Comparing the Sensitivity of Generic Effectiveness Measures With Symptom Improvement in Persons With Schizophrenia

JEFFREY M. PYNE, MD,* GREER SULLIVAN, MD, MSPH,* ROBERT KAPLAN, PHD,† AND D. KEITH WILLIAMS, PHD‡

OBJECTIVE. The purpose of this study was to compare the sensitivity of four generic effectiveness measures with clinically meaningful symptom improvement in persons with schizophrenia.

METHOD. Baseline and 6-month interviews were conducted with 134 subjects diagnosed with schizophrenia or schizoaffective disorder. The design was observational. The four generic effectiveness measures included the Quality of Well-Being scale (QWB), a quality-adjusted index score based on the SF-36 VAS, Veterans SF-36 mental health component summary score (MCS), and the World Health Organization Disablement Assessment Schedule (WHO-DAS). Symptom measures included the Positive and Negative Syndrome Scale (PANSS) and Calgary Depression Scale (CDS). The side effect measure was the Extrapyramidal Symptom Rating Scale (ESRS). Data analysis included correlations between symptom, side effect, and generic effectiveness change scores; and an effect size calculation to detect a clinically significant improvement in the total PANSS.

RESULTS. All four effectiveness measures were correlated with changes in side effects. All but the SG-36 VAS were correlated with changes in depression. Only the QWB was correlated with changes in PANSS scores. The QWB required at least three times fewer subjects (n = 61) to detect a clinically significant improvement in total PANSS compared with the other effectiveness measures (n = 201–324).

CONCLUSIONS. It is recommended that clinicians and researchers use the QWB to demonstrate the effectiveness and cost-effectiveness of schizophrenia interventions. The QWB allows for direct comparison of the effectiveness and cost-effectiveness of schizophrenia interventions with other mental and physical health interventions and may contribute to a greater recognition of the value of mental health interventions. (Med Care 2003;41:208–217)

In 1993 the US Public Health Service convened a panel on cost-effectiveness analysis and recommended procedures to standardize the methods used for economic evaluations. The panel urged

*From the South Central VA Healthcare Network, and the Department of Psychiatry, Central Arkansas Veterans Healthcare System, the University of Arkansas for Medical Sciences, Little Rock, Arkansas.
†From the Department of Family and Preventive Medicine, University of California San Diego, La Jolla, California.
‡From the Department of Biometry, University of Arkansas for Medical Sciences, Little Rock, Arkansas.

Supported by a VA Career Development Award from the Department of Veterans Affairs Health Services Research & Development Service (Grant #CD-97310-2) and the South Central (VISN 16) MIRECC.

An earlier version of this work was presented at the 2001 APA Annual Meeting, New Orleans, LA.

Address correspondence and reprint requests to: Jeffrey M. Pyne, MD, Central Arkansas Veterans Healthcare System, Department of Psychiatry, 116F2/NLR, 2200 Fort Roots Drive, North Little Rock, AR 72114. E-mail: pynejeffreym@uams.edu

Received January 9, 2002; initial review March 15, 2002; accepted July 16, 2002.
the use of generic measures of effectiveness, such as the quality-adjusted life year (QALY), to facilitate the direct comparison of outcomes across disorders.\textsuperscript{1} Demonstrating the value of mental health interventions using a metric common to both mental and physical disorders is especially important in an increasingly competitive health care marketplace.\textsuperscript{2} Currently, cost-effectiveness analyses (CEAs) of mental health interventions are relatively uncommon compared with CEAs of physical health interventions. Without data that uses a common effectiveness metric, arguments for maintaining or increasing funding for mental health interventions will lack the advantage of direct comparisons with physical health interventions.

Pharmacological and psychosocial treatment advances for schizophrenia are typically associated with higher costs. CEAs can determine the degree to which additional costs are associated with greater treatment effectiveness. Few studies in the schizophrenia literature have evaluated the cost-effectiveness of interventions and these studies do not use generic effectiveness measures. Instead, they are typically based on outcomes specific to persons with schizophrenia.\textsuperscript{3–7} The downside to this approach is that the results can only be compared with other schizophrenia interventions, not with interventions for other mental or physical illnesses.

