

Outcome assessment with psychiatric inpatients diagnosed as Severely and Persistently Mentally Ill: A guide for instrument selection

Gary M. Burlingame and Todd W. Dunn
Brigham Young University
Steven Chen and August Lehman
Utah State Hospital
Reid Axman
Brigham Young University
Dallas Earnshaw
Frank Rees
Utah State Hospital

Abstract

Objective: A procedure to assist in selecting outcome measures for inpatients treated at a state psychiatric hospital is described. It combines evidence-based criteria from the literature, instruments shown to be sensitive to change in clinical trials, and the perspectives of a multi-disciplinary team of researchers, administrators, providers and patient advocates.

Methods: Recent efficacy and effectiveness studies were used to identify recurrently used outcome instruments. Next, comparisons of the most frequently used instruments are made on seven criteria proposed as “best practice” indicators, including sensitivity to change and robust psychometrics.

Results and Conclusions: Rater-completed instruments are represented more often than patient-completed; however considerable variability across both methods was found on the criteria. The limited resources associated with publicly-funded inpatient facilities led to a recommendation to select at least one rater- and one patient-completed instrument.

Introduction

In 2000, national health care expenditures were approximately 1.3 trillion dollars, with a projected two-fold increase by 2010 (1). Over 7% were for mental health care totaling an estimated 91 billion dollars (2). To curb this alarming cost increase and insure quality care, a greater emphasis on treatment accountability emerged from legislative and accreditation bodies, public agencies, and consumers (3,4,5). In this “age of accountability”, it is becoming standard practice for mental health providers to implement outcome management programs to understand the relationship between services, cost, and patient change (6,7). Such programs require instruments that are standardized, psychometrically sound, easy to use, practical, and available at a low cost (6,8,9,10).

Those interested in employing outcome management are met with a cornucopia of possibilities. As Hermann and colleagues note (11,12), more than 50 stakeholders have proposed over 300 measures for quality assessment leading some to recommend common measures or methods (13,14,15,16). A vital distinction to maintain as one enters this literature is Donabedian’s tripartite framework (17) that categorizes quality assessment into measures of structure, process or outcome.

Measures that assess the process of health care delivery are plentiful and have received more attention than outcome measures (11). This has been attributed to the lower cost of process measures and their ability to provide quick feedback to administrators and clinicians (18,19). However, measures of outcome may be a more direct indicator of quality than either structure or process (20). Moreover, recent advances in computerized outcome management systems now provide the same real-time feedback that was once the domain of process measures (21,22,23). Ellwood (24) opines that outcome measures empower psychiatrists' management of patient care, while others suggest that:

Psychiatrists and mental health care administrators who use outcome assessment to study and apply principles of continuous quality management daily, will probably experience better efficiency, greater effectiveness, lower costs and more satisfied patients (18) (p. 489).

The less frequent attention to outcome assessment may also be due to confusion around definitions (25) coupled with an enormous number of measures from which to select (26). One promising method to deal with the bewildering number of outcome measures is to restrict focus to those that are designed for the target patient population. Unfortunately, clinical reality often demands that the target population be defined beyond a single diagnosis or grouping; i.e. depression or mood disorder. For instance, Erbes and colleagues (27) evaluate and recommend outcome instruments for a Veterans Administration hospital necessarily considering a broad range of diagnoses. In this paper we focus upon psychiatric inpatients diagnosed as severely and persistently mentally ill and propose a similar method.

Outcome Assessment with Psychiatric Inpatients

Individuals diagnosed as severely and persistently mentally ill have been considered a difficult population to track from an outcomes perspective (28,29). The terms severe and persistent have been operationalized as “functional limitations in activities for daily living, social interaction, concentration, and adaptation to change in the environment” and likelihood to “last for twelve months or more”, respectively (30). These patients number between one and five million, are diagnosed with schizophrenia, schizo-affective, bipolar, major depression, autism, and obsessive compulsive disorders and have been estimated to cost the health care system \$148 billion annually (31).

Despite the extensive impact on resources and the increasing focus on accountability, active debate persists regarding which outcome measures to use with this patient population. Several studies have evaluated the utility and effectiveness of individual measures (e.g., 32,33,34,35,36) while others have focused upon comparisons between a limited number of instruments (e.g., 37,38,39,40,41). Still, little consensus exists on which measures to select. Particular challenges include norms that provide interpretive meaning, sensitivity to change in patients that are expected to demonstrate little improvement, and robust psychometrics needed to capture subtle patient change.

If an outcome measure succeeds in addressing these challenges, it must also meet a reasonable standard of clinical utility by minimizing time devoted to data collection and other direct costs. While the challenges associated with selecting an outcome measure is the focus herein, the pragmatics of implementing an outcome management program in a large state-run psychiatric hospital are addressed in the State Mental Health Policy column of this issue (42).

Overview of Framework for Selecting Outcome Measures

While a host of resources for selecting quality measures are available (6,11,14,18,43,44), their direct application pose complex challenges. Accordingly, we propose and illustrate an approach used to guide a multi-disciplinary work group at a state psychiatric hospital in its selection of outcome measures (Figure 1). We don't presume to evaluate the enormous number of outcome measures (26), nor are we recommending a specific set of measures. Rather, we describe a method that proved fruitful in wading through a myriad of recommendations and measures.

