# USING OUTCOME MEASURES TO PROMOTE BETTER OUTCOMES

Mark Zimmerman, Iwona Chelminski, Diane Young, Kristy Dalrymple

## Abstract

The previous manuscript suggested that clinicians must be aware of available psychotropics, the strength of their clinical trial data, and when that fails be aware of basic neuroscience principles in order to work towards clinical remission in the their patients. In order to use available psychotropics clinicians may have to embrace complex polypharmacy also addressed previously. A fourth concept needed to improve the quality of care and promote symptom remission is to measure outcome. Psychiatry is the only medical discipline in which quantified measurements of outcome are not the standard of care. In mental health clinical settings outcome evaluations are typically based on unstructured interactions that yield *unquantified* judgments of progress. This is at variance with other areas of medical care in which outcome is determined, in part, on the change of a numerical value. Body temperature, blood pressure, cholesterol values, blood sugar levels, cardiac ejection fraction, and white blood cell counts are examples of quantifiable variables that are used to evaluate treatment progress. In treating psychiatric disorders, standardized, quantifiable outcome measures exist for most major psychiatric disorders, yet they are rarely used in routine clinical practice. Recently, the term "measurement-based care" has been coined in reference to the use of standardized scales to measure the outcome of psychiatric treatment. In this article we review perceived obstacles in adopting a measurement-based care treatment approach, and illustrate how the use of self-report depression scales is feasible, acceptable to patients, and may improve outcome. A web-based system of outcome assessment is described that allows response and remission rates to be calculated in a group of patients. The tools and technology now exist to overcome the challenges posed by a measurement-based care approach towards care, and this will hopefully accelerate the incorporation of measurement into routine clinical practice and enable quality improvement efforts to be tested in a cost-effective manner.

**Key Words**: clinical remission, complex polypharmacy, measurement-based care, self-report depression scales

**Declaration of interest**: None

Mark Zimmerman, M.D., Iwona Chelminski, Ph.D., Diane Young, Ph.D., Kristy Dalrymple, Ph.D. From the Department of Psychiatry and Human Behavior, Brown Medical School, and the Department Psychiatry, Rhode Island Hospital, Providence, R.I.

**Corresponding author**
Mark Zimmerman, M.D., Bayside Medical Center, 235 Plain Street, Providence, RI 02905
e-mail: mzimmerman@lifespan.org

## Lack of measurement in clinical practice – An inadequate current standard of care

Imagine going to your primary care doctor with a fever and symptoms of an upper respiratory track infection. Your primary care provider puts his or her palm to your forehead and agrees that you feel warm. A course of treatment is recommended, you return in a couple of days, and he or she again feels your forehead and notes that you are cooler. Would you be happy with this approach towards care? Would you continue to see a doctor who evaluated your body temperature in this way? We would not accept this level of care from an internist, family practitioner, or pediatrician, and yet this is the community standard of care provided by most behavioral health clinicians when treating depression and other psychiatric disorders.

To determine the impact of treatment it is necessary to evaluate outcome. In mental health clinical settings this typically is based on unstructured interactions that yield unquantified judgments of progress. This is at variance with other areas of medical care in which outcome is determined, in part, on the change of a numerical value. Body temperature, blood pressure, cholesterol values, blood sugar levels, cardiac ejection fraction, and white blood cell counts are examples of quantifiable variables that are used to evaluate treatment progress. In the mental health field, standardized, quantifiable outcome measures exist for most major psychiatric disorders, yet they are rarely used in routine clinical practice. Thus, to determine the impact of treatment it is not simply a matter of *evaluating* outcome, but rather a matter of *measuring* outcome.

Let us take depression as an example. Despite its high prevalence, high morbidity, and high consumption of health care resources, the standard of care for evaluating the efficacy of treatment for depression in clinical practice is based on unquantified, nonstandardized, clinical impressions. The lack of systematic assessment, or measurement, can impede treatment outcome because patients might report clinical improvement without also noting the presence of residual symptoms. Clinicians may therefore be unaware of an inadequate or incomplete treatment

Submitted June 2010, Accepted November 2010

response and will incorrectly conclude that no changes in treatment are needed. If clinicians are not making appropriate treatment recommendations, then outcome is likely to be poorer. Routine measurement of outcome with reliable and valid instruments may improve outcome by providing clinicians with information that will enable them to modify their treatment approach with the individual patient. The results of two surveys of psychiatrists, however, suggest that standardized scales are not being used to evaluate outcome in clinical practice.

Gilbody et al. (2002) surveyed 340 psychiatrists in the United Kingdom regarding their use of outcome measures. Only 11.2% of the psychiatrists routinely used standardized measures to assess outcome when treating depression and anxiety disorders. More than half of the clinicians indicated that they never used standardized measures to evaluate outcome. The authors did not ask the respondents why they were disinclined to use scales to measure outcome; however, they noted that several respondents included comments on the questionnaires indicating that they thought such scales were simplistic, not useful in clinical practice, of questionable reliability and validity, or overly burdensome and costly to implement routinely.