One reason for the lack of generic cost-effectiveness studies for schizophrenia interventions is the concern that generic measures may not be sensitive to symptom changes considered important in the treatment of mental illness\textsuperscript{8} or schizophrenia in particular.\textsuperscript{9} We therefore designed this study to directly compare the sensitivity to clinically meaningful schizophrenia symptom change of generic effectiveness measures. Specifically, we compared the sensitivity of four generic effectiveness measures. Two measures were derived from existing preference-weighted health-related quality of life (HRQL) measures: the Quality of Well-Being scale (QWB) and a quality-adjusted index score based on the SF-36 (SF-36 VAS).\textsuperscript{10} Preference-weighted measures incorporate the desirability of HRQL outcomes on a scale bounded by death (0.0) and perfect health (1.0) and are used to calculate QALYs. The other two measures were HRQL measures that used standardized scoring: the mental health component summary score (MCS) from the Veterans SF-36 and the World Health Organization Disablement Assessment Schedule (WHO-DAS). Standardized measures directly convert responses to a HRQL questionnaire to another metric (eg, a 0–100 scale) and are not preference-weighted.

These particular measures were chosen because they were either designed as a QALY measure (QWB), can be converted into a QALY score (SF-36), or currently provide a standardized index score with plans to develop an algorithm for QALY scoring (WHO-DAS). We hypothesized that the QWB would be the most sensitive to schizophrenia symptom change because it includes a wider range of mental health symptoms than the other measures.

Materials and Methods

Overview of Study Design

The study design was observational and longitudinal; the treatment interventions were not controlled or influenced by the research team. Subjects were interviewed at baseline and 6 months. Inpatients were interviewed within 3 days of admission. All the symptom, side effect, and effectiveness measures were administered to each subject at baseline and 6 months. The order of generic effectiveness measure administration was randomized at the time of each interview. Separate interviewers conducted the symptom and HRQL interviews to maintain the independence of symptom and generic effectiveness assessment.

Subjects

A convenience sample of subjects was recruited from the South Central Veterans Health care System inpatient and outpatient mental health units. Inpatients and outpatients were recruited to capture a wide range of schizophrenia symptom severity at baseline. Subject selection criteria required that patients (1) had a Structured Clinical Interview for DSM-IV diagnosis of a primary psychotic disorder (schizophrenia or schizoaffective disorder); (2) were at least 18 years old; (3) planned to stay in the Central Arkansas area for the next year; (4) provided the name and phone number of a contact person; and (5) signed informed consent. The use of patients with schizophrenia (n = 85) or schizoaffective disorder (n = 49) ensured a more typical sample of patients.
with a primary psychotic disorder. Of the 173 subjects that were screened for the study, 139 were eligible, 134 completed the baseline interview (five subjects were enrolled then refused to complete the baseline interview), and 112 (84%) completed the 6-month interview. Of those subjects who did not complete the 6-month interview, eight were not located, eight were located but did not keep follow-up appointments, five were located but withdrew from the study, and one subject died.

Measures

All the measures were interviewer-administered. The symptom measures included the Positive and Negative Syndrome Scale (PANSS) and Calgary Depression Scale (CDS). The PANSS is a 30-item schizophrenia-specific scale that includes a total score and subscale scores for positive (e.g., hallucinations, delusions, and paranoia), negative (e.g., social and emotional withdrawal and blunted affect) and general (e.g., anxiety, depression, poor attention) symptoms during the past week. The PANSS interrater reliability for this study was 0.92. The CDS is a 9-item scale that includes items that are unique to depression and do not overlap with symptoms of schizophrenia covers the past 2 weeks.

The side effect measure was the subjective extrapyramidal symptom subscale from the Extrapyramidal Symptom Rating Scale (ESRS). The subjective extrapyramidal symptom subscale includes 11 items and measures the subjective severity of medication side effects such as tremor, stiffness, restlessness, and abnormal involuntary movements during the past week.

