks (BSN), pp. 1–6. IEEE (2013)
7. Moore, E., Clements, M., Peifer, J., Weisser, L.: Analysis of prosodic variation in speech for clinical depression. In: Proceedings of the 25th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, pp. 2925–2928. IEEE (2003)
8. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, D.M.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**, 829–837 (2000)
9. Quatieri, T.F., Malyska, N.: Vocal-source biomarkers for depression: a link to psychomotor activity. In: Interspeech, pp. 1059–1062
10. Vicsi, K., Sztaho, D., Kiss, G.: Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), pp. 511–515. IEEE (2012)
11. Harel, B., Cannizzaro, M., Snyder, P.J.: Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study. Brain Cogn. **56**, 24–29 (2004)
12. Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R.: Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiatry **72**, 580–587 (2012)
13. Nilsonne, Å., Sundberg, J., Ternström, S., Askenfelt, A.: Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Am. **83**, 716–728 (1988)
14. Scripture, E.: A study of emotions by speech transcription. Vox **31**, 179–183 (1921)
15. Ooi, K.E.B., Low, L.-S.A., Lech, M., Allen, N.: Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4613–4616. IEEE (2012)
16. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: Detecting depression: a comparison between spontaneous and read speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7547–7551. IEEE (2013)
17. Kroencke, K., Spitzer, R., Williams, J.: The phq-9: validity of a brief depression severity measure [electronic version]. J. Gen. Intern. Med. **16**, 606–613 (2001)
18. Hönig, F., Batliner, A., Nöth, E., Schnieder, S., Krajewski, J.: Automatic modelling of depressed speech: relevant features and relevance of gender. In: INTERSPEECH, pp. 1248–1252
19. DUTIR. http://ir.dlut.edu.cn/Group.aspx?ID=4
20. ShuJuTang. http://www.datatang.com/data/43216
21. Bandura, A., Pastorelli, C., Barbaranelli, C., Caprara, G.V.: Self-efficacy pathways to childhood depression. J. Pers. Soc. Psychol. **76**, 258 (1999)
22. Zhou, G., Hansen, J.H., Kaiser, J.F.: Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. **9**, 201–216 (2001)
23. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., Parker, G.: A comparative study of different classifiers for detecting depression from spontaneous speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8022–8026. IEEE (2013)
24. Kurniawan, H., Maslov, A.V., Pechenizkiy, M.: Stress detection from speech and galvanic skin response signals. In: 2013 IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS), pp. 209–214. IEEE (2013)

25. De Jong, N.H., Wempe, T.: Praat script speech rate (2008). Accessed 14 Oct 2008
26. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
27. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005)
28. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recogn. Lett. **15**, 1119–1125 (1994)
29. Mitchell, T.M.: Machine Learning. WCB/McGraw-Hill, Boston (1997)
30. Quinlan, J.R.: C4. 5: programs for machine learning. Elsevier (2014)
31. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. **13**, 415–425 (2002)
32. Farzanfar, R., Hereen, T., Fava, J., Davis, J., Vachon, L., Friedman, R.: Psychometric properties of an automated telephone-based PHQ-9. Telemed. e-Health **20**, 115–121 (2014)
33. Spitzer, R.L., Kroenke, K., Williams, J.B., Patient Health Questionnaire Primary Care Study Group: Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. JAMA **282**, 1737–1744 (1999)
34. Löwe, B., Unützer, J., Callahan, C.M., Perkins, A.J., Kroenke, K.: Monitoring depression treatment outcomes with the patient health questionnaire-9. Med. Care **42**, 1194–1201 (2004)