Zimmerman and McGlinchey (2008b) conducted a similar survey of 314 psychiatrists in the United States. They, too, found that the vast majority of psychiatrists did not routinely use scales to monitor outcomes of treating depression (**table 1**). More than half of the psychiatrists indicated that they never or rarely used scales to monitor outcome, and less than 10% almost always used scales to monitor outcome of depression treatment. They compared the characteristics of psychiatrists who reported using scales frequently or almost always to the rest of the group and found no differences between the two groups in gender, age, years of practice, or practice setting.

Subjects who reported never, rarely, or only sometimes using scales to monitor outcome were asked the reasons for not routinely using scales in their clinical practice. More than one-quarter of the subjects indicated that they did not believe using scales would be clinically helpful, that they take too much time to use, and that they were not trained in their use (**table 2**).

The results of these two surveys found that psychiatrists typically do not use standardized scales of established reliability and validity when treating patients with depression. One issue identified as an obstacle in their use is the perceived burden of scale completion. If the payers of the delivery of mental health treatment increasingly encourage, or require, the measurement of outcome, then the user friendliness of measurement tools, as well as their reliability and the validity, will be critical to their widespread adoption. Clinicians are already overburdened with paperwork, and adding to this load by requiring repeated detailed evaluations with such instruments as the Hamilton Rating Scale for Depression (Hamilton 1960) is unlikely to meet with success. Self-report questionnaires are a cost-effective option because they are inexpensive in terms of professional time needed for administration, and they correlate highly with clinician ratings. To be sure, there are also limitations with self-report questionnaires such as response set biases, and their use may be limited by the readability of the scale and literacy of the respondent. However, self-report scales are free of clinician bias, and are therefore free from clinician overestimation of patient improvement (which might occur when there are incentives to document treatment success).

Suggestions of the beneficial impact of measuring outcome come from the STAR*D trial, the largest study of the treatment of depression ever conducted. In the acute phase component of STAR*D, during which patients were treated with citalopram for up to 12 weeks, the rates of response and remission were similar to rates typically reported in controlled efficacy studies. Trivedi et al. (2006) suggested that an adequate treatment response might have been more difficult to achieve in STAR*D than typical industry-funded efficacy studies because patients with comorbid disorders, who are less responsive to treatment, were not excluded. They attributed the better than expected (albeit modest) response and remission rates to the adoption of a system of measurement-based care. That is, they indicated that the use of frequent, standardized, quantitative assessments to guide treatment decision-making contributed to an increased likelihood of a positive outcome, and they recommended that a measurement-based care approach towards clinical management be adopted in routine clinical practice.

**Table 1.** *Reported Frequency of Use of Standardized Scales to Measure Outcome in the Treatment of Depression by 314 Psychiatrists Attending a CME Conference*[a]

| Frequency | % | N |
|---|---|---|
| Never | 28.8 | 88 |
| Rarely | 32.0 | 98 |
| Sometimes | 21.2 | 65 |
| Frequently | 11.4 | 35 |
| Almost all the time | 6.5 | 20 |

[a] Data was excluded for 8 subjects because it was missing (n=7), or more than one response was checked (n=1)

**Table 2.** *Reasons Given by Psychiatrists for not Using Standardized Scales to Measure Outcome in the Treatment of Depression (n=248)*[a]

| Reasons | % | N |
|---|---|---|
| Do not believe it would be clinically helpful | 27.8 | 69 |
| Do not know what measure to use | 20.6 | 51 |
| Takes too much time | 33.9 | 84 |
| Too disruptive of clinical practice | 19.0 | 47 |
| Wasn't trained to use them | 34.3 | 85 |
| Other | 28.6 | 71 |

[a] Three subjects who indicated that they never, rarely, or sometimes used scales did not respond to this question.

The call for measurement-based care is consistent with the Center for Medicare and Medicaid Services Physician Quality Reporting Initiative, which is intended to increase clinicians' motivation to systematically evaluate outcome by providing financial incentives to monitor outcome. At present, the level of financial incentive is modest (1.5% of fees).

## One reason why measurement may be important in treating depression – improved detection of residual symptoms

As noted previously, the authors are using major depressive disorder as a prototype psychiatric disorder where outcomes may be measured. Readers should also note that these principles could apply to any of the myriad of psychiatric disorders that psychopharmacologists treat. Again, would a physician treat diabetes without measuring glucose levels? Or treat hypertension without measuring blood pressure? Or treat a febrile illness without measuring body temperature? Of course not. Measurement provides the clinician with information regarding the degree and completeness of treatment success, and suboptimal outcome in the treatment of diabetes, hypertension, hypercholesterolemia, or an infection would prompt intervention. The same should be true in the treatment of depression. As noted in the first paper of this issue, there may not be any blockbuster, cure all psychotropics in the research pipeline, and psychopharmacologists must get better at using the agents we have safely and aggressively. Outcome measures will aid us in detected residual symptoms of patients in non-remitted states. This will ultimately prompt clinicians to increase their level of care to aim for remission.