Figure 1 begins by identifying the targeted population. This is critical when evaluating instruments because many have not been extended or tested with distinct patient populations (e.g., norms, construct validity, sensitivity to change, etc). Invariably, the number of instruments is reduced at step one. Step two (Figure 1) further limits the universe of instruments by considering measures that are repeatedly used in efficacy or effectiveness studies with the targeted population. There are obvious advantages and disadvantages with this step. Advantages include the increased probability of selecting measures that will successfully capture change in a targeted population. Indeed, average change across studies on a measure can be quantified by using a pre-to-post treatment effect size. This metric provides one index to compare sensitivity to change which is useful since instruments vary on sensitivity to change (44), especially when used with different populations (45). An average effect size also provides clinicians with a baseline to "benchmark" expectations for patient gains.

There are notable disadvantages associated with the empirical filter in step two. If widely employed, it could lead to stagnation in the field by disadvantaging the use of promising new instruments. However, "older" measures have presumably survived the test of time because they are effective and often serve as the "gold standards" for new instruments. This

approach may also underemphasize certain domains (e.g., functioning) and overemphasize others (e.g., symptom). However, our proposal is not intended to balance all potential outcome domains and we defer to sources on such (18,46). Rather, our goal is pragmatic; identifying measures successfully used with a target population to capture meaningful patient change.

Step three acknowledges the numerous and competing criteria proffered as selection guidelines (e.g., broad domain coverage, robust psychometrics, cost and clinical utility). Indeed, in the illustration that follows we identified 24 criteria offered by experts. Once again, the pragmatics of clinical practice require a restricted set that are highly relevant to the clinical setting. A frequently endorsed measure in step two may be simply impractical for a particular clinical setting. Thus, step 3 criteria temper decision making by considering the clinical setting. Step 4 applies these criteria to the measures uncovered in step two and often reorders which measures are considered best. A measure identified in step 2 as highly endorsed may drop precipitously in rank after considering the criteria in step 3. The integration (step 4) of empirical performance (step 2) and clinical setting (step 3) is the end result of selection (step 5).

Representing multiple perspectives may produce better decisions regarding outcome systems (6,27). Accordingly, a group of academically based researchers, hospital administrators, mental health care providers and patient advocates teamed together to use the aforementioned method to select optimal measures for a state hospital. More specifically, our principle aims were to: a) survey treatment studies to identify measures in frequent use for our target population, b) identify relevant literature-based selection criteria and c) review the outcome measures using the proposed criteria.

Applying the Framework to a State Hospital: An Illustration

Step One: Identifying Target Population

Patients at our facility are typically diagnosed (47) with psychotic illnesses (54% schizophrenia, delusional or schizoaffective disorders) and mood disorders (23% major depression, bipolar or dysthymia). Thus, outcome measures not directly normed for this population were excluded; a procedure that biased our sample in favor of the target population.

Step Two: Survey of Relevant Literature

Our interest in calibrating our selection process with measures used in efficacy or effectiveness treatment studies led to a computerized search of over 30 bibliographic databases (e.g., PsychINFO, MEDLINE, Social SciSearch, ERIC). This yielded nearly 500 citations published between 1990 and 2002 using the search terms: a) severe and persistent and mental and ill or SPMI, b) severe and mental and ill or SMI, and c) schizophrenia and outcome and inpatient. Studies before 1990 were excluded to limit bias towards older instruments, and instruments were not counted more than once if they were used in multiple publications resulting from a single investigation.

Only 94 citations (20%) were treatment evaluation studies that used standardized outcome measures (Table 1). Excluded citations included conceptual or policy papers, reviews, studies of process or structure measures, or unstandardized outcomes measures (e.g., drop out rates, cost, recidivism, etc.). The sample produced 110 measures, with 11 (10%) used four or more times, 3 (3%) used thrice, 15 (14%) used twice, 81 (74%) used just once. Interestingly, 25 (23%) were investigator-created which have been argued as providing meaningless comparative information (48).

Three observations were drawn from the results of the survey (Table 1). First, the clinician-rated measures used most often included: a) the Brief Psychiatric Rating Scale (BPRS) (49), b) the Global Assessment of Functioning Scale (GAF) (50), c) the Positive and Negative Syndrome Scale (PANSS) (51), and d) the Scale for the Assessment of Negative Symptoms (SANS) (52). The BPRS was employed twice as frequently (44.1%) as the PANSS and SANS.

Second, very few self-report instruments found repeated use. Only two self report outcome measures surfaced in four or more studies, including the Symptom Checklist-90-Revised (SCL-90-R) (53) and the Quality of Life Interview (QOLI) (54). Third, multiple measure assessment was preferred over single measure protocols by a ratio of 2 to 1. This finding mirrors recommendations that a single source may be less reliable because each source contributes a valid yet potentially divergent perspective (55). While the average number of measures used per study was 2.8 (range from 1 to 13), what was striking was the number of studies ($N = 30$) that used a single outcome instrument.

Step 3: Selection of Criteria

The absence of a “gold standard” for selecting outcome instruments led to an integration of criteria suggested by six sources (6,9,10,55,56,57). These experts offer 24 criteria for selecting optimal outcome measures, from which seven were chosen based on frequency of endorsement and fit with our setting. In no particular hierarchical order, the criteria were: a) applicability to the target population, b) availability of training protocol and materials, c) appropriate norms to insure interpretability of scores, d) psychometric integrity (i.e., possess adequate reliability and validity), e) cost, f) administration time, and g) sensitivity to change. Each is discussed more fully below.

