ELSEVIER

Journal of Neurolinguistics 20 (2007) 50–64

Journal of NEUROLINGUISTICS

www.elsevier.com/locate/jneuroling

Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology

James C. Mundt[a,*], Peter J. Snyder[b,c], Michael S. Cannizzaro[b,d], Kara Chappie[b], Dayna S. Geralts[a]

[a]*Healthcare Technology Systems, Inc., 7617 Mineral Point Road, Suite 300, Madison, WI 53717, USA*
[b]*Pfizer Global Research & Development, Groton Laboratories, Groton, CT, USA*
[c]*Department of Psychology, University of Connecticut, Storrs, CT, USA*
[d]*Department of Communication Sciences, University of Vermont, Burlington, VT, USA*

Received 10 January 2006; received in revised form 27 March 2006; accepted 4 April 2006

Abstract

Efforts to develop more effective depression treatments are limited by assessment methods that rely on patient-reported or clinician judgments of symptom severity. Depression also affects speech. Research suggests several objective voice acoustic measures affected by depression can be obtained reliably over the telephone. Thirty-five physician-referred patients beginning treatment for depression were assessed weekly, using standard depression severity measures, during a 6-week observational study. Speech samples were also obtained over the telephone each week using an interactive voice response system to automate data collection. Several voice acoustic measures correlated significantly with depression severity. Patients responding to treatment had significantly greater pitch variability, paused less while speaking, and spoke faster than at baseline. Patients not responding to treatment did not show similar changes. Telephone standardization for obtaining voice data was identified as a critical factor influencing the reliability and quality of speech data. This study replicates and extends previous research with a larger sample of patients assessing clinical change associated with treatment. The feasibility of obtaining voice acoustic measures reflecting depression severity and response to

*Corresponding author. Tel.: +1 608 827 2467; fax: +1 608 827 2444.
E-mail address: jmundt@healthtechsys.com (J.C. Mundt).

0911-6044/$ - see front matter
doi:10.1016/j.jneuroling.2006.04.001

treatment using computer-automated telephone data collection techniques is also established. Insight and guidance for future research needs are also identified.

Keywords: Depression assessment; Methodology; Speech; Voice acoustics; Telephone; Interactive voice response (IVR)

1. Introduction

1.1. Depression

Depression is a common, disabling, and life-shortening condition that occurs throughout the lifespan. More than 1 in 6 persons in the US experience a major depressive episode in their lifetime (Kessler et al., 1994). Depression is associated with half of all suicides and present an annual national economic burden exceeding $43 billion (Greenberg, Stiglin, Finkelstein, & Berndt, 1993; Mann, 2002). Billions of dollars are invested developing more effective and faster acting treatments with fewer adverse side effects. However, assessment tools currently used to measure depression severity and clinical change present limitations.

1.2. Assessment of depression by clinicians

Research in any domain is limited by the accuracy and precision of available measurement instruments. Despite recent advances in understanding the neurophysiology associated with depression, the clinical tools used in research of depression treatments have remained basically unchanged over the past 40 years.

Depression measures have historically relied upon either patient self-reported or clinician judgments of symptom severity. Both approaches attempt to derive reliable, objective measures of depression severity using subjective judgments that depend upon the training and experience of the clinician and patient. Such measures are critical for patient screening, severity assessment, measurement of change over time, identification of treatment benefits and remission, and detection of relapse or recurrence (Rush & Ryan, 2002).

The Hamilton Depression Rating Scale (HAMD; Hamilton, 1960) is the most widely used assessment of depression in clinical trial research. Despite noted psychometric weaknesses of the HAMD (e.g., Bech, Grosby, Husum, & Rafaelsen, 1984; Faries et al., 2000; Gibbons, Clark, & Kupfer, 1993; Santor & Coyne, 2001; Zimmerman, Posternak, & Chelminski, 2005), it is the most common 'gold standard' depression assessment. Efforts to standardize HAMD assessment (e.g., Williams, 1988), as well as intensive training of HAMD raters to improve rating reliability (Demitrack, Faries, Herrera, DeBrota, & Potter, 1998), have shown limited benefits in reducing measurement error. Other widely used clinician-based rating scales include the Montgomery–Asberg Depression Rating Scale (Montgomery & Asberg, 1979), the Inventory of Depressive Symptomatology (Rush, Gullion, Basco, Jarrett, & Trivedi, 1996), and the Quick Inventory of Depressive Symptomatology (Rush et al., 2003). Each requires rater training, practice, and certification to produce acceptable inter-rater reliabilities.

In addition to 'simple' measurement error problems inherent in clinician-based assessments of depression, systematic bias of such ratings in clinical trial research has