The acceptability of the generic effectiveness measures from the subjects’ perspective was determined using the Patient Acceptance Scale (PAS). The PAS is a 6-item scale that includes three subscales: four items for negative affect burden (including feeling embarrassed, upset, annoyed, and uncomfortable), and one item each for the questionnaire length and ease of answering the questionnaire. The PAS was administered immediately after the subject completed each of the effectiveness measures.

The generic effectiveness measures included the Veterans SF-36 MCS, SF-36 VAS, WHO-DAS, and QWB (Table 1). The Veterans SF-36 is a VA-adapted version of the original Medical Outcomes Study SF-36 and is reported to be more sensitive than the original SF-36. Adaptations to the original SF-36 included changing two role limitation response sets from yes/no to five-point ordinal responses ranging from “no, none of the time” to “yes, all the time.” The output from the Veterans SF-36 includes the same eight subscales, physical health component scale (PCS), and mental health component scale (MCS) as the original SF-36. The period of time covered for most items is 4 weeks. Scoring for the Veterans SF-36 is based on the original SF-36 formula. The Veterans SF-36 has been used to measure the HRQL associated with the following chronic illnesses: hypertension, diabetes, chronic lung disease, osteoarthritis, low back pain, and alcohol related disorders.

The SF-36 VAS was scored by converting the Veterans SF-36 responses into a quality-adjusted index score using a visual analogue scale (VAS) conversion formula developed for the original SF-36. The VAS is a method for obtaining preference weights for health states using a scale from 0 (equal to death) to 100 (equal to perfect health) and a prop that is similar to a thermometer. For example, an SF-36 health state with a VAS score of 50 can be thought of as half way between death and perfect health and would be equal to a preference weight of 0.5. As mentioned above, the primary difference between the SF-36 and Veterans SF-36 was the expansion of the response sets for two questions. To use the SF-36 VAS conversion formula, we modified the Veterans SF-36 response sets to reflect those in the original SF-36 and then calculated the quality-adjusted index score as described by Brazier et al. The VAS conversion formula was developed by Brazier et al in three steps: (1) create a simplified health state classification system from the SF-36; (2) obtain preference weights corresponding to the desirability of various descriptions of health and functioning (health states) based on the SF-36; and (3) develop a formula for estimating the preference-weighted index score for any set of responses to the SF-36. The preference-weighted score was bounded by 0.0 (death) and 1.0 (perfect health) to facilitate calculation of QALYs. One QALY equals 1 year of life free of any health problems or functional impairments.

The WHO-DAS was developed by the World Health Organization (WHO) as an international instrument for measuring disability. The current scoring system provides a standardized total score and the following subscales scores: understanding and communicating, getting around, taking care of

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
yourself, getting along with people, household activities, and participation in society. Preference weights for calculating a quality-adjusted index score are planned for the near future, at which time the WHO-DAS will be the WHO measure for QALYs.

The QWB scale was designed to measure QALYs.\textsuperscript{20} Using this method patients are classified according to objective levels of functioning represented by mobility, physical activity, and social activity subscales. In addition to classification by functioning, patients are also classified by chief symptom or problem. Symptoms or problems may be severe, such as trouble learning, remembering, or thinking clearly; or minor such as taking medication. The QWB preference weights were collected from a general population sample. The QWB has been used in a wide variety of different studies including cancer, Alzheimer disease, chronic obstructive pulmonary disease, AIDS, cystic fibrosis, diabetes mellitus, sinusitis, atrial fibrillation, lung transplantation, arthritis, cancer, schizophrenia, and several other conditions.\textsuperscript{20–24} Further, the QWB was designed to inform health care resource allocation decisions and initially served as the basis for an innovative experiment on rationing health care by the state of Oregon.\textsuperscript{25}

