The value of computer-administered self-report data in central nervous system clinical trials

Bill Byrom1* & James C Mundt2

Abstract

Over the last decade, significant research has been conducted into the development of computer-delivered interviews and scoring methods to assess disease severity in a number of central nervous system indications. In particular, equivalence has been demonstrated between gold standard assessments (such as the Hamilton depression and Hamilton anxiety rating scales) delivered either by computer or by clinicians, in a number of clinical studies. Such computer-based approaches have a number of advantages including the elimination of inter-rater variance and prevention of inclusion bias when inclusion criteria are based upon baseline ratings. In addition, some studies have demonstrated that computer-administered ratings can be more sensitive to the direction of treatment-related changes than clinician ratings. Much of this evidence has led to recent US Food and Drug Administration acceptance of certain computer-administered assessments as primary endpoints in depression clinical trials.

Keywords Anxiety clinical trials, baseline score inflation, computerized clinical assessments, depression clinical trials, interactive voice response, placebo effect

Introduction

This article reviews the development and validation of computerized clinical assessments for use in clinical trials of drugs affecting the central nervous system (CNS), with emphasis on depression and anxiety. Key validation studies are discussed, demonstrating the equivalence of computerized assessments to clinician-made subjective assessments, with particular focus on the Hamilton depression (HAMD) and Hamilton anxiety (HAMA) rating interviews. The results of studies performed over the last five years using these techniques are reported, highlighting their particular utility in limiting the potential for inclusion bias in CNS clinical trials and providing particularly sensitive measures of change. Commentary is also given on the regulatory acceptance of this methodology and the direction of future research in this area.

Acceptability of self-report data

In recent years, patient-reported outcomes have become increasingly used and accepted as important endpoints in clinical drug submissions, and a number of successful drug approvals have been achieved using patient self-report data. For example, sertraline for premenstrual dysphoric disorder, paroxetine for social anxiety disorder, topiramate for migraine/headache and zalephon for sleep disorders, were all approved using data reported and collected by patients. There are a multitude of such examples, but more recently, there has been an increasing trend toward collecting such data using electronic means rather than paper diaries and questionnaires. The US Food and Drug Administration (FDA) approved eszopiclone for the treatment of insomnia based upon primary sleep-diary data collected using a telephone-based patient diary (via interactive voice response (IVR)). In addition, the approved FDA submission for ketorolac tromethamine as a treatment for ocular pain included patient self-report data collected using a personal digital assistant (PDA), and the prescribing information for the estradiol/levonorgestrel transdermal system Climara Pro, a hormonal transdermal patch for the treatment of postmenopausal symptoms, included patient-reported bleeding/spotting data collected using IVR during a one-year clinical trial.

The European Medicines Agency (EMEA) makes direct reference to the use of electronic means for the collection of patient self-report data in two of their clinical investigation guidance documents, for asthma [1] and oral contraceptives [2]. More significantly, in August 2004 the FDA stated that it would accept self-report data in general, as an approach to assessing symptom severity in clinical drug trials focusing on major depressive disorder in outpatients [3•]. The agency identified as acceptable approaches HAMD performed using a validated computer interview delivered by IVR, and the inventory of depressive symptoms (IDS) [4] and quick IDS (QIDS), in their various forms. This is particularly significant, as primary endpoints in depression trials have previously been exclusively based upon subjective ratings performed by the clinician.

In the remainder of this review, the use of computer-administered self-report data in CNS clinical trials and...
evidence of the validity of measuring clinical endpoints in this way, are explored.

**Rationale for the electronic application of self-report instruments**

There is a growing body of evidence in the literature illustrating the limitations of using paper diaries and questionnaires in recording patient-reported outcomes. Paper instruments are subject to missing data, conflicting entries, ambiguous completion and invented values [5,6]. One methodology study even detected patients completing their paper diary entries prior to the actual days for which the entries were ‘reported’ [7].