Research has consistently demonstrated that residual symptoms of depression in patients who have been identified as treatment responders are at increased risk for relapse. For example, Paykel et al. (1995) followed up 64 treatment responders for 15 months. Treatment response was defined as failure to meet full major depression criteria for 2 months. Patients who scored above 8 on the HRSD were three times more likely to relapse during the follow-up interval than patients scoring 8 or below (76% vs. 25%). Thase et al. (1992) followed 48 depressed patients who responded to 16 weeks of cognitive-behavior therapy for one year after the completion of treatment. Responders scored 10 or less on the HRSD and their scores improved at least 50% from baseline. The responders were subdivided into those who did (remission) and did not (response) score 6 or less on the HRSD for the last two months of treatment. Patients who scored 6 or less were significantly less likely to relapse than patients who scored 7 through 10. Several other follow-up studies have similarly found that the presence of residual symptoms in patients who responded to treatment predicted poorer outcome (Evans et al. 1992, Judd et al. 2000, Simons et al. 1986).

The data is clear—the presence of residual symptoms in depressed patients who have improved with treatment predicts poorer long-term outcome. How well do clinicians detect such residual symptoms? We are not aware of studies that have addressed this question. However, as demonstrated in the STAR*D study, residual symptoms are common. Trivedi et al (2006) found that two-thirds of the patients experienced mild-moderate levels of symptoms at the end of the acute phase of treatment with citalopram. To be sure, the remission rate during the acute phase in the STAR*D trial was modest despite the use of measurement to guide treatment decision-making. However, cumulative remission rates after multiple levels of treatment were greater than 60% (Quitkin et al. 2005, Rush et al. 2006), and we agree with the STAR*D researchers' speculation that quantified measurement enhanced outcome because incomplete response could not be ignored.

Changing the standard of care in the treatment of depression, or other psychiatric disorders, to incorporate a validated assessment tool would raise the standard to the level accepted in the treatment of other chronic medical disorders such as diabetes and hypertension. The use of a measurement tool should reduce the likelihood of under recognition of the residual symptoms which leaves patients at greater risk for relapse.

## Other reasons why measurement might improve outcome

Routine outcome assessment with self-report scales can enhance therapeutic effectiveness for different reasons depending on the stage of treatment. The completion of self-administered scales increases patients' active participation in their care, and this might facilitate participation in other therapeutically beneficial activities such as exercise or pleasant activities. Patients who are more active in their treatment, and who believe that their clinicians better understand their clinical status, may be more likely to continue with treatment. Valid symptom assessment may help clinicians identify for patients areas of improvement that had not been recognized. For example, consider a patient who is still depressed, pessimistic, amotivated, and self-deprecatory who, at the beginning of the follow-up visit states that they are no better, but who in fact is sleeping better, feeling somewhat more energetic, and concentrating better. Identification of some areas of improvement could reduce patients' therapeutic nihilism, thereby increasing treatment retention. Thus, more accurate symptom assessment might not only improve detection of residual symptoms in patients who report that they are feeling better, but it can also improve detection of mild improvement, which might be a harbinger of future improvement (Gelenberg and Chesen 2000, Nagayama et al. 1991, Szegedi et al. 2003), in patients who are not yet doing well.

Patients followed longitudinally, over the course of years, may uniquely benefit from routine use of scales. For example, it may be easier to detect seasonal patterns of symptom fluctuation when looking at graphs of symptom scores. Patients who relapse, and distort the effectiveness of treatment by minimizing or overlooking periods of sustained remission because of state-dependent cognitive biases, may be more open to more accurate views of their longitudinal course when shown the forms they had completed months earlier

which indicated minimal levels of symptoms.

There are thus multiple theoretical reasons as to why measurement-based care might improve treatment outcome.

## Prior Studies of the Impact of Measurement on Outcome

Lambert et al. (2001, 2002) have conducted a series of studies of the impact of measurement and feedback on psychotherapy outcome of mildly ill outpatients. Following the work of Howard et al. (1996), a patient's progress was compared to the expected course of symptomatic and functional improvement. The expected level of improvement was based on benchmarking studies of thousands of patients who received psychotherapy in diverse settings and who completed the same outcome measure. From these benchmarking studies "recovery curves" were derived which graphically illustrated the expected rate and level of improvement. Patients completed the outcome scale before each therapy visit, and a research assistant scored the scale and compared the results to the empirically derived recovery curves. Based on this comparison, a colored dot was placed on the patient's chart indicating the adequacy of improvement. Inadequate levels of improvement were accompanied by a message suggesting either that treatment should be intensified or perhaps changed altogether.

In their initial study, Lambert et al. (2001) randomly assigned 609 clients treated in a university counseling center to the feedback or treatment as usual groups. All patients completed the outcome measures. The group was relatively mild in severity with one-third receiving a V code diagnosis or a diagnosis of adjustment disorder. In the group making inadequate levels of improvement (approximately 10% of the entire sample), those randomized to the feedback condition received significantly more therapy visits than the patients randomized to the no feedback condition, scored significantly lower on the outcome questionnaire at the end of treatment, were more likely to improve by the end of treatment (26% vs. 16%), and less likely to deteriorate by the end of treatment (6% vs. 23%).