### Table 1. Summary Description of Generic Effectiveness Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>No. items</th>
<th>No. days\textsuperscript{\dagger}</th>
<th>Subscales</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>QWB</td>
<td>41–54\textsuperscript{*}</td>
<td>3</td>
<td>Symptom/problem</td>
<td>Preference weighted</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mobility</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Physical activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Social activity</td>
<td></td>
</tr>
<tr>
<td>WHO-DAS</td>
<td>64–73\textsuperscript{*}</td>
<td>30</td>
<td>Understanding/communicating</td>
<td>Standardized</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Getting around</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Taking care of yourself</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Getting along with people</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Life activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Participation in society</td>
<td></td>
</tr>
<tr>
<td>SF-36 VAS</td>
<td>36</td>
<td>28</td>
<td>Physical functioning</td>
<td>Preference weighted</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Role limitation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bodily pain</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Energy/vitality</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Social functioning</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mental health</td>
<td></td>
</tr>
<tr>
<td>Veterans SF-36 MCS</td>
<td>36</td>
<td>28</td>
<td>Physical functioning</td>
<td>Standardized</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Role limitations–physical</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bodily pain</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>General health perceptions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Energy/vitality</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Social functioning</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Role limitations–emotional</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mental health</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Summary scales:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Physical component (PCS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mental component (MCS)</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{*}The range of items is due to the branching pattern of the questionnaire and is dependent on subject responses.

\textsuperscript{\dagger}The number of days assessed by the scale.

QWB indicates Quality of Well-Being Scale; WHO-DAS, World Health Organization Disablement Assessment Schedule; SF-36 VAS, Medical Outcome Study SF-36 visual analogue scale conversion to quality-adjusted index score; Veterans SF-36 MCS, Veterans version of the SF-36 mental health component summary score.
Statistical Analysis

The distributions for completion time and patient acceptance of each the effectiveness measures were skewed. Therefore, completion time and patient acceptance of each effectiveness measure were compared using the Wilcoxon rank sum test. Change scores were calculated by subtracting baseline from 6-month scores. Pearson correlation coefficients were used to measure the relationship between change scores of the effectiveness, symptom, and side effect measures. The effect size was estimated by subtracting the mean change score of subjects who did not demonstrate clinically significant symptom improvement from the mean change score of subjects who demonstrated clinically significant symptom improvement and dividing this result by the pooled SD for the two groups. Clinically significant symptom improvement was defined by a 20% improvement in total PANSS score between the baseline and 6-month interview because this definition is commonly used in schizophrenia intervention studies.11,26,27 Effect sizes were used to (1) translate the change scores for effectiveness measures into a common unit of measurement sensitivity and (2) to calculate the number of subjects needed per group to detect a clinically significant change. We defined the effect size as the mean difference in change scores divided by the pooled SD as is used in two-tailed t tests.

Results

Subjects

Sixty inpatient and 74 outpatient subjects completed baseline interviews. There were no significant baseline differences in age, gender, ethnicity, marital status, education, employment status, PANSS scores, CDS, ESRS, Veterans SF-36, SF-36 VAS, WHO-DAS, or QWB between those who did or did not complete the 6-month interview. However, outpatients at baseline were more likely to complete the 6-month interview than inpatients (90% vs. 75%, P = 0.02). The mean age of subjects was 46.8 years (SD = 8.1). Ninety-four percent (126/134) of subjects were male, 56% (75/134) were Black, 42% (56/134) were never married, 74% (99/134) had a high school education or greater, and 90% (120/134) were not working. Twenty-six percent (29/112) of subjects met clinically significant symptom improvement criteria of 20% improvement in total PANSS score. There were no significant baseline sociodemographic, symptom, or generic effectiveness measure differences between the 20% symptom improved and symptom not improved groups, except that the improved symptom group had a higher (more symptomatic) mean baseline total PANSS score (P = 0.002).

Acceptability of Scales

The negative affect burden reported by subjects using the PAS was significantly greater for the Veterans SF-36 (P = 0.03) and WHO-DAS (P = 0.007) compared with the QWB. There were no statistically significant differences between the effectiveness measures when subjects were asked about the length or the ease of answering the questionnaires.