Electronic solutions, whether delivered using a personal computer, FDA or IVR, offer a number of advantages over paper approaches including in-built logic checking, automated questionnaire branching, time and date stamped entries and defined entry windows to prohibit retrospective or prospective entries. These features mean that patient-reported data collected electronically have greater quality and integrity than paper-recorded data. However, many of the instruments discussed in this review are not simply patient diaries converted from a paper version to computerized delivery, but are validated computerized clinical assessments, comprising a computer-delivered interview and scoring algorithm, that can replace the appraisals normally conducted by a clinician.

**Factors contributing to failed CNS trials**

The main challenge facing clinical trials in many CNS indications is in designing studies that provide the best opportunity to demonstrate true drug efficacy compared with placebo. High-placebo response rates in depression and anxiety studies are suspected to be a key reason for failure to detect significant separation of active treatment arms from placebo. In an analysis of new drug applications approved by the FDA between 1985 and 1997, Khan et al reported that only 47.7% of adequate-dose new antidepressant treatment arms demonstrated statistically significant separation from placebo, with 20.5% being supportive and 31.8% equivocal (ie, no significant difference from placebo) [8]. In a similar analysis of anxiolytic drug approvals by the FDA between 1985 and 2000, it was reported that only 48% of adequate-dose new drug treatment arms demonstrated statistically significant separation from placebo [9]. These results do not include development programs that may have been unnecessarily halted due to poor results prior to regulatory submission.

There may be many causes for the high proportion of failed clinical trials in depression, but baseline disease severity has been identified as a prominent factor [10] - the greater the severity of the baseline depression, the greater the active treatment effects are, compared with placebo. In this analysis, flexible dosing designs and the inclusion of fewer treatment arms and fewer female patients were also demonstrated to improve treatment-related differences. Patients entering depression clinical trials, and also clinical trials for CNS and other disorders, generally improve on placebo. Optimal study design may minimize placebo response and thereby reduce the likelihood of study failure.

Computerized administration of clinical assessments has been suggested as one method for accomplishing this aim [11,12]. Subjective investigator ratings have formed the basis of primary efficacy endpoints in clinical drug trials in depression, anxiety, schizophrenia and many other CNS disorders. Assessments are made following semi-structured interviews, and clinicians rate the severity of symptoms exhibited by patients on an ordinal scale against a number of individual scale items. Item scores, subscale scores and total scores form the basis of clinical measurements. Gold standard instruments include: (i) HAMD [13], the Montgomery-Åsberg depression rating scale (MADRS) [14], and the long and short forms of the IDS [4] for depression; (ii) HAMA [15] and the Leibowitz social anxiety scale (LSAS) [16] for anxiety; (iii) the Yale-Brown obsessive-compulsive scale (Y-BOCS) [17] for obsessive-compulsive disorder; and (iv) the brief psychiatric rating scale (BPRS) [18], the scale for assessment of negative symptoms (SANS) [19] and the positive and negative symptoms scale (PANSS) [20] for schizophrenia. These instruments are valuable in determining both the baseline disease severity and in measuring treatment-related changes.

Inter-rater differences in assessment reliability contribute to measurement error, and consequently to the inability to separate treatment arms. Regulatory guidance requires study sponsors to measure and control the effects of inter-rater variability by conducting pre-study training to standardize ratings [21]. Demitrack et al reported that such training has minimal effects [22]. In their study, the difference between minimum and maximum total scores of 86 raters, obtained during a training event that consisted of four recorded Hamilton interviews, varied from 14 to 21 points. This variability is significant relative to the range of scores likely in a depressed population - where most patients would be scored between 10 and 35. In addition, the time spent by investigators performing patient interviews, such as for HAMD, is often far less than the recommended 25 to 30 min; in one report 35% of interviews were conducted in < 10 min [23]. These limitations can be addressed by validated computer interviews, which deliver completely standardized assessments on each application.

**The validity of computerized assessments compared with clinician ratings**

There is a wealth of literature reporting the equivalence of electronic delivery of assessment instruments to conventional pen and paper versions. However, our focus in this review is the use of computerized clinical assessments as an alternative or adjunct to human assessments (in particular clinician subjective assessments).