Lambert et al. (2002) conducted a replication study, again in a university counseling center treating mildly ill clients (more than 40% with a V code diagnosis or adjustment disorder). In this larger study of 1,020 clients, those in the feedback condition improved significantly more than those in the no feedback condition, though this difference was limited to clients who did not manifest the expected level of improvement. In the group that did not achieve expected levels of improvement during the course of treatment (approximately 24% of the sample), those in the feedback condition were significantly more likely to have improved by the time of treatment termination (32% vs 18%). The authors also combined these results with those from their first study and reported that the improvement rate across both studies in the patients failing to achieve expected improvement was significantly higher in the feedback group (30.5% vs. 17.5%), and deterioration rates were significantly lower (15% vs. 23%). As in the initial study, clients in the feedback condition received more therapy sessions. Other studies from this group have been consistent with these initial results (Harmon et al. 2007, Hawkins et al. 2004, Whipple et al. 2003).

Lambert's work has demonstrated that measurement and feedback is associated with improved outcome, and can influence a therapist's behavior insofar as more therapy visits are conducted with clients who are known to be not doing as well as expected. A limitation of these studies is that all except one small study of 200 psychiatric outpatients (Hawkins et al. 2004) have been based on mildly ill clients receiving psychotherapy at a university counseling center. Only one-quarter of the patients were diagnosed with some type of mood disorder (the exact nature of the disorder was not indicated). Also, these studies have ostensibly examined the impact of feedback, not measurement per se. Measurement in the absence of feedback is a sterile, clinically meaningless exercise that perhaps is counter-therapeutic. The subjects in Lambert et al's studies completed the outcome scale on a weekly basis, and one wonders what the clients in the no feedback group thought when their responses were not discussed in the treatment sessions. Perhaps some clients were frustrated, confused, or dissatisfied with treatment because the information provided on the outcome scale was not raised in treatment. In fact, the study by Hawkins et al. (2004) included a condition in which patients received explicit feedback based on their questionnaire responses, and they noted that patients were interested in this information.

Measurement-based care approaches need to use scales that are readily interpretable to the clinicians who use them. Lambert et al.'s study relied on research assistants to score the measure and alert clinicians to the results. This approach is cost prohibitive for implementation in clinical practice. As Lambert et al. (2002) themselves noted in the conclusion of their first replication study "if client-focused outcome research is to have any applicability it must remain simple and easy to implement in day-to-day clinical practice." Lambert's group has since developed easier, patient-entered and software-scored models where clinicians receive rating scores immediately during sessions.

## Potential obstacles in measuring outcome

There are five obstacles to be overcome in the implementation of outcomes assessment when treating psychiatric disorders in clinical practice: 1) Patient acceptability. If measurement is overly burdensome to patients, then they may be dissatisfied with their care and either drop out from treatment or seek care elsewhere. 2) Clinician acceptability. If measurement interferes with a clinician's usual work flow, then it is less likely to be adopted. 3) Clinical utility. Measurement that is clinically meaningful and improves the efficiency of conducting clinical evaluations will be more likely adopted than tools that do not inform clinical decision making. 4) Cost. Instruments that have higher acquisition costs, or support staff costs, are less likely to be utilized. 5) Data aggregation. If the goal of measuring outcome is to improve quality, then it will be necessary to aggregate information on outcome

across patients in order to determine the impact of a change in service delivery. Each of these obstacles is discussed below.

## Acceptability of outcome measurement

There is no shortage of measures to be used to monitor outcome. Two perspectives are of primary importance in deciding which measure to choose—that of the patient and the clinician. Patients should find the measures user-friendly and the directions easy to follow. The questions should be understandable and relevant to the patient's problem. The scales should be brief, taking no more than 2 to 3 minutes to complete, so that upon routine administration at follow-up visits patients are not inconvenienced by the need to come for their appointment 10-15 minutes early in order to complete the measure.

The instrument should provide clinicians with clinically useful information and improve the efficiency of conducting their clinical evaluation; thus, the measure should have practical value to the practicing clinician. Of course, clinicians need to be able to trust the information provided by any instrument they use. Consequently, outcome measures should have a sound basis in psychometrics, demonstrating good reliability, validity, and sensitivity to change. Clinicians and clinics should also find the instrument user-friendly; it should be easy to administer and score with minimal training. **Table 3** presents a list of desirable features of a typical depression outcome scale.

Our group examined the feasibility and acceptability of using a self-administered depression questionnaire to measure outcome in routine clinical practice in two studies of depressed psychiatric outpatients who were in ongoing treatment (Zimmerman and McGlinchey 2008a). The patients completed a questionnaire assessing how burdensome it was to complete the scale during the visit (0=very little burden; 3=a large burden), and their willingness to complete the scale at every visit to help monitor the progress of their treatment (0=not at all willing; 3=very willing to fill it out at every visit). Almost all patients considered questionnaire completion very little or only a little burdensome (98.0%, n=49), and no patient perceived it as very burdensome. More than 90% of patients indicated a willingness to complete the scale at every visit in the future if their clinician believed that it was helpful (94.0%, n=47).