Step 4: Comparison of Frequently Used Outcome Instruments

A summary of our evaluation on six of the most frequently used clinician and self-report instruments is presented in Table 2. (This summary is restricted to six instruments due to space limitations.) A brief examination of each measure and greater explication of the seven criteria follows.

Brief Psychiatric Rating Scale (BPRS). The BPRS satisfied our target population since it was created to provide rapid assessments of psychopathology for inpatient populations. Its' extensive use in the literature has produced ready-made norms for a variety of populations. There are two revisions of the original 16-item version (49); an 18-item (58) and 24-item expanded version (BPRS-E) (59). Each version has produced four similar symptom factors; manic hostility, withdrawal-retardation (negative symptoms), thinking disturbance (positive symptoms), and depression-anxiety that match typical patient characteristics of state psychiatric hospitals (47).

Clinician-rated scales can provide greater consistency across patients and diagnoses than self-report measures thereby producing more reliable system wide evaluations (60). However, this consistency is directly related to the quality of the training material available to insure adequate inter-rater reliability; a clear strength of the BPRS (44,57). Indeed, good to moderate inter-rater reliability is evident (37,61,62), along with moderate test-retest reliability (61) and good internal consistency reliability (62). The literature was also largely supportive of the it's construct and concurrent validity (61,63,64,65,66,67). We ranked the clinical utility of the BPRS as high since it was normed on normal and clinical populations, available at no cost, and very sensitive to change with an average $d = 1.21$. It's greatest shortcoming was the resource drain associated with a clinician-

rated instrument, an issue addressed by Earnshaw et al. in this issue (42).

Global Assessment of Functioning Scale (GAF). As a standard part of the diagnostic protocol (68), the GAF is the most widely used measure of psychiatric patient function (33) with the extant literature providing a wealth of normative data. Introduced as a revised version of the Global Assessment Scale (69), clinicians rate global patient function on a single scale ranging from 100 (e.g., absence of symptoms to minimal symptoms) to 1 (e.g., persistent danger of severely hurting self or others). Research has reported inter-rater reliability coefficients that range from modest to excellent (70,71,72,73) and moderate to high concurrent validity estimates (73,74). From a clinical utility perspective, the GAF was viewed as comparable to the BPRS being normed on inpatient and outpatient populations, available at no cost, very quick to administer, and very sensitive to change ($d = 1.10$). Like the BPRS, consistency of ratings requires the implementation of rater training and periodic consistency checks.

Positive and Negative Syndrome Scale (PANSS). The PANSS was developed "as an instrument for measuring the prevalence of positive and negative syndromes in schizophrenia" (p. 270) (51). It consisted of the BPRS-18 (58) plus 12 items from the Psychopathology Rating Scale (75). Clinicians rate patient symptomatology using 30-items that aggregate on four scales (positive symptoms, negative symptoms, composite, and general psychopathology). Research has reported evidence of acceptable construct and concurrent validity (76,77), good internal consistency reliability, moderate test-retest reliability (51), and inter-rater reliability coefficients that range from high to moderate (37,77). Clinical utility was rated lower since it is lengthier to administer, more costly to use (\$32 for a set of 25 questionnaires) and normed on a narrower population. Nevertheless it

appears to be very sensitive to change in our analysis ($d = 1.23$).

Scale for the Assessment of Negative Symptoms (SANS). The SANS was developed by Andreasen (52,78) as a measure of negative symptoms in patients with schizophrenia. Clinicians use 30-items that are aggregated on five subscales (affective flattening/blunting, alogia, apathy, asociality, and inattention). It possesses adequate construct and concurrent validity coefficients (35,77), good internal consistency reliability (52), moderate 24-month test-retest reliability (79), and interrater reliability coefficients that range from moderate to high (52,66,79). It was ranked the lowest because of moderate clinical utility, as it was normed on a single population, somewhat lengthy and moderately sensitive to change ($d = .68$). However, it is available at no cost.

Symptom Checklist-90-R (SCL-90-R). Originally designed for use with psychiatric outpatients, the SCL-90-R (53) has enjoyed wide spread use in clinical and research settings producing a wealth of normative data. Patients respond to 90-items that are aggregated on nine symptom dimensions (somatization, obsessive-compulsivity, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism) and three global scales (Global Severity Index, Positive Symptom Distress Index, and Positive Symptom Total). Research has reported little evidence of construct validity (53,80), although it shows good internal consistency, test-retest reliability (80,81,82), and moderate concurrent validity with the BPRS (83).

The SCL-90-R was viewed as having moderate clinical utility, being normed on community, outpatient and inpatient populations, quick to administer and moderately sensitive to change ($d = .69$). Considerations lowering its rank were cost (\$41 per 50 hand-scored answer sheets) and the nature of self-report with the target

population. Self-report measures require less staff time, they also permit consumer-focused outcome assessment, as patients are empowered to report on their symptoms and expectations about treatment (84). Disadvantages include an insufficient clinical picture due to the dependence on the patient's ability to accurately describe their condition, which at times is doubtful due to denial, minimization of symptomatology, or responder bias (85).

Lehman's Quality of Life Interview (QOLI). The QOLI (51) is a highly structured interview developed to assess current quality of life and global well being among chronically mentally ill populations. It is made up of objective and subjective questions that allow the patient to rate their current situation and satisfaction with life. It possesses good construct validity (86,87), moderate concurrent validity (88), moderate test-retest reliability (51), and high internal consistency reliability (89). It was ranked low on clinical utility due to its length, cost (pay-per-use) and low sensitivity to change ($d = .02$). However, the latter was based on a single study leading to considerable caution, and norms were available for community and inpatient populations.