Time to Complete Scales

The median minutes to complete the Veterans SF-36, QWB, and WHO-DAS were 10 (range 4–30), 17 (range 5–35), and 20 (range 8–60), respectively. The median times to complete the QWB and WHO-DAS were significantly greater (P <0.0001) than the median time to complete the Veterans SF-36.

Correlations Between Symptom and Effectiveness Change Scores

The correlation matrix for the effectiveness, symptom, and side effect change scores is shown in Table 2. All the effectiveness measure change scores (QWB, WHO-DAS, SF-36 VAS, and Veterans SF-36 MCS) were significantly correlated with the side effect change score. The QWB, WHO-DAS, and SF-36 MCS change scores were significantly correlated with depression change scores. The QWB change score was the only effectiveness measure that was significantly correlated with PANSS change scores.

Effect Size and Number Needed Per Group

Table 3 shows the change scores, standard deviations, effect sizes, and number of subjects needed to detect clinically significant symptom
improvement (20% improvement in total PANSS) during a 6-month period. The effect size for the QWB (0.36) was approximately twice that of the other effectiveness measures. Further subgroup analysis found that the effect size was greater for outpatients than inpatients for all effectiveness measures. For example, the effect sizes for inpatients using the same 20% symptom improvement criteria were all less than 0.10. The effect sizes for outpatients were 0.70 (QWB), 0.27 (Veterans SF-36 MCS), 0.27 (WHO-DAS), and 0.26 (SF-36 VAS). We also compared the effect sizes for more and less severely ill subjects based on the baseline total PANSS scores. The more severe group was defined by a baseline total PANSS score greater than 71 (n = 56) and the less severe group was defined by a baseline total PANSS score less than or equal to 71 (n = 56). In general, the effect sizes were greater for subjects in the less severe group (data not shown but available from first author on request).

### Table 2. Change Score Correlation Matrix for Symptoms, Side Effect, and Effectiveness Measures

<table>
<thead>
<tr>
<th></th>
<th>QWB</th>
<th>WHO DAS</th>
<th>SF-36 VAS</th>
<th>SF-36 MCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>−0.17</td>
<td>−0.05</td>
<td>−0.06</td>
<td>−0.009</td>
</tr>
<tr>
<td>Negative</td>
<td>−0.20*</td>
<td>−0.15</td>
<td>−0.18</td>
<td>−0.11</td>
</tr>
<tr>
<td>General</td>
<td>−0.22*</td>
<td>−0.12</td>
<td>−0.06</td>
<td>−0.10</td>
</tr>
<tr>
<td>Total</td>
<td>−0.30†</td>
<td>−0.16</td>
<td>−0.13</td>
<td>−0.12</td>
</tr>
<tr>
<td>Depression</td>
<td>−0.26†</td>
<td>−0.30†</td>
<td>−0.15</td>
<td>−0.27‡</td>
</tr>
<tr>
<td>Side effects</td>
<td>−0.33‡</td>
<td>−0.28‡</td>
<td>−0.25‡</td>
<td>−0.22*</td>
</tr>
</tbody>
</table>

*P < 0.05, †P < 0.01, ‡P < 0.001.

SF-36 VAS is the quality-adjusted SF-36 index score based on visual analogue scale weights.10
SF-36 MCS is the mental health component summary scale score of SF-36.

PANSS positive symptom subscale measures the severity of hallucinations, delusions, and paranoid thoughts. The negative symptom subscale measures social and emotional withdrawal, rapport, and blunted affect. The general symptom subscale measures depression, anxiety, orientation, and judgment. The total score is a linear combination of the three subscale scores.