Of course, acceptability and feasibility may vary by scale. Some measures consist of 100 or more statements, whereas others contain fewer than 10 items. The study summarized above was of a scale we developed in the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project—the Clinically Useful Depression Outcome Scale (CUDOS) (Zimmerman et al. 2008, Zimmerman et al. 2004), a brief measure that asks respondents to rate 16 symptom items on a 5 point Likert scale. In a second study of feasibility, a separate sample of 50 depressed outpatients completed both the CUDOS and the Beck Depression Inventory (BDI) (Beck et al. 1961) during a follow-up visit. In contrast to the CUDOS, the BDI has respondents read groups of 4 statements and select the item that best describes how they had been feeling during the preceding week. Thus, the BDI includes more information to read than the CUDOS. After completing the two questionnaires the patients completed a questionnaire asking which of the two measures took less time to complete, was easier to understand, less burdensome to complete, and more acceptable to complete at every follow-up appointment. Significantly more patients indicated that the CUDOS took less time to complete than the BDI, and was less of a burden to complete. Nearly three times as many patients indicated that they would prefer to complete the CUDOS than the BDI to monitor the outcome of treatment. These studies suggest that patients did not find scale completion burdensome, especially when the scale is brief, but they do appreciate scales being simpler and shorter in nature.

**Table 3.** *Desirable Features of a Self-Report Depression Outcome Scale*

| | |
|---|---|
| 1. | Brief |
| 2. | Acceptable to patients |
| 3. | Covers all DSM-IV diagnostic criteria for major depressive disorder |
| 4. | Reliable (internal consistency and test-retest reliability) |
| 5. | Convergent validity |
| 6. | Discriminant validity |
| 7. | Indicator of symptom severity |
| 8. | Indicator of remission status |
| 9. | Case-finding capability as a screening instrument |
| 10. | Assesses psychosocial function |
| 11. | Assesses quality of life |
| 12. | Assesses suicidal thoughts |
| 13. | Sensitive to change |
| 14. | Easy to score |
| 15. | Inexpensive |

## The data aggregation problem

Every year the U.S. News and World Report publishes a list of the top programs in psychiatry (and other medical specialties). Noteworthy in the accompanying article is the absence of a discussion of data describing treatment response rates. Are psychiatric patients treated in programs at the top of this list more likely to achieve better outcomes than patients treated in programs lower on the list, and do patients who are treated in programs that are not on the list have even poorer outcome? Perhaps in some medical specialties, for some procedures or for some disorders, data exist demonstrating that a positive outcome is more likely in some medical centers than others, but we are not aware of such data in psychiatry.

Considering the issue of outcome more broadly, our field is only beginning to ask fundamental questions regarding the effectiveness of our currently available treatments in real-world clinical practice. How well do they work? For whom do they work best, and for whom are they ineffective? How many patients are and are not receiving evidence-based care, and is the provision of evidence-based care associated with better outcome?

Are some clinicians more effective than others, and, assuming there are differences in patient outcomes between clinicians, is it possible to improve outcomes in patients treated by those clinicians who perform below average? Each of these, and other, questions can only be addressed if outcome is measured in clinical practice and the data is aggregated across patients.

While data aggregation is not necessary to realize the clinical benefit of measurement-based care in the treatment of individual patients, data aggregation is necessary for quality improvement efforts in which outcome across a caseload is compared before and after a change in service delivery. Data aggregation could be labor intensive, and thus costly. To address the potential obstacle posed by the cost of data entry, we developed a web-based system to administer and score the administration of the CUDOS and aggregate data across a clinician's caseload to determine response and remission rates at the end of the acute phase of treatment.

## Web-based assessment

A web-based platform for the administration of outcome assessments offers several advantages over paper-and-pencil assessments such as patient convenience, reduced missing data, reduced administrative burden and associated costs, automatic scoring, and generation of large data bases. Web-based scales can be completed by patients at their convenience in their home rather than arriving early or staying after their clinical appointment to complete the measure. A computer administered survey can prompt respondents to ensure all questions are answered thereby reducing missing data. The administrative costs associated with the copying, handing out, and scoring of paper questionnaires are reduced with a web-based system. Similarly, the high costs of establishing and maintaining a data base to evaluate treatment outcome for a large sample of patients based on administration, scoring, and data entry of paper questionnaires could be markedly reduced with a web-based system. Moreover, because the data collected via the Internet is automatically entered into a data base, data entry errors are reduced (Coles et al. 2007).

Studies comparing paper and computer or Internet scale administration have found high correspondence between the two assessment methods (Cook et al. 2007, Merten and Ruch 1996, Ogles et al. 1998, Peterson et al. 1996). However, few studies examined web-based scale administration, and we are unaware of a single study of psychiatric patients in real-world ongoing treatment. Thus, caution should remain before extrapolating the results of paper-and-pencil scale administration to web-based administration, and the prevailing recommendation of demonstrating equivalence should continue. As part of the MIDAS project we have conducted the first study comparing paper and Internet administration of a depression scale to depressed psychiatric outpatients in ongoing treatment.

We present here the preliminary findings of an ongoing study of the comparability of paper and web versions of the CUDOS. Forty-eight hours before their appointment, 30 depressed patients in ongoing outpatient treatment received an email reminding them of their forthcoming appointment and directed them to complete the web-administered version of the CUDOS (CUDOS-W). At the end of their appointment the patients were asked if they would complete the paper version of the CUDOS. It was explained to them that the circumstances and setting of scale completion sometimes influences responses to a scale, therefore, it was important to examine the comparability of computer and paper administrations of a scale. At the visit, the clinician completed the Montgomery-Asberg Depression Rating Scale (MADRS) (Montgomery and Asberg 1979) and rated the Global Assessment of Functioning (GAF) and Clinical Global Index of Severity (CGI-S) (Guy 1976).