Step 5: Selecting Measures

Clinician-rated instruments clearly outnumbered self-report instruments in our analysis. Lachar and colleagues (64) explain that such measures have recently achieved an advantage over self-report in hospitals because of the disabling psychopathology patients now must exhibit to justify hospitalization. The impairment of newly admitted patients negatively impacts patients' ability to complete even a brief self-report measure. However, the accuracy of information from clinician-completed measures must be balanced by the resource drain. The BPRS and GAF require the least time to administer, and the BPRS, GAF, and PANSS appear to be equally sensitive to change, yet all required mastery of training

materials and demonstrated reliability to produce meaningful outcome information.

The team openly acknowledged the limitations of the sample, including a time frame that may have disadvantaged newer instruments (e.g., Multnomah Community Abilities Scales and Outcome Questionnaire). Further, while the frequency count allowed us to easily calibrate against findings in the extant literature, it may have inadvertently excluded potentially useful instruments because of their infrequent use in our sample (e.g., the Medical Outcomes Study SF-36 and Addictions Severity Index appeared in two studies each). However, infrequent use portend of unknown properties such as sensitivity to change and normative characteristics.

Three issues affected our final recommendations. First, global single scale assessments (e.g., GAF) are frequently used because they are simple to administer and provide immediate feedback (70). However, these scales suffer limitations in accuracy due to the combining of patient symptomatology and function in a single rating (90) leading some to question their accuracy with our target population (33,91).

Second, like most publicly funded facilities, we have limited resources. As a mental health agency focused on improving service delivery from both an organizational and consumer-oriented perspective (84), we were aware of the considerable discussion regarding the importance and effectiveness of self-report and clinician-rated instruments (e.g., 92,93,94,95). At face value, our survey suggests the BPRS and SCL-90-R as the best clinician-rated and self-report outcome instruments. However, concerns over financial resources, administration time, staff support, staff competence, and training led to active debate. Our adoption of the BPRS led to infrastructure realignment to address these concerns, which was detailed by Earnshaw and colleagues in this issue (42).

Finally, our survey suggested the SCL 90-R as a self-report tool, but its use raised two concerns; meaningfulness of the outcome data due to patient impairment and cost. When patients are physically unable or unwilling (due to malingering or resistance) to complete a self-report assessment, data may be too erratic (item endorsement at both ends of the range) to facilitate meaningful interpretation. Nevertheless, we adopted an alternative self-report measure due to cost and Earnshaw et al. (42) details how we dealt with data accuracy concerns.

Conclusion

Publicly funded facilities have clinical services as the primary focus with sparse resources available to allocate for outcome assessment. Nonetheless, all of us are faced with evidence-based accountability requirements to demonstrate the effectiveness of clinical services. Outcome measures may provide a valuable supplement to other measures of quality (structure & process) that are more frequently employed. The method employed herein to evaluate extant outcome measures for a target population provides one guide for instrument selection that may prove useful with other target populations. Leveraging against existing clinical trial literature focuses discussion upon a limited set of instruments with an estimated sensitivity to change and positively biases discussion upon domains and measures that have a proven empirical track record. The proposed method is not without limitations, foremost of which is its empirical versus clinical bias.

References

1. U.S. Census Bureau: Statistical Abstract of the United States. Retrieved July 18, 2003, from <http://www.census.gov/prod/2003pubs/02statab/health.pdf>
2. Coffey RM, Mark T, King E, et al: National Estimates for Expenditures

- for Mental Health the Substance Abuse Treatment, 1997. Rockville, MD, Substance Abuse and Mental Health Services Administration, 2000
3. Lyons JS, Howard KI, O'Mahoney MT, et al: *The Measurement and Management of Clinical Outcomes in Mental Health*. New York, John Wiley & Sons, 1997
 4. Mirin S, Namerow M: Why study treatment outcome? *Hospital and Community Psychiatry* 42:1007-1013, 1991
 5. Teague GB, Ganju V, Hornik JA, et al: The MHSIP mental health report card: consumer-oriented approach to monitoring the quality of mental health plans. *Evaluation Review* 21:330-341, 1997
 6. Burlingame GM, Lambert MJ, Reisinger CW, et al: Pragmatics of tracking mental health outcomes in a managed care setting. *Journal of Mental Health Administration* 22:226-236, 1995
 7. Burlingame GM, Mosier JI, Wells MG, et al: Tracking the influence of mental health treatment: The development of the Youth Outcome Questionnaire. *Clinical Psychology and Psychotherapy* 8:361-379, 2001
 8. Blank MB, Koch, JR, Burkett BJ: Less is more: Virginia's performance outcomes measurement system. *Psychiatric Services* 55:643-645, 2004
 9. Newman FL, Ciarlo JA, Carpenter D: Guidelines for selecting psychological instruments for treatment planning and outcome assessment, in *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment*, 2nd ed. Edited by Maruish ME. Mahwah, NJ, Lawrence Erlbaum Associates, 1999
 10. Vermillion JM, Pfeiffer SI: Treatment outcome and continuous quality improvement: Two aspects of program evaluation. *Psychiatric Hospital* 24:9-14, 1993
 11. Hermann RC, Palmer RH: Common ground: A framework for selecting core quality measures for mental health and substance abuse care. *Psychiatric Services* 53:281-287, 2002
 12. Hermann RC, Leff H, Palmer RH, et al: Quality measures for mental health care: Results from a national inventory. *Medical Care Research and Review* 57:135-153, 2000
 13. Manderscheid R, Henderson M: The field needs to agree on quality measures. *Behavioral Health Accreditation and Accounting Alert*, February:4-5, 2001
 14. Brook, RH: The RAND/UCLS appropriateness method, in *Clinical practice guideline development: methodology perspectives*. Edited by McCormick KA, Moore SR & Siegel RA. Rockville, MD, Agency for Health Care Policy and Research, 1994
 15. McGlynn EA, Kerr EA, Asch SM: New approach to assessing the clinical quality of care for women: The QA tool system. *Womens Health Issues* 9:184-192, 1999
 16. McGlynn EA: Choosing and evaluating clinical performance measures. *Joint Commission Journal of Quality Improvement* 24:470-479, 1998
 17. Donabedian A: *Explorations in Quality Assessment and Monitoring: The Definition of Quality and Approaches to its Assessment*. Ann Arbor, Michigan, Health Administration Press, 1980
 18. McGrath BM, Tempier RP: Implementing quality measures in psychiatry: From theory to practice—shifting from process to outcome. *Canadian Journal of Psychiatry* 48:467-474, 2003
 19. Hermann RC, Finnerty M, Provost S, et al: Process measures for the assessment and improvement of

