Validity and reliability of an inpatient severity of psychiatric illness measure

BENTSON H. MCFARLAND,¹ ANNE E. KOVAS,¹ SHELBY L. HAUGAN,¹ DAVID A. POLLACK,¹ JO M. MAHLER¹

¹ Department of Psychiatry, Oregon Health and Science University, Portland OR, USA

Abstract

Inpatient psychiatric severity measures are often used but few psychometric data are available. This study evaluated the psychometric properties (reliability and validity) of a measure used to assess severity of psychiatric illness among inpatients. Using the severity measure, minimally trained raters conducted retrospective patient record reviews to assess medical necessity for psychiatric hospitalization. The data analysis compared 135 civilly committed psychiatric inpatients with a heterogeneous group of 248 psychiatric inpatients at a general hospital. The severity measure showed acceptable inter-rater reliability in both populations. Two-way analysis of variance showed that the intra-class correlation coefficient for the total score was 0.65 for general hospital subjects and 0.63 for civilly committed subjects. Differences in mean scores were substantial (15 out of a possible 75 points for general hospital subjects versus 42 for civilly committed subjects, Mann-Whitney $U = 562$, $p < 0.001$). As expected, all civilly committed subjects were well above admission cut-off score of 12, versus only 64% of the general hospital patients. The measure is appropriate for retrospective severity assessment and may also be useful for pre-admission screening.

Key words: health status indicators, hospitalizations, mental health, psychometrics, severity of illness

Introduction

The decision to admit a person to psychiatric inpatient treatment is complex and requires consideration of clinical, social, and contextual variables, including the severity of the individual's illness. Measures of severity are therefore useful for justifying admission to psychiatric inpatient programmes, for addressing utilization reviewers' concerns about medical necessity, and for research or program evaluation.

A frequently employed measure is the 12-item checklist developed by Flynn and Henisz (1975), which has been used to justify hospitalization (Newmark et al., 1977; Henisz et al., 1981), to predict hospitalization (Costello et al., 1991), and to rate inpatient severity retrospectively (McFarland et al., 2002). An attractive feature of the Flynn and Henisz

measure is its ease of use by lay people who have had minimal training. However, little information is available about psychometric properties of this measure. In particular, it is unknown whether or not the instrument's reliability depends on characteristics of the patient population.

In contradistinction to the Flynn and Henisz instrument, many severity measures require administration by a physician or a mental health professional. For example, Roy-Byrne and associates (1998) examined the reliability, validity, and clinical utility of their 11-item pre-hospitalization scale in a project that correlated scores with medical necessity as defined by managed care organizations (MCOs). Similarly, Glazer and Gray (1996) developed a pre-hospitalization decision-support scale that showed good inter-rater

reliability and validity when used by clinical raters to evaluate the level of care necessary for patients requiring psychiatric services in a health maintenance organization (HMO). Some mental health professional-administered measures have been developed for prospective and retrospective assessment. For example, the Severity of Psychiatric Illness (SPI) scale developed by Lyons and co-workers has shown acceptable validity for retrospectively assessing medical necessity in psycho-geriatric patients (Lyons et al., 1995) and for predicting psychiatric hospital admission, readmission, utilization, and outcome (Lyons et al., 1995; Lansing et al., 1997; Lyons et al., 1997). However, these instruments generally require administration by clinicians.

Severity measures have also been designed for lay use (see, for example, Flynn and Henisz, 1975; Durbin et al., 1999). Often, trained non-physician raters can complete these assessments retrospectively. Checklists that trained raters can use have several advantages over other forms of severity measures. For example, non-medical staff with minimal qualification and brief training can complete the checklist, which summarizes most of the criteria that suggest hospitalization. A lay person can also fill out a checklist without reference to diagnosis and without taking an extensive history because the rating is based on the patient's current behaviour and mental status (Flynn and Henisz, 1975). Therefore, these checklists are efficient time savers in both clinical and research settings.

The Flynn and Henisz severity measure used in this study evolved from Whittington's (1966) weighted checklist. This earlier checklist allowed relatively inexperienced workers (research assistants) at a mental health centre to evaluate a patient's need for hospitalization. Flynn and Henisz (1975) modified Whittington's scale into the current measure, which is a 12-item checklist for retrospective chart review. Henisz et al. (1981) expanded on the results of Flynn and Henisz's preliminary study by correlating checklist scores with inpatient versus outpatient status. Another study showed significantly higher scores for inpatients versus outpatients and correlated higher scores with significantly greater psychopathology as measured by the Minnesota Multiphasic Personality Inventory (MMPI) (Newmark et al., 1977). Additionally, Costello and associates (1991) examined the valid-

ity of a modified version of Flynn and Henisz' checklist for use with children. In the pediatric study, 389 children were clinically evaluated and recommended for hospitalization versus non-hospital treatment. Immediately following the evaluation and recommendation, a psychiatrist completed the checklist. In 95% of cases, the checklist correctly predicted hospitalization using a cutoff score of five. In the children's study, the validation data were obtained from the same patients as were used to derive the cutoff score.