The formats of the paper and Internet versions of the CUDOS were identical. A copy of the scale is included in the Appendix. The web administration presents the entire scale at once, rather than one item at a time. Patients are able to change their responses after answering a question. The questionnaire cannot be submitted unless all items are completed.

The sample included 6 (20.0%) men and 24 (80.0%) women who ranged in age from 22 to 85 years (M = 47.0, SD =12.2). The mean score on the MADRS (11.9, SD=12.3) and CGI-S (1.2, SD=1.3) indicated a mild level of depression severity. The mean score on the GAF was 66.1 (SD=10.2).

The average interval between the completion of the paper and Internet versions of the scale was 1.1 days (SD=0.8). The correlation between the CUDOS and CUDOS-W was high (ICC=.95, $p<.001$). The mean scores were similar on the paper and Internet administrations ($18.9 \pm 14.5$ vs. $18.5 \pm 13.8$, paired $t = 0.4$, n.s.). In our previous validation studies of the paper version of the CUDOS, we found that a cutoff of 20 identified patients who were in remission. Based on this cutoff score there was 94.7% agreement between the paper and Internet administrations in determining patients' remission status ($k = .90$).

The internal consistency of the paper and Internet administrations of the CUDOS was high (Cronbach's alpha = .94 and .93, respectively). For each item the correlation between the CUDOS and CUDOS-W was significant (median = .86).

Both the paper and Internet versions of the CUDOS were significantly correlated with the MADRS (r= .92 and .89, respectively), CGI-S (r= .94 and .93, respectively), and GAF (r= .94 and .90, respectively). None of the differences in the correlations between the paper and Internet administrations and the validity scales were significant.

The preliminary results of this ongoing study support the reliability and validity of Internet administration of the CUDOS. Internal consistency, item-scale correlations, and correlations with external validators were as high with Internet administration as with paper administration of the scale. The website, which includes both the CUDOS and our Clinically Useful Anxiety Outcome Scale (Zimmerman et al. 2010), can be found at www.outcometracker.org and is currently available for clinicians to use at no cost.

## Are you using scales to monitor outcome when treating depressed patients?

We hope that the reader will ask him/herself the question heading this section. If you are not using a scale to monitor outcome, and the chances are very high that you are not, ask yourself why not. Consumer-friendly, reliable and valid self-administered questionnaire can improve the efficiency of the clinical encounter, and allow clinicians to spend more time discussing topics other than symptoms. In this era when many clinical encounters are 15-minute (or briefer) medication visits, increased efficiency can make the visit more meaningful and beneficial to both clinicians and patients. Of course, there is also a risk that use of self-administered scales completed before the visit could be responsible for the clinician reducing the amount of time spent with the patient and not conducting an adequate clinical assessment of the patients' status. Self-administered questionnaires are not a substitute for clinical assessment, but instead could enhance the efficiency of such evaluations.

There are many self-administered depression scales, though some are less appealing as outcome tools for use in routine clinical practice because they are either too long (Beck et al. 1961, Rush et al. 1996, Zimmerman et al. 1986), lack adequate coverage of the DSM-IV diagnostic criteria (Radloff 1977, Zung 1965), are expensive to purchase (Beck et al. 1961), or are somewhat complicated to score (Zung 1965). Because of ease of use considerations, we would recommend that either the CUDOS or the 9-item Patient Health Questionnaire (PHQ-9) (Kroenke et al. 2001) be used by clinicians at every visit to monitor the course of depression. Both scales take less than 2 minutes, on average, to complete, and both assess all of the DSM-IV criteria for major depressive disorder. Because it contains fewer items than the 16-item CUDOS, the PHQ-9 probably takes a little less time to complete. However, the advantage offered by being somewhat briefer is offset by some loss of information. The PHQ-9 adheres to the construction of the DSM-IV criteria; thus compound DSM-IV criteria which refer to more than one symptom (e.g., insomnia or hypersomnia; increased or decreased appetite) are represented by a single item on PHQ-9. Since treatment decision-making might be influenced by whether a patient has insomnia or is sleeping too much, or has a reduced appetite or is eating too much, the PHQ-9 does not capture potentially clinically significant information. However, more important than which scale is used to monitor outcome is that some measure is used. Measures such as the CUDOS or PHQ-9 have clearly identified cutoff scores to identify remission, and therefore should not require any special training to be adopted by non-mental health professionals.