- quality of care for schizophrenia. *Schizophrenia Bulletin* 28:95-104, 2002
20. Srebnik D, Hendryx M, Stevenson J, et al: Development of outcome indicators for monitoring the quality of public mental health care. *Psychiatric Services* 48:903-909, 1997
 21. Brown GS, Burlingame GM, Lambert MJ, et al: Pushing the quality envelope: A new outcomes management system. *Psychiatric Services* 52:925-934, 2001
 22. Modai I, Ritsner M, Silver H, et al: A computerized patient information system in a psychiatric hospital. *Psychiatric Services* 52:476-478, 2002
 23. Lambert MJ: Emerging methods for providing clinicians with timely feedback of effective treatment. In *Sessions: Journal of Clinical Psychology* 61, in press
 24. Ellwood PM: Shattuck lecture outcomes management: A technology of patient experience. *New England Journal of Medicine* 318:1549-56, 1988
 25. Bachrach LL: Assessment of outcomes in community support systems: results, problems and limitations. *Schizophrenia Bulletin* 8:39-60, 1982
 26. Froyd JE, Lambert MJ, Froyd JD: A review of practices of psychotherapy outcome measurement. *Journal of Mental Health* 5:11-15, 1996
 27. Erbes C, Polusny MA, Billig J, et al: Developing and applying a systematic process for evaluation of clinical outcome assessment instruments. *Psychological Services* 1:31-39, 2004
 28. Rowan T, O'Hanlon WH: *Solution-Oriented Therapy for Chronic and Severe Mental Illness*. New York, John Wiley & Sons, 1999
 29. Soreff SM: *Handbook for the Treatment of the Seriously Mentally Ill*. Kirkland, WA, Hogrefe and Huber Publishers, 1996
 30. Rothbard AB, Schinnar AP, Goldman H: The pursuit of a definition for severe and persistent mental illness, in *Handbook for the Treatment of the Seriously Mentally Ill*. Edited by Soreff SM. Kirkland, WA, Hogrefe & Huber Publishers, 1996
 31. Carey MP, Carey KB: Behavioral research on severe and persistent mental illnesses. *Behavioral Therapy* 30:345-353, 1999
 32. Page AC, Hooke GR, Rutherford EM: Measuring mental health outcomes in a private psychiatric clinic: Health of the Nation Outcome Scales and Medical Outcomes Short Form SF-36. *Australian and New Zealand Journal of Psychiatry* 35:377-381, 2001
 33. Piersma H, Boes J: The GAF and psychiatric outcome: A descriptive report. *Community Mental Health Journal* 33:35-41, 1997
 34. Piersma HL, Reaume WM, Boes JL: The Brief Symptom Inventory (BSI) as an outcome measure for adult psychiatric inpatients. *Journal of Clinical Psychology* 50:555-563, 1994
 35. Sayers SL, Curran PJ, Mueser KT: Factor structure and construct validity of the Scale for the Assessment of Negative Symptoms. *Psychological Assessment* 8:269-280, 1996
 36. Wallace CJ, Liberman RP, Tauber R, et al: The Independent Living Skills Survey: A comprehensive measure of the community functioning of severely and persistently mentally ill individuals. *Schizophrenia Bulletin* 26:631-658, 2000
 37. Bell M, Milstein R, Beam-Goulet J, et al: The Positive and Negative Syndrome Scale and the Brief Psychiatric Rating Scale. *Journal of*