On the other hand, little information is available about psychometric properties of the Flynn and Henisz measure within the population of hospitalized patients. Specifically, little or no data on the inter-rater reliability of the Flynn and Henisz severity measure have been published, with one exception from our own research. When two raters independently used this checklist to assess the charts of psychiatric inpatients, the instrument had an acceptable intra-class correlation coefficient (a measure of inter-rater reliability) for the overall severity score (McFarland et al., 2002). However, it remained to be seen whether or not adequate reliability would be found when the instrument was used with heterogeneous groups of psychiatric inpatients.

The present study was designed to assess the inter-rater reliability of the Flynn and Henisz instrument in diverse groups and to examine the validity of the measure for evaluating severity of psychiatric illness within the overall population of hospitalized psychiatric patients.

Materials and methods

Subjects

We reviewed the records of two groups of psychiatric inpatients in Portland, Oregon. Civilly committed subjects ($n = 150$) had received court assignment to psychiatric inpatient programs for individuals who had been found dangerous to themselves or others because of mental illness. Civilly committed subjects were a 10% random sample selected from a list of individuals who were involuntarily admitted for psychiatric treatment in Portland hospitals and who had an index hospital discharge between 1 January 1995 and 31 December 1999. Judges had committed all civilly committed subjects within 2 weeks of the index admission. The general hospital group members ($n = 248$) were all Medicaid clients with voluntary or involuntary psy-

chiatric admission during 1994 or 1996 at a teaching hospital in Portland, Oregon. Inpatient records from both the civilly committed and general hospital groups were abstracted to obtain information such as demographics, legal status, diagnoses, and length of stay. Table 1 shows demographic and clinical information for both groups of subjects. The project was approved by the Oregon Health and Science University Institutional Review Board.

Clinical assessment and statistical analysis

Two pairs of raters reviewed the records of the civilly committed and general hospital subjects, respectively. The four raters were second year medical students who were trained on four hospital discharge records that were rated in conjunction with two board-certified psychiatrists. In addition, the raters collaboratively evaluated a dozen records and reached consensus on the ratings under the supervision of a board certified psychiatrist.

For both groups, the raters graded each subject's severity on domains such as suicidality, homicidality, and aggression using the 12-item Flynn and Henisz instrument (shown in Table 2). Each item was scored from 0 to 3, according to patient severity. Each item's score was then weighted from 1 through 4. The weights were specified by Flynn and Henisz (1975) and represent their opinions about the relative clinical importance of each item. A weighted average then produced the overall severity rating. According to Flynn and Henisz (1975), patients with scores of 12 or more are said to need hospitalization.

The abstractors randomly selected 20 charts from the general hospital patients and an additional 20 charts from the civilly committed subjects. Two abstractors then rated the same 20 general hospital charts independently. Another two abstractors rated the twenty civilly committed charts independently.

Inter-rater reliability was calculated as the intra-class correlation coefficient. Two approaches were used to compute the intra-class correlation: a one-way and a two-way analysis of variance, respectively, as recommended by Bartko (1991) and Snedecor and Cochran (1989). Searle's (1971) one-way analysis of variance method was used to compute 95% confidence intervals for all intra-class correlation coefficient estimates. In addition, for each item the method of Fleiss et al. (2003) was used to compute weighted Kappa statistics, their standard errors, and 95% confidence intervals.

Diagnoses were obtained from discharge summaries. In general, diagnoses were provided by board certified psychiatrists, according to the fourth edition of the American Psychiatric Association's Diagnostic and Statistical Manual (DSM-IV).

Results

Of the 150 civilly committed subjects' records, 15 did not supply enough information to rate patients' severity of psychiatric illness. These subjects were not included in any analyses (final $n = 135$). Age and race data were missing for two civilly committed subjects for whom all other data were available; these subjects were included in all analyses, but are not reported in descriptions of age and race. All general hospital chart records provided sufficient information to assess psychiatric severity (final $n = 248$). Patient records also provided demographic data for all 248 general hospital patients.

Table 1 shows the demographics of both the civilly committed and the general hospital groups. Overall, both groups of study subjects were middle-aged or younger adults, however, age did differ between the groups ($t[379] = 3.35, p < 0.001$). Gender and race/ethnicity did not differ between the groups. Both groups had severe mental disorders such as schizophrenia (the most common diagnosis), although there were differences in the distribution of primary diagnosis ($\chi^2[6] = 80.47, p < 0.001$). When primary diagnosis categories were examined, there were differences in bivariate analyses between the groups in percent of subjects diagnosed with schizophrenia (28% of general hospital subjects versus 64% of civilly committed subjects, $\chi^2[1] = 50.59, p < 0.001$), substance abuse (20% of general hospital versus 4% of civilly committed, $\chi^2[1] = 17.58, p < 0.001$), and adjustment disorder (13% of general hospital versus 0 civilly committed, $\chi^2[1] = 17.46, p < 0.001$). There were no statistically significant differences in diagnoses of depression (15% of general hospital versus 6% of civilly committed), bipolar disorder (10% of general hospital versus 18% of civilly committed), or 'other' psychiatric illness (14% of general hospital versus 7% of civilly committed).