### Table 3. Effectiveness Measure Effect Sizes and Number of Subjects Needed to Detect Clinically Significant Symptom Improvement

<table>
<thead>
<tr>
<th>Measure</th>
<th>Change Scores</th>
<th>Standard Deviation</th>
<th>Effect Size</th>
<th>No. Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Improved</td>
<td>Not Improved</td>
<td>Difference</td>
<td>Improved</td>
</tr>
<tr>
<td>QWB</td>
<td>0.023</td>
<td>−0.025</td>
<td>0.048</td>
<td>0.082</td>
</tr>
<tr>
<td>WHO DAS</td>
<td>39.37</td>
<td>9.76</td>
<td>29.61</td>
<td>102.51</td>
</tr>
<tr>
<td>SF-36 VAS</td>
<td>0.028</td>
<td>−0.015</td>
<td>0.043</td>
<td>0.197</td>
</tr>
<tr>
<td>SF-36 MCS</td>
<td>3.44</td>
<td>0.57</td>
<td>2.87</td>
<td>14.02</td>
</tr>
</tbody>
</table>

Clinically significant symptom improvement defined as 20% improvement in PANSS total score.

Effect size calculation is defined as the mean difference in change scores divided by the pooled standard deviation used in two-tailed t tests.

Number of subjects needed to detect 20% improvement in PANSS total score.

### Discussion

This is the first study to directly compare the sensitivity of generic effectiveness measures with symptom change in persons with schizophrenia. Among the effectiveness measures included in this study, the QWB was the only measure to be significantly correlated with schizophrenia-specific symptoms and the QWB resulted in a larger effect size compared with the other measures. These findings provide useful data to inform the increasingly important decision about which generic effectiveness measure to use in mental health outcomes studies.28 The four generic effectiveness measures tested were chosen because of their widespread use in the United States and internationally, and because each can be used to calculate QALYs (QWB and SF-36) or will soon be used to calculate QALYs (WHO-DAS).

The 26% symptom response rate in this study is similar to response rates reported for VA patients with schizophrenia treated with clozapine (26% at 6 months and 37% at 1 year) or haloperidol (12%...
at 6 months and 32% at 1 year) in a randomized double-blind study but lower than the 6-month response rate of 61% from an earlier open trial of clozapine treatment.27 From the subject perspective, the QWB appears to be preferred. The QWB was associated with significantly less negative affect compared with the other effectiveness measures. The median time to complete the QWB was approximately 7 minutes longer than the Veterans SF-36. This may have some impact on the cost of administering the QWB; however, the additional questionnaire administration time will be far less than the time required to recruit and interview additional subjects.

The change score correlation matrix results indicated that only the QWB change score was correlated with schizophrenia-specific symptom change scores (negative, general, and total PANSS change scores). There are at least three possible explanations for the QWB being more sensitive than the other measures to schizophrenia-specific symptom change. First, the QWB explicitly asks about a larger number of symptoms than the other effectiveness measures and includes cognitive items such as trouble learning, remembering, or thinking clearly which are common problems associated with schizophrenia and are highly weighted by the general public (meaning low desirability) in the QWB scoring program. Second, the QWB response set is Yes/No, which may be easier for patients with schizophrenia to answer than ordinal response sets. Third, the QWB asks subjects about the past 3 days, which is more similar to the 1-week timeframe of the PANSS than the past 4 weeks or 30 day timeframe used in the other effectiveness measures.

None of the effectiveness measures were significantly correlated with changes in positive symptoms. Other cross-sectional research has found negative symptoms to be consistently and significantly correlated with generic effectiveness measures and positive symptoms to be either similarly correlated, less strongly correlated, or not correlated with generic effectiveness measures. These results have lead some researchers to suggest that negative symptoms may be a more important determinant of generic HRQL than positive symptoms.31 The effect sizes for the generic effectiveness measures reported in Table 3 were all in the small effect size range. However, the QWB effect size was approximately twice the size of the other effectiveness measures. In subgroup analyses, it appeared that the effect sizes for all the effectiveness measures were larger for outpatients than inpatients and larger for the less severely symptomatic at baseline than the more severely symptomatic at baseline. For example, among outpatients (who are the typical subjects recruited for intervention trials) the number of subjects needed per group to detect a 20% symptom improvement using the QWB was only 22. These results indicate that the effectiveness measures are not as sensitive to a 20% symptom improvement in more severely ill patients. This finding could be explained by a threshold effect, meaning that more severely ill patients require a greater percent symptom improvement to improve their HRQL as measured by a generic effectiveness questionnaire than less severely ill patients. When we attempted to examine this explanation, however, the number of more severely ill subjects at baseline with a greater symptom improvement (eg, 30%) in our sample was too small (inpatients n = 4, and more severely ill patients regardless of patient status n = 5) to calculate reliable effect sizes for these groups. Future work among more severely ill patients is needed to further examine this possible threshold effect.