- Nervous and Mental Disease 180:723-728, 1992
38. Brekke JS: An examination of the relationships among three outcome scales in schizophrenia. *Journal of Nervous & Mental Disease* 180:162-167, 1992
39. Cramer JA, Rosenheck R, Xu W, et al: Quality of life in schizophrenia: A comparison of instruments. *Schizophrenia Bulletin* 26:659-666, 2000
40. Green RS, Gracely EJ: Selecting a rating scale for evaluating services to the chronically mentally ill. *Community Mental Health Journal* 23:91-102, 1987
41. Welham J, Stedman T, Clair A: Choosing negative symptom instruments: Issues of representation and redundancy. *Psychiatry Research* 87:47-56, 1999
42. Earnshaw D, Rees F, Dunn TW, et al: Implementing a multi-source outcome assessment protocol in a state psychiatric hospital: A case study from the public sector. In this issue
43. American Psychiatric Association: *Handbook of psychiatric measures*. Washington, DC, Author, 2000
44. Maruish M (ed): *The use of psychological testing for treatment planning and outcome assessment*, 3rd ed. Mahwah, NJ, Lawrence Erlbaum Associates, 2004
45. Hill CE, Lambert MJ: Methodological issues in studying psychotherapy processes and outcomes, in Bergin and Garfield's *Handbook of Psychotherapy and Behavior Change*, 5th ed. Edited by Lambert MJ. New York, John Wiley, 2004
46. Rosenblatt A, Wyman N, Kingdon D, et al: Managing what you measure: Creating outcome driven systems of care for youth with serious emotional disturbances. *Journal of Behavioral Health Services Research* 25:177-193, 1998
47. Burlingame GM, Earnshaw D, Hoag M, et al: A systematic program to enhance clinician group skills in an inpatient psychiatric hospital. *International Journal of Group Psychotherapy* 52:555-587, 2002
48. Bednar R, Burlingame GM, Masters K: Systems of family treatment: Substance or semantics? In *Annual Review of Psychology*, vol 39. Edited by Rosenzweig MR & Porters LW. Palo Alto, CA, Annual Reviews Inc, 1988
49. Overall JE, Gorham DR: The Brief Psychiatric Rating Scale. *Psychological Reports* 10:799-812, 1962
50. American Psychiatric Association: *Diagnostic and statistical manual of mental disorders*, 4th ed. Washington, DC, Author, 1994
51. Kay SR, Fiszbein A, Opler LA: The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* 13:261-276, 1987
52. Andreasen NC: Negative symptoms in schizophrenia: Definition and reliability. *Archives of General Psychiatry* 39:784-788, 1982
53. Derogatis LR, Cleary PA: Confirmation of the dimensional structure of the SCL-90: A study in construct validation. *Journal of Clinical Psychology* 33:981-989, 1977
54. Lehman AF: A Quality of Life Interview for the chronically mentally ill. *Evaluation and Program Planning* 11:51-62, 1988
55. Ciarlo JA, Brown TR, Edwards DW, et al: *Assessing Mental Health Treatment Outcome Measurement Techniques*: DHHS Publication No. 86-1301. Washington, DC, US Government Printing Office, 1986
56. Ogles BM, Lambert MJ, Masters KS: *Assessing outcome in clinical*

- practice. Needham Heights, MA, Allyn & Bacon, 1996
57. Ventura J, Green MF, Shaner A, et al: Training and quality assurance with the Brief Psychiatric Rating Scale: The drift busters. *International Journal of Methods in Psychiatric Research* 3:221-244, 1993
58. Overall JE: The Brief Psychiatric Rating Scale, in ECDEU Assessment Manual. Edited by Guy W. Rockville, MD, National Institute of Mental Health, 1976
59. Lukoff D, Nuechterlein KH, Ventura J: Manual for the expanded BPRS. *Schizophrenia Bulletin* 12:594-602, 1986
60. Cone JD: *Evaluating Outcomes: Empirical Tools for Effective Practice*. Washington, DC, American Psychological Association, 2001
61. Hedlund JL, Vieweg BW: The Brief Psychiatric Rating Scale (BPRS): A comprehensive review. *Journal of Operational Psychiatry* 11:49-65, 1980
62. Hafkenscheid A: Psychometric evaluation of a standardized and expanded Brief Psychiatric Rating Scale. *Acta Psychiatrica Scandinavica* 84:294-300, 1991
63. Ventura J, Nuechterlein KH, Subotnik KL, et al: Symptom dimensions in recent-onset schizophrenia and mania: A principal components analysis of the 24-item Brief Psychiatric Rating Scale. *Psychiatry Research* 97:129-135, 2000
64. Lachar D, Bailey SE, Rhoades HM, et al: New subscales for an anchored version of the Brief Psychiatric Rating Scale: Construction, reliability, and validity in acute psychiatric admissions. *Psychological Assessment* 13:384-395, 2001
65. Mueser KT, Curran PJ, McHugo GJ: Factor structure of the Brief Psychiatric Rating Scale in schizophrenia. *Psychological Assessment* 9:196-204, 1997
66. Thiemann S, Csernansky JG, Berger PA: Rating scales in research: The case of negative symptoms. *Psychiatry Research* 20:47-55, 1987
67. Newcomer JW, Faustman WO, Yeh W, et al: Distinguishing depression and negative symptoms in unmedicated patients with schizophrenia. *Psychiatry Research* 31:243-250, 1990
68. American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed, Text Revision. Washington, DC, Author, 2002
69. Endicott J, Spitzer RL, Fleiss JL, et al: The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry* 33:766-771, 1976
70. Hall R: Global Assessment of Functioning: A modified scale. *Psychomatics* 36:267-275, 1995
71. Jones SH, Thornicroft G, Coffey M, et al: A brief mental health outcome scale: Reliability and validity of the Global Assessment of Functioning (GAF). *British Journal of Psychiatry* 166:654-659, 1995
72. Schwartz RC, Cohen BN, Grubaugh A: Does insight affect long-term inpatient treatment outcome in chronic schizophrenia? *Comprehensive Psychiatry* 38:283-288, 1997
73. Hilsenroth MJ, Ackerman SJ, Blagys MD, et al: Reliability and validity of DSM-IV axis V. *American Journal of Psychiatry* 157:1858-1963, 2000
74. Startup M, Jackson MC, Bendix S: The concurrent validity of the Global Assessment of Functioning (GAF). *British Journal of Clinical Psychology* 41:417-422, 2002
75. Singh MM, Kay SR: A comparative study of haloperidol &