Millard and Carver reported a comparison of quality-of-life data collected via a live telephone interview or collected using an automated interview delivered via IVR [24]. This study employed the SF-12 quality-of-life instrument for 229 individuals suffering from mild back pain. In their study, patients were randomly assigned to one of the assessment modalities and assessed only once. Despite methodological limitations, the study concluded that both human and
computerized approaches were equivalent when assessing the physical component summary scale of the quality-of-life instrument. However, the researchers concluded that patients were less reserved in reporting emotional concerns and mood to a computer compared with a human. This finding concurs with other studies. For example, Kobak et al reported that IVR had a greater sensitivity in detecting alcohol problems [25••], and Turner et al indicated a higher prevalence of reporting of sexual behaviors and drug abuse among adolescents when audio-assisted computer interviewing was used [26••]. There is evidence that in comparison to face-to-face interviews, patients may exhibit increased honesty when responding to sensitive, emotional or embarrassing questions using computer assessments.

Turning to more complex clinical assessments used in CNS studies, such as the HAMD, MADRS and HAMA, much research has been conducted into translating the semi-structured interviews and item scoring performed by trained clinicians, into fully structured computerized interviews with non-linear scoring algorithms to produce equivalent measures. Considerable data has been collected over more than ten years to support the clinical utility of this approach.

Significantly, Kobak et al provided a summary of validation evidence for a computer interview version of the HAM D from ten comparative studies performed from 1990 to 1999 [27••]. In the studies reported, computer interviews were delivered using PC and IVR technology. The IVR technology used recorded prompts to question the individual, who then selected response options using the telephone touchtone keypad. This simple interface is especially convenient, enabling assessment from the home as well as the clinic, and largely overcomes issues of functional illiteracy that might limit the ability to deliver complex clinical assessments via other electronic media. It is not uncommon, for example, for researchers to read out site-based questionnaires to patients unable to understand the written version. There have been few examples for the delivery of complex CNS clinical assessments using any other electronic solution in place of clinicians, and much of the review that follows refers exclusively to instruments delivered via IVR technology. In their review, Kobak et al reported a high correlation between clinician and computer HAMD scores (n = 1791, r = 0.81, p < 0.001) [27••]. Considering all 17 scale items, internal consistency measures of scale reliability [28] supported the use of the computer-administered HAMD. Across ten studies, Cronbach’s α ranged from 0.60 to 0.91 for the computerized HAMD, compared to a range of 0.41 to 0.91 for clinician ratings. In eight out of ten studies, Cronbach’s α for computer rating exceeded that for clinician rating. Test-retest reliability was also reported for scores obtained by computer interview in the same range as clinician test-retest correlations. Because of the high number of questions delivered during computer interview, it is likely that the computerized assessment approach is less prone to memory artificially inflating correlations between assessments repeated over a short time period, compared with clinician memory of individual item ratings.

In a study comparing clinician and computerized HAMA ratings in patients with a diagnosed anxiety disorder (n = 86) or major depression (n = 128) and healthy controls (n = 78) [29•], the computer-administered HAMA demonstrated good internal consistency (Cronbach’s α = 0.90) and test-retest reliability (r = 0.96), and a correlation of 0.92 (p < 0.001) was found between computer and clinician versions. No significant difference between mean HAMA scores obtained by computer or clinician was observed among the patients diagnosed with anxiety disorders, although a small significant difference in total scores was detected for the sample overall, which was not clinically relevant as it was within the order of the test-retest variability. A more recent study of 72 patients, comparing computer (IVR) to clinician HAMA, reported no significant differences in scores obtained using either method (mean scores = 15.53 and 16.13 for IVR and clinician, respectively), and similar measures of internal consistency and test-retest reliability were observed as in the earlier studies [30]. Additional validity was evidenced by good concordance between IVR HAMA scores and Beck anxiety inventory assessments [31] of the same patients.

More recently, research has been conducted into translating the semi-structured clinician interview and rating using the MADRS for assessment of depression severity, into an IVR interview and scoring algorithm [32••]. Depressed subjects (n = 60) were assessed by clinician and IVR in random order, and the two sets of resulting scores were not significantly different. Scale reliability measures, ie, internal consistency and test-retest reliability, were comparable between the methods. Although promising, the study was conducted in a relatively small sample of patients and further research is required to investigate the sensitivity of the IVR MADRS to clinical change over the time associated with different treatment regimens.