Figure 1 shows the distribution of severity in civilly committed subjects and general hospital subjects. General hospital subjects' scores ranged from 2 to 45. The mean, weighted average severity score for the general hospital subjects was 14.8 with a standard deviation of 7.7. Civilly committed subjects' weighted

Table 1. Demographic and clinical comparisons of subject groups

N	General hospital 248		Civily committed 135	
Gender				
Female	136	(55%)	54	(40%)
Male	112	(45%)	81	(60%)
Age (SD) ^a	36.8	(10.5)	41.1	(14.7)
Race/ethnicity				
White	205	(83%)	111	(82%)
American Indian	3	(1%)	6	(4%)
Asian American	12	(5%)	2	(2%)
African American	19	(8%)	16	(12%)
Hispanic	7	(3%)	0	
Other	2	(1%)	0	
Primary diagnosis ^b				
Schizophrenia ^c	69	(28%)	87	(64%)
Substance-use disorder ^d	50	(20%)	5	(4%)
Major depressive disorder	37	(15%)	8	(6%)
Adjustment disorder ^e	31	(13%)	0	
Bipolar disorder	25	(10%)	24	(18%)
Other	36	(14%)	9	(7%)
Severity (SD) ^f	14.8	(7.7)	41.7	(9.0)
Severity range	2–45		24–62	
Above admission cutoff score of 12	159	(64%)	135	(100%)

^a two-tailed *t* test with 379 degrees of freedom = 3.35, *p* < 0.001

^b χ^2 2 with 6 degrees of freedom = 80.47, *p* < 0.001

^c χ^2 2 with 1 degree of freedom = 50.59, *p* < 0.001

^d χ^2 2 with 1 degree of freedom = 17.58, *p* < 0.001

^e χ^2 2 with 1 degree of freedom = 17.46, *p* < 0.001

^f two-tailed *t* test with 381 degrees of freedom = 30.75, *p* < 0.001; Mann-Whitney *U* = 562.0, *p* < 0.001

total severity scores ranged from 24 to 62. The mean, weighted average severity score for the civily committed subjects was 41.7 with a standard deviation of 9.0. Differences in mean, weighted severity score were substantial ($t[381] = 30.75$, *p* < 0.001). The non-parametric Mann-Whitney test also showed notable differences in severity (*U* = 562.0, *p* < 0.001). Using other terminology, the effect size for the comparison between civily committed and general hospital patients was 3.28. As expected, all civily committed subjects were well above the admission cutoff score of 12 recommended by Flynn and Henisz (1975). Interestingly, only 64% of the general hospital patients scored above the cut-off.

Multiple regression of severity measure scores showed differences between general hospital and civily committed subjects even after adjustment for demographics. Predictor variables that were entered stepwise into the multiple regression equation included the indicator variable (civily committed versus general hospital), age, gender, ethnicity (white versus other), and primary diagnosis (schizophrenia versus other). Of these predictors, the indicator variable (standardized coefficient = -0.862 , $t = -29.24$, *p* < 0.001), primary diagnosis of schizophrenia (standardized coefficient = -0.073 , $t = -2.51$, *p* < 0.05), and gender (standardized coefficient = -0.063 , $t = -2.28$, *p* < 0.05), were statistically significant. Adjustment for schizophrenia diagnosis and

Figure 1. Severity measure weighted total score distributions.

gender had little impact on the difference in severity between general hospital and civilly committed patients (27.3 points of difference after adjustment).

Two-way analysis of variance showed that the intra-class correlation coefficient for the weighted total score was 0.65 for general hospital subjects and 0.63 for civilly committed subjects. Detailed results are shown in Table 2. Most items in the checklist showed good inter-rater reliability. However the intra-class correlation coefficient was low, 0, or even negative for a few items. Specifically, item 6 (antisocial behaviour) had an intra-class correlation coefficient of 0 when the checklist was used to assess general hospital subjects; item 10 (noxious family situation or disordered state) had a negative intra-class correlation coefficient for civilly committed subjects; and item 12 (need for 24-hour observation or treatment of substance dependence) had an intra-class correlation coefficient of 0 for general hospital subjects. These low intra-class correlation coefficients reflected lack of variation in patients' scores (items 6 and 12) or missing information in their charts (item 10).

Table 2 also shows the 95% confidence intervals for the intra-class correlation coefficients calculated according to Searle's (1971) one-way analysis of variance method. Confidence intervals calculated using the approach of Fleiss et al (2003) were similar to those obtained with the Searle (1971) method (data not shown).

Discussion

Although the Flynn and Henisz measure or close relatives thereof have been used in numerous projects, until

now few psychometric data pertaining to inpatient populations have been available. We used Flynn and Henisz's modification of Whittington's checklist to evaluate severity in civilly committed inpatients versus general hospital psychiatric inpatients. As expected, the two-way analysis of variance model usually yielded intra-class correlation coefficients slightly higher than those obtained with