- chlorpromazine in terms of clinical effects & therapeutic reversal with benzotropine in schizophrenia: Theoretical implications for potency differences among neuroleptics. *Psychopharmacologia* 43:103-113, 1975
76. Cuesta MJ, Peralta V: Psychopathological dimensions in schizophrenia. *Schizophrenia Bulletin* 21:473-482, 1995
77. Norman RM, Malla AK, Cortese L, et al: A study of the interrelationship between and comparative interrater reliability of the SAPS, SANS and PANSS. *Schizophrenia Research* 19:73-85, 1996
78. Andreasen NC: Scale for the Assessment of Negative Symptoms (SANS). Iowa City, Department of Psychiatry, University of Iowa College of Medicine, 1984
79. Schuldberg D, Quinlan DM, Morgenstern H, et al: Positive and negative symptoms in chronic psychiatric outpatients: Reliability, stability, and factor structure. *Psychological Assessment* 2:262-268, 1990
80. Hafkenscheid A: Psychometric evaluation of the Symptom Checklist (SCL-90) in psychiatric patients. *Personality and Individual Differences* 14:751-756, 1993
81. Derogotis LR, Melisaratos N: The Brief Symptom Inventory: An introductory report. *Psychological Medicine* 13:595-605, 1983
82. Schmitz N, Hartkamp N, Franke GH: Assessing clinically significant change: Application to the SCL-90-R. *Psychological Reports* 86:263-274, 2000
83. Morlan KK, Tan S: Comparison of the Brief Psychiatric Rating Scale and the Brief Symptom Inventory. *Journal of Clinical Psychology* 54:885-894, 1998
84. Howard PB, El-Mallakh P, Rayens, MK, et al: Consumer perspectives on quality of inpatient mental health services. *Archives of Psychiatric Nursing* 17:205-217, 2003
85. Eisen SV, Leff HS, Schaefer E: Implementing outcome systems: Lessons form a test of the BASIS-32 and the SF-36. *The Journal of Behavioral Health Services and Research* 26:18-27, 1999
86. Uttaro T, Lehman A: Graded response modeling of the Quality of Life Interview. *Evaluation & Program Planning* 22:41-52, 1999
87. McNary SW, Lehman AF, O'Grady KE: Measuring subjective life satisfaction in persons with severe and persistent mental illness: A measurement quality and structural model analysis. *Psychological Assessment* 9:503-507, 1997
88. Corrigan PW, Buican B: The construct validity of subjective quality of life for the severely mentally ill. *Journal of Nervous and Mental Disease* 183:281-285, 1995
89. Russo J, Roy-Byrne P, Reeder D, et al: Longitudinal assessment of quality of life in acute psychiatric inpatients: Reliability and validity. *Journal of Nervous and Mental Disease* 185:166-175, 1997
90. Phelan M, Wykes T, Goldman H: Global function scales. *Social Psychiatry and Psychiatric Epidemiology* 29:205-211, 1994
91. Moos R, McCoy L, Moos B: Global Assessment of Functioning (GAF) ratings: Determinants and role as predictors of one-year treatment outcomes. *Journal of Clinical Psychology* 56:449-461, 2000
92. Hoyt WT: Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods* 5:64-86, 2000
93. Baigent L, Ostbye T, Femando MLD: Feasibility of client reports to measure treatment outcome in schizophrenia. *Canadian Journal of Psychiatry* 44:94-95, 1999

- | | |
|--|---|
| <p>94. Hamera EK, Schneider JK, Potocky M, et al: Validity of self-administered symptom scales in clients with schizophrenia and schizoaffective disorders. <i>Schizophrenia Research</i> 19:213-219, 1996</p> <p>95. Rogers R: <i>Clinical Assessment of Malingering and Deception</i>, 2nd ed. New York, Guilford Press, 1997</p> <p>96. Hu L, Bentler PM: Cutoff criteria for fit indexes in covariance</p> | <p>structure analysis: Conventional criteria versus new alternatives. <i>Structural Equation Modeling</i> 6:1-55, 1999</p> <p>97. Cohen J: A power primer. <i>Psychological Bulletin</i> 112:155-159, 1992</p> <p>98. Lipsey MW, Wilson DB: <i>Practical Meta-Analysis</i>. Thousand Oaks, CA, Sage, 2001</p> |
|--|---|

Table 1

Most Frequently Used Outcome Measures in Populations Diagnosed with SPMI

Names of Outcome Measures ^a	Number of Times Used	Percentage of Use ^b
Brief Psychiatric Rating Scale (BPRS)	30	44.1
Global Assessment of Functioning Scale (GAF) ^c	26	38.2
Positive and Negative Syndrome Scale (PANSS)	13	19.1
Scale for the Assessment of Negative Symptoms (SANS)	13	19.1
Clinical Global Impression (CGI)	7	10.3
Abnormal Involuntary Movement Scale (AIMS)	6	8.8
Lehman's Quality of Life Interview (QOLI)	6	8.8
Scale for the Assessment of Positive Symptoms (SAPS)	5	7.4
Symptom Checklist-90-Revised (SCL-90-R)	5	7.4
Nurses' Observation Scale for Inpatient Evaluation (NOSIE)	4	5.9

^a Reports outcome measures used four or more times. ^b Percentage of use was calculated using the number of articles that employed an outcome instrument used four or more times (N=68). The 26 remaining articles were not included in this calculation because they used more specialized or infrequency used outcome instruments. ^c The number of times used for the GAF includes nine uses of the GAS, an earlier version of the GAF.