Other studies have examined and demonstrated good equivalence between electronic- and clinician-rated scores for the Y-BOCS [33] and LSAS [34]. These instruments, however, are essentially checklists, which makes the equivalence between clinician and self-report less surprising.

The different types of assessment relevant to CNS studies that have been adapted and used via IVR are listed in Table 1.

Validated computer assessments may offer many advantages to researchers. For example, the standardized interview delivered via computer eliminates the inter-rater variability observed in multicenter clinical trials. As discussed later in this review, it can also eliminate the sources of rater bias, such as inclusion bias and bias introduced by side-effect profile unblinding. The ability to perform clinical assessments from home rather than in the clinic, using remote electronic solutions such as IVR or PDA, provides many benefits to study design. Efficacy data may be collected more frequently to investigate, for example, speed of onset, and data may be collected in study phases in which regular clinical assessments are not convenient, such as during long-term safety extensions.
A final aspect of the validity of these techniques involves assessment of whether these computer-administered ratings are practical for use in clinical trials. It should be noted that they in no way replace the standard requirements for a clinician-patient meeting. Computer-administered ratings can, however, reduce the time required for lengthy rating interviews, permitting more time for clinicians to focus on the care and treatment of the patient.

A study of social anxiety disorder patients (n = 874) using IVR versions of the HAMD and LSAS [35] reported that 90% of patients rated computer assessment as ‘very easy’ or ‘easy’ [12•]. Similar results were reported in another study using IVR HAMD assessment (Figure 1), in which 76% of patients (n = 74) agreed that the questions were clearly stated and understandable (12% had no opinion), 93% agreed that the assessment was easy to complete (4% had no opinion), 79% found the IVR assessments convenient (12% had no opinion), and 48% indicated that the computer assessment allowed them to express their true feelings (30% had no opinion) [36•].

Baseline assessment of disease severity

Most CNS clinical trials using clinician-assessed subjective rating scales as endpoints require patients to have a baseline rating above or below a specified threshold to ensure that only those of an appropriate disease severity are included in the study. For example, depression studies using the HAMD endpoint commonly require a baseline HAMD score ≥ 20 as an inclusion criterion. Due to the considerable enrollment pressure on clinicians, and the fact that compensation is normally contingent upon randomization, conscious or subconscious bias can be introduced into baseline ratings to satisfy inclusion criteria. Inflation of scores at baseline introduces treatment-independent improvements post-baseline, which contribute to the placebo effect and

---

**Table 1. Assessments that have IVR adaptations relevant to CNS clinical trials.**

- Brief social phobia scale (BSPS)
- Changes in sexual functioning questionnaire (CSFQ)
- Cognitive function test battery (CFTB; Cognitive Drug Research Ltd/ClinPhone Group Ltd)
- Daily telephone assessment (DTA) - for depression
- Davidson trauma scale (DTS©)
- HAMA
- HAMD
- LSAS
- MADRS
- Memory-enhanced retrospective evaluation of treatment (MERET®)
- Mental health screener (MHS)
- Quality-of-life enjoyment and satisfaction questionnaire (Q-LES-Q)
- SF-12, SF-20 and SF-36 quality-of-life scales
- Symptoms of dementia screener (SDS)
- Telephonic remote evaluation of neuropsychological deficits (TREND)
- Work and social adjustment scale (WSAS)
- Y-BOCS

---

This figure is based on data from reference [36•].
exacerbate the problem of demonstrating treatment-related differences. A number of studies presented at scientific meetings have implicated this phenomenon by comparing ratings performed by clinicians with those produced by computer assessment. In one such study, reported by DeBrota et al (Figure 2) [37••], the researchers not only compared clinician HAMD ratings with computer scores collected using IVR at each of 8 weekly treatment visits, but also blinded the investigators to the true start of double-blind treatment by including a variable duration phase of placebo treatment after randomization without the knowledge of the investigators. At baseline, the researchers observed that the distribution of clinician scores was not normal and was truncated at the inclusion threshold score of 20. In fact, only four of the 291 patient assessments (1.4%) made by clinicians at week 0 were below 20. In contrast, the distribution of computerized assessment scores at this time point was normal, with 38% of patients rated below 20. Clinician ratings before randomization were, on average, 2.2 points higher than those obtained via IVR (p < 0.001). Following randomization, the concordance between clinician and computer ratings returned. During the final four visits after treatment assignment the correlation between methods was 0.84 (p < 0.01) and the mean clinician scores were not significantly different from the IVR-derived scores (they were 0.31 points higher than the IVR scores). The large differences in the distributions of ratings at baseline and the convergence of assessment methods following randomization is suggestive of inflated clinician ratings of disease severity at baseline.