## Conclusions

It is time for the clinical management of psychiatric disorders such as depression to more closely resemble the management of other chronic medical conditions, and this means that outcome should be measured in a quantifiable manner at each clinical encounter. There is suggestive evidence that measurement-based care improves outcome, though this has not been studied using a method that can be incorporated into routine clinical care. If measurement-based care is to be adopted in clinical practice it is essential that it not be burdensome to the practicing clinician. Brief, yet valid, scales exist that can be readily incorporated into clinical practice. Routine assessment is well received by patients (Zimmerman and McGlinchey 2008a). If the results of well-designed, randomized controlled studies demonstrate that measurement-based care improves outcome, improves treatment retention, and reduces more costly and intensive levels of service, this could potentially have a profound impact on the treatment of depression in clinical practice because of how easy it will be for clinicians to adopt this care management approach. There may be only limited data suggesting that measurement might improve outcome when treating depression, but there is no reason to wait until the studies have been done to prove the benefit of measurement-based care in the treatment of depression. There is little downside to adopting this approach when treating depressed patients, unless clinicians shorten the therapeutic encounter because they overly rely on information from self-report questionnaires and do not conduct adequate clinical assessments.

To highlight an overlying model that runs throughout this special issue, psychopharmacologists must have a better command of the data (approvals, off-label approaches, pharmacodynamic mechanisms, etc.) and now push their practice envelopes to rationally use outcome measures. In the era of pay for performance initiatives and for psychiatry to be accepted and appreciated as a medical specialty, outcomes should be measured while we strive for remission in our patients. These outcome measures become even more important as clinicians use rational polypharmacy techniques where drug costs and side effect burdens likely increase. Clinicians and patients may now assign outcome numbers to the intervention employed. For example, both clinician and patient will know that the patient's chronic, resistant depression improved 24% after a third drug augmentation strategy was added which also prompted dry mouth and cost a $40 a month co-pay. Both parties may now assess if the added cost and side effect was clinically meaningful.

## References

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961). An inventory for measuring depression. *Archives of General Psychiatry* 4, 561-571.

Coles ME, Cook LM, Blake TR (2007). Assessing obsessive compulsive symptoms and cognitions on the internet: evidence for the comparability of paper and Internet administration. *Behaviour Research and Therapy* 45, 9, 2232-2240.

Cook IA, Balasubramani GK, Eng H, Friedman E, Young EA, Martin J, Nay WT, Ritz L, Rush AJ, Stegman D, Warden D, Trivedi MH, Wisniewski SR (2007). Electronic source materials in clinical research: acceptability and validity of symptom self-rating in major depressive disorder. *Journal of Psychiatric Research* 41, 9, 737-743.

Evans M, Hollon S, DeRubeis R, Piasecki J, Grove W, Garvey M, Tuason V (1992). Differential relapse following cognitive therapy and pharmacotherapy for depression.