Table 2
Evaluation of Frequently Used Clinician-Rated and Self-Report Outcome Measures

Name of Measure ^a	Specialty Population ^b	Type of Training Material	Normative Data Groups ^c	Reliability Coefficients ^d	Validity Coefficients ^d
Clinician-Rated Measures					
BPRS	Psychiatric inpatients, & schizophrenia	Training manual, videos & quality assurance programs	Normal, psychiatric inpatient & outpatient	Inter-rater: from .73 to .87 Test-retest: $r > .70$ for 8 items of BPRS-18 Internal consistency: from .79 to .83	Content: Goodness of Fit Index $> .90$ for BPRS and BPRS-18 Concurrent: SANS ($r = .70$), Hamilton Rating Scale for Depression ($r = .80$)
GAF	Psychiatric inpatients	Training manual	Psychiatric inpatient & outpatient	Inter-rater: from .62 to .96	Concurrent: Global Severity Index of the SCL-90-R ($r = -.46$), SANS ($r = -.63$), SAPS ($r = -.48$)
PANSS	Schizophrenia	Training video and manual	Psychiatric inpatient	Inter-rater: from .53 to .91 Test-retest: .67 Internal consistency: from .73 to .83	Content: Goodness of Fit Index $> .90$ Concurrent: SANS ($r = .54$), SAPS ($r = .68$)
SANS	Schizophrenia	Training manual	Schizophrenia	Inter-rater: from .53 to .93 Test-retest: .50 Internal consistency: .89	Construct: Goodness of Fit Index $> .90$ Concurrent: PANSS ($r = .54$)
Self-Report Measures					
SCL-90-R	Psychiatric outpatients	Training video and manual	Normal, psychiatric inpatient & outpatient	Test-retest: from .66 to .91 Internal consistency: from .71 to .97	Concurrent: BSI ($r = .92$ to .99)
QOLI	Psychiatric inpatients	Training video and manual	Normal & psychiatric inpatient	Test-retest: .68 Internal consistency: from .80 to .92	Construct: Comparative Fit Index $> .90$ Concurrent: BPRS Depression Subscale ($r = -.37$), Social Functioning and Support Scale ($r = .48$)

^a BPRS=Brief Psychiatric Rating Scale, GAF=Global Assessment of Functioning Scale, PANSS=Positive and Negative Syndrome Scale, SANS=Scale for the Assessment of Negative Symptoms, SCL-90-R=Symptom Checklist-90-Revised, QOLI=Lehman's Quality of Life Interview. ^b The term "specialty population" seeks to capture the specific population for which each instrument was developed. ^c Identifies populations for which normative data is available. ^d References of this data are provided in the text. Cutoff scores have been assigned by Burlingame et al. (6), who established the cutoff for internal consistency reliability (coefficient alphas) at .80, test-retest reliability at .70, and concurrent validity coefficients at or above .50, and Ventura et al. (57), who decided that inter-rater reliability scores should be at least .80 or higher for outcome instrumentation. Also, in order to possess construct validity, Goodness of Fit indices or Comparative Fit Indices should be .90 or above (96).

Name of Measure ^a	Availability of Instrument	Time to Administer Measure	Sensitivity to Change ^e (<i>d</i> / # of studies)
Clinician-Rated Measures			
BPRS	Public Domain	10-30 minutes	1.21 / 18
GAF	Proprietary	5-15 minutes	1.10 / 10
PANSS	Proprietary	30-40 minutes	1.23 / 7
SANS	Public Domain	15-30 minutes	.68 / 5
Self-Report Measures			
QOLI	Proprietary	20-45 minutes	.02 / 1
SCL-90-R	Proprietary	12-15 minutes	.69 / 2

^eSensitivity to change was determined for each measure by calculating effect sizes across patient scores before and after treatment. Specifically, means and standard deviations of pre and post-treatment scores or probability values were used to calculate *d* values, which have been operationalized as small (*d* = .20), medium (*d* = .50), and large (*d* = .80) (97,98). All published and unpublished studies (*N* = 32) over the last ten years were included in the analysis if they: (1) used one of the six instruments, (2) assessed patient change in a populations diagnoses with SPMI, (3) reported sufficient statistics to calculate an effect size, and (4) employed a sample size of at least 10. Study design, length of treatment for individuals with SPMI, sample size, population type (acute vs. chronic SMPI), treatment type (pharmaceutical vs. psychosocial), and choice of instrument did not significantly influence effect sizes. The average effect size is reported for each measure, along with the number of studies used in the average effect size calculation for each measure.

Figure 1: Flow Chart Depicting Proposed Method for Selecting Outcome Measures