Figure 2. Concordance of IVR and clinician HAMD ratings pre- and post-baseline for a study requiring baseline HAMD scores ≥ 20.

To qualify for this study, patients required a clinician-assessed HAMD score of ≥ 20 at visits 1 and 2. It was observed that clinician scores were poorly correlated with IVR scores at these visits. Moreover, the distribution of clinician scores at visits 1 and 2 was truncated at a score of 20, illustrating the inflation of baseline scores to meet the study inclusion criteria. Concordance of IVR and clinician assessments was re-established at later visits where study eligibility did not depend upon the HAMD ratings. This figure is based on data from reference [37••].

\[ \text{p} \] Patient self-assessment (IVR), \[ \text{c} \] clinician assessment.
In a second study, Rayamajhi et al compared clinician and IVR HAMD assessments obtained from 307 patients who participated in a double-blind placebo-controlled study for major depressive disorder [38••]. A lower baseline minimum score of 15 was required by this study and clinician HAMD scores were only an average of 1.09 points higher than IVR assessments at screening/baseline, with the agreement between assessment methods improved following randomization.

In a third study, IVR HAMD assessments were used to pre-qualify individuals prior to clinician assessment in a study of major depressive disorder [Mundt J, unpublished results]. Patients were qualified for clinician assessment if they were rated with an IVR HAMD rating of ≥ 20 at both screening and baseline. To enter the study, patients also required a subsequent baseline clinician rating of ≥ 22. Patients (n = 582) were qualified by IVR assessment and subsequently rated by the clinicians. Overall, a total of 1593 IVR HAMDs (1011 at screening visits, 582 at baseline) were performed. Of all rated patients with HAMD scores ≥ 20, 188 had computer HAMD scores of 20 or 21. Of these, only one patient was rated below 22 by clinicians and hence disqualified from entering the study. Interestingly, clinicians were not informed that the threshold for qualification using the IVR assessments was lower than the clinician rating inclusion criterion. The fact that only one marginal IVR-rated patient (scored 20 to 21) was excluded by clinicians could be evidence of confirmation bias among raters. However, it could be argued that IVR HAMDs are simply underestimating the disease severity, which is more accurately assessed by the clinician. The data from the two previous studies indicated lower IVR ratings at screening and baseline before concordance was restored between clinician and IVR ratings at subsequent assessments where inclusion/exclusion did not depend upon the ratings [37••,38••]. However, in the third study the mean IVR rating of qualified patients at baseline was significantly higher than that obtained via clinician rating (26.8 versus 25.4, respectively, p < 0.001), despite the use of a lower IVR qualification threshold, indicating that IVR does not systematically under evaluate depression severity.

An opposite effect to this observation of baseline score inflation among clinician ratings in studies requiring a minimum score for study entry has also been reported. Inclusion bias has been observed in studies requiring patients to present with a baseline score no greater than a defined threshold. A recent placebo-controlled pilot study of St John's Wort for the treatment of social phobia required a HAMD score ≤ 16 for inclusion into the study [39•]. Of the 40 patients that were assessed at baseline, none were excluded based upon clinician assessed HAMD. However, self-report HAMD data (obtained by paper questionnaire [40]) indicated that 16 out of 40 (40%) patients had a self-report HAMD score > 16, with self-report scores being an average of 5.05 points higher than clinician ratings (p < 0.001).