Based on the results reported in Table 3, the number of subjects required to detect a 20% symptom improvement change using the QWB was less than one-third the number of subjects required for the other generic effectiveness measures. The implication of this result is that the QWB can save substantial amounts of time and money in the design and conduct of effectiveness studies. For example, if a study includes a baseline and follow-up interview and it costs $100 to complete each interview then an estimate of the data collection cost would be $24,400 for the QWB and $80,400 to $129,600 for the other measures.

A common concern when interviewing patients with schizophrenia is their ability to reliably report their quality of life and level of functional impairment. However, numerous investigators have found that patients with schizophrenia are able to reliably report their quality of life or level of functional impairment. Another concern is that generic measures of effectiveness may not be appropriate treatment outcome measures for schizophrenia. We think that generic effectiveness measures are useful when the goal is to use a common metric to compare outcomes across disorders (eg, in cost-effectiveness analyses). We ac-
knowledge that not all outcomes for a given disorder will necessarily register on a given generic effectiveness measure and therefore we recommend the use of generic and disease-specific outcome measures when evaluating the effectiveness of an intervention. The generic measures will facilitate between disorder comparisons and the disease specific measures will facilitate within disorder comparisons.

There are important limitations to this study. The study used a convenience sample of predominantly male subjects from one clinical setting and therefore may not generalize to other settings. However, compared with the sample of patients from the Schizophrenia Patient Outcomes Research Team (PORT) study, the demographics of the current study are similar to a more general population of patients with schizophrenia (eg, 42% black in PORT, 56% black in current study; 70% high school education or greater in PORT, 74% in current study; and 54% never married in PORT, 42% never married in current study). In addition, most subjects were unemployed and this is typical of samples of patients with schizophrenia. Inpatients at baseline were less likely to complete the follow-up interview, which is most likely because they are generally less adherent to care than outpatients. Other generic effectiveness measures exist that were not tested in this study including the EuroQol, Health Utilities Index, and Assessment of Quality of Life. However, the EuroQol is based on a relatively simple descriptive system and therefore would likely be insensitive to schizophrenia symptom outcomes. The Health Utilities Index intentionally focuses on “within the skin” disability and therefore does not include social functioning, which is an important domain for schizophrenia outcomes. The Assessment of Quality of Life instrument appears promising, but was published after this study began. Our study design was observational, therefore we do not know if the results would be similar in an experimental study. In an experimental study, such as an intervention trial, the magnitude of symptom change may be greater because of more stringent subject selection criteria and tighter control of intervention conditions. We would not expect greater symptom change to alter the overall results of this study; in fact, an observational design may provide a more conservative estimate of the sensitivity of these measures. However, further study of generic effectiveness measures in experimental studies is warranted.

**Conclusion**

We recommend the use of the QWB to demonstrate the effectiveness and cost-effectiveness of schizophrenia interventions. The QWB allows for the direct comparison of the effectiveness and cost-effectiveness of schizophrenia interventions with other mental health and physical health interventions and may contribute to a greater recognition of the value of mental health interventions.

**Acknowledgments**

The authors thank Audrey Burnam, PhD, for helpful comments on an earlier manuscript and Guyla Craft and Marlyn Lipton for assistance with data collection.

**References**


216

PYNE ET AL