*Archives of General Psychiatry* 49, 10, 802-808.
Gelenberg AJ, Chesen CL (2000). How fast are antidepressants? *Journal of Clinical Psychiatry* 61, 10, 712-721.
Gilbody S, House A, Sheldon T (2002). Psychiatrists in the UK do not use outcomes measures. *British Journal of Psychiatry* 180, 101-103.
Guy W, (1976). *ECDEU Assessment Manual for Psychopharmacology - Revised*. U.S. Department of Health and Human Services, Rockville, pp. 534-537.
Hamilton M (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry* 23, 56-62.
Harmon S, Lambert M, Smart D, Hawkins E, Nielsen S, Slade K, Lutz W (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research* 17, 4, 379-392.
Hawkins E, Lambert MJ, Vermeersch D, Slade K, Tuttle K (2004). The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy Research* 14, 3, 308-327.
Howard KI, Moras K, Brill PL, Martinovich Z, Lutz W (1996). Evaluation of psychotherapy.Efficacy, effectiveness, and patient progress. *American Psychologist* 51, 10 1059-1064.
Judd L, Paulus M, Schettler P, Akiskal H, Endicott J, Leon A, Maser J, Mueller T, Solomon D, Keller M (2000). Does incomplete recovery from first lifetime major depressive episode herald a chronic course of illness? *American Journal of Psychiatry* 157, 1501-1504.
Kroenke K, Spitzer R, Williams J (2001). The PHQ-9. Validity of a brief depression severity measure. *Journal of General Internal Medicine* 16, 9, 606-613.
Lambert MJ, Whipple JL, Smart DW, Vermeersch DA, Nielsen SL, Hawkins EJ (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research* 11, 1, 49-68.
Lambert MJ, Whipple JL, Vermeersch DA, Smart DW, Hawkins EJ, Nielsen SL, Goates M (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy* 9, 2, 91-103.
Merten T, Ruch W (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences* 20, 3, 281-291
Montgomery SA, Asberg M (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry* 134, 382-389.
Nagayama H, Nagano K, Ikezaki A, Tashiro T (1991). Prediction of efficacy of antidepressant by 1-week test therapy in depression. *Journal of Affective Disorders* 23, 4, 213-216.
Ogles BM, France CR, Lunnen KM, Bell MT, Goldfarb M (1998). Computerized depression screening and awareness. *Community Mental Health Journal* 34, 1, 27-38.
Paykel ES, Ramana R, Cooper Z, Hayhurst H, Kerr J, Barocka A (1995). Residual symptoms after partial remission: An important outcome in depression. *Psychological Medicine* 25, 6, 1171-1180.
Peterson L, Johannsson V, Carlsson S (1996). Computerized testing in a hospital setting: Psychometric and psychological effects. *Computers in Human Behavior* 12, 3, 339-350.
Quitkin F, McGrath P, Stewart J, Deliyannides D, Taylor B, Davies C, Klein D (2005). Remission rates with 3 consecutive antidepressant trials: Effectiveness for depressed outpatients. *Journal of Clinical Psychiatry* 66, 6, 670-676.
Radloff LS (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1, 3, 385-401.
Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH (1996). The Inventory of Depressive Symptomatology (IDS). *Psychological Medicine* 26, 3, 477-486.
Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, Niederehe G, Thase ME, Lavori PW, Lebowitz BD, McGrath PJ, Rosenbaum JF, Sackeim HA, Kupfer DJ, Luther J, Fava M (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *American Journal of Psychiatry* 163, 1905-1917.
Simons A, Murphy G, Levine J, Wetzel R (1986). Cognitive therapy and pharmacotherapy for depression. Sustained improvement over one year. *Archives of General Psychiatry* 43, 1, 43-48.
Szegedi A, Muller MJ, Anghelescu I, Klawe C, Kohnen R, Benkert O (2003). Early improvement under mirtazapine and paroxetine predicts later stable response and remission with high sensitivity in patients with major depression. *Journal of Clinical Psychiatry* 64, 4, 413-420.
Thase ME, Simons AD, McGeary J, Cahalane JF, Hughes C, Harden T, Friedman E (1992). Relapse after cognitive behavior therapy of depression: Potential implications for longer courses of treatment. *American Journal of Psychiatry* 149, 1046-1052.
Trivedi M, Rush A, Wisniewski S, Nierenberg A, Warden D, Ritz L, Norquist G, Howland R, Lebowitz B, McGrath P, Shores-Wilson K, Biggs M, Balasubramani G, Fava M (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D implications for clinical practice. *American Journal of Psychiatry* 163, 28-40.
Whipple JL, Lambert MJ, Vermeersch DA, Smart DW, Neilsen SL, Hawkins EJ (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology* 50, 1, 59-68.
Zimmerman M, Chelminski I, McGlinchey JB, Posternak MA (2008). A clinically useful depression outcome scale. *Comprehensive Psychiatry* 49, 2, 131-140.
Zimmerman M, Chelminski I, Young D, Dalrymple KL . A clinically useful anxiety outcome scale. *Journal of Clinical Psychiatry*. 71, 5, 534-42.
Zimmerman M, Coryell W, Corenthal C, Wilson S (1986). A self-report scale to diagnose major depression disorder. *Archives of General Psychiatry* 43, 11, 1076-1081.
Zimmerman M, Mattia JI (2001). A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire (PDSQ). *Archives of General Psychiatry* 58, 8, 787-794.
Zimmerman M, McGlinchey J (2008a). Depressed patients' acceptability of the use of self-administered scales to measure outcome in clinical practice. *Annals of Clinical Psychiatry* 20, 3, 125-129.
Zimmerman M, McGlinchey JB (2008b). Why don't psychiatrists use scales to measure outcome when treating depressed patients? *Journal of Clinical Psychiatry* 69, 12,1916-1919.
Zimmerman M, Posternak M, Chelminski I (2004). Using a self-report depression scale to identify remission in depressed outpatients. *American Journal of Psychiatry* 161, 1911-1913.
Zung WWK (1965). A self-rating depression scale. *Archives of General Psychiatry* 12, 63-70.

# Appendix

**INSTRUCTIONS**

This questionnaire includes questions about symptoms of depression. For each item please indicate how well it describes you during the PAST WEEK, INCLUDING TODAY. Circle the number in the columns next to the item that best describes you.

**RATING GUIDELINES**

**0=not at all true (0 days)**
**1=rarely true (1-2 days)**
**2=sometimes true (3-4 days)**
**3=often true (5-6 days)**
**4=almost always true (every day)**

**During the PAST WEEK, INCLUDING TODAY....**

1. I felt sad or depressed 0 1 2 3 4
2. I was not as interested in my usual activities 0 1 2 3 4
3. My appetite was poor and I didn't feel like eating 0 1 2 3 4
4. My appetite was much greater than usual 0 1 2 3 4
5. I had difficulty sleeping 0 1 2 3 4
6. I was sleeping too much 0 1 2 3 4
7. I felt very fidgety, making it difficult to sit still 0 1 2 3 4
8. I felt physically slowed down, like my body was stuck in mud 0 1 2 3 4
9. My energy level was low 0 1 2 3 4
10. I felt guilty 0 1 2 3 4
11. I thought I was a failure 0 1 2 3 4
12. I had problems concentrating 0 1 2 3 4
13. I had more difficulties making decisions than usual 0 1 2 3 4
14. I wished I was dead 0 1 2 3 4
15. I thought about killing myself 0 1 2 3 4
16. I thought that the future looked hopeless 0 1 2 3 4

17. Overall, how much have symptoms of depression interfered with or caused difficulties in your life during the past week?

0) not at all
1) a little bit
2) a moderate amount
3) quite a bit
4) extremely

18. How would you rate your overall quality of life during the past week?

0) very good, my life could hardly be better
1) pretty good, most things are going well
2) the good and bad parts are about equal
3) pretty bad, most things are going poorly
4) very bad, my life could hardly be worse