This phenomenon of inclusion bias is not confined to depression ratings. A study of 624 social anxiety disorder patients was presented at the 2002 American College of Neuropsychopharmacology meeting [41••]. In this study, patients required a clinician rated HAMA score of ≥ 20 to enter the open-label (active) 6-week run-in phase. Responders at week 6, defined as achieving a HAMA score ≤ 11 and exhibiting a minimum change from baseline HAMA of 50%, were then included in the subsequent 6-month double-blind relapse prevention phase. By comparison of IVR and clinician rated HAMA scores, the researchers reported inflation of clinician scores at baseline to meet the entry criterion of HAMA score ≥ 20, and deflation of clinician scores at the end of the open-label phase to meet the continuation criterion of HAMA score ≤ 11. Clinician scores were, on average, 2.7 points higher at baseline and 2.9 points lower at the end of the open-label phase compared with computerized assessments (p < 0.05).

**Sensitivity to detection of treatment effects**

In addition to the elimination of inclusion bias, computer-administered ratings have been associated with the measurement of smaller changes from baseline and larger treatment effects relative to placebo, compared with those observed using clinician ratings. The smaller changes from baseline can be explained by the elimination of baseline score inflation/deflation observed among clinician ratings, but the increased sensitivity to demonstrate treatment separation is a significant finding that may benefit future study design.

Rayamajhi et al reported the effects of 8 weeks of treatment with duloxetine, fluoxetine or placebo in 307 major depressive disorder patients (Figure 3) [38••]. Clinician-rated baseline adjusted least-square mean changes from baseline to endpoint HAMD were -5.99, -8.11 and -6.59 for placebo, duloxetine and fluoxetine, respectively. The corresponding mean changes from baseline calculated using IVR HAMDs were -4.04, -6.29 and -5.84, respectively. As discussed above, the lower changes from baseline within each treatment group may be associated with clinician score inflation at baseline. Petkova et al proposed the simultaneous use of both patient self-ratings and clinician ratings in non-psychotic and non-demented populations, so that the similarity between clinician ratings and patient self-ratings can provide a measure of bias in clinician ratings [42•]. Despite the smaller changes from baseline, analysis of IVR-rated assessments in this study gave evidence of increased relative treatment effects and greater power to demonstrate treatment separation. The analysis of both clinician- and IVR-rated data highlighted significant improvements in depression severity attributable to duloxetine compared with placebo: baseline adjusted least-square mean changes from baseline compared to placebo were -2.12 (p = 0.008) and -2.25 (p = 0.006) for clinician and IVR ratings, respectively. For fluoxetine-treated patients, however, clinician ratings showed little evidence of improvement relative to placebo, whereas the IVR ratings indicated a much larger improvement that approached statistical significance: baseline adjusted least-square mean changes from baseline compared to placebo were -0.6 (p = 0.536) and -1.8 (p = 0.076) for clinician and IVR ratings, respectively. The fact that the fluoxetine arm was underpowered in this study, having half as many patients as the duloxetine arm, makes this result all the more impressive.
This is not an isolated result. For example, using the IVR HAMA clinical assessment, similar indications of increased sensitivity to treatment-related effects have been reported [41••].

**Conclusions**

The data reported in the literature and at scientific meetings over the past decade, predominantly in the last five years, have demonstrated the potential of using computerized assessments in clinical trials. More significantly, they provide evidence of the equivalence between computer-delivered patient interviews and ratings, and the corresponding clinician-delivered versions. Examples cited in this review include the HAMD, HAMA and MADRS instruments commonly used in depression and anxiety clinical trials. Such evidence has been pivotal in recent FDA
acceptance of IVR HAMD as a primary endpoint in depression outpatient trials.

The studies reviewed here have highlighted the advantages and utility of computerized clinical assessments, in particular in elimination of the inclusion bias observed when subjective clinician ratings form the basis of patient eligibility. In addition, the reduction of inter-rater variability and reduced placebo effect due to elimination of baseline score inflation gives computerized clinical assessments increased sensitivity to detect treatment-related differences from placebo.

Because of the success in developing and validating equivalent measures of severity using computerized clinical assessments, approaches for taking greater advantage of the unique capabilities of computer-administered self-assessment continue to be developed.

One such development that has provided favorable results is the development of assessments to measure the speed of onset of CNS drugs. Assessments commonly used in depression studies, such as HAMD and MADRS, are designed for weekly use, but recent research suggests that patients receiving newer antidepressants may exhibit symptom improvement before the end of the first week of treatment [43]. Computerized assessments that can be self-completed at home using electronic diary solutions facilitate the reliable collection of such data, and can do so with minimal reporting burdens for patients [44]. For example, in a 12-week open-label trial using daily reports during the first week of treatment, 137 depressed patients randomized to one of two doses rated their pain symptoms, and emotional and physical improvements using an IVR system. Results demonstrated that patients receiving the higher dose reported significant symptom improvement as early as one day after treatment [45]. The daily telephone assessment (DTA©) utilizes the same approach to daily patient self-report of the severity of core depressive symptoms that might provide better precision of therapeutic onset of action [46]. Further research is required to develop computerized self-assessment instruments that can measure speed of onset of efficacy and adverse events for use in other indications.

Another area in which telephone-based clinical assessments are showing promise is in self-assessment of improvement. Although retrospective assessments of change can be more sensitive than serial assessments made at multiple time points [47], retrospective recall is problematic as memory is largely a reconstructive process and the accuracy of reconstruction is profoundly affected by the passage of time. Memory enhanced retrospective evaluation of treatment (MERET©) uses an IVR interface to permit patients enrolled in clinical studies to describe feelings and experiences related to their condition at baseline in their own voice and words. These descriptions are recorded and then played back to patients prior to making subsequent retrospective reports concerning change after treatment. Preliminary results with the technique have been encouraging [48], suggesting that the process may increase apparent treatment effects relative to current methods for obtaining patient global impressions of improvement.

A final area of ongoing research involves the capture and acoustical analysis of the speech patterns of patients. Recent studies of voice acoustical analysis of patients with extremely early-stage Parkinson's disease have produced exciting findings [49], and pilot research with depressed patients also appears promising [50]. A preliminary methodology study comparing the acoustical measures made from recordings obtained using state-of-the-art laboratory recording equipment and simultaneous recording over the telephone using an IVR system was also encouraging, indicating that the data obtained by both methods were highly comparable [51]. This opens the possibility of using such inexpensive techniques in large-scale clinical trials. Ongoing research, currently being funded by the National Institute on Mental Health, will continue to explore and develop these techniques.

These future directions will complement the existing achievements in this area to ensure that computerized clinical assessment, in particular using IVR, make a valuable contribution to current CNS clinical trials.

References


- Factor analysis identifying baseline disease severity as the major factor in antidepressant study failure. The paper also cites flexible dosing regimens, number of treatment arms and proportion of female patients as important factors.


- Useful review of factors contributing to failed CNS trials and patient acceptability of IVR computerized assessments.


- Short article describing the inter-rater variation in HAMD scores following an investigator training event.


- Describes computerized diagnostic assessments for psychiatric disorders. Patients demonstrated increased honesty to sensitive questions using computerized (IVR) assessment, leading to a higher prevalence of diagnosis of alcohol/drug abuse.


- Computer-assisted interviews provide increased reporting of sensitive issues including injection drug use and sexual contact.


- Valuable review and meta-analysis of validation studies comparing clinician- and IVR-administered HAMDs.


- Evidence for the concordance between clinician- and IVR-administered HAMA ratings.


- Results of a pilot study highlighting the concordance between clinician- and IVR-administered MADRS ratings.


- Good patient acceptability data using IVR clinical assessments.


- Study providing strong evidence of baseline score inflation with clinician-administered HAMD ratings.


- Results from a study highlighting baseline score inflation and increased sensitivity to detecting treatment effects using the IVR-administered HAMD compared with clinician-administered ratings.


- Report from a study demonstrating inclusion bias when a maximum HAMD score is required for study entry (baseline score deflation).


- A study demonstrating baseline score inflation with clinician-administered HAMA, and improved sensitivity to detect treatment effects using computerized assessment.


- Article suggesting that self-assessment should be used as a measure of clinician bias.


- Study identifying rapid onset treatment effects using daily IVR assessments.


- A study showing increased sensitivity of patient impression of improvement measures when using a telephone-based assessment with audio replay of a self-recorded anchor message describing baseline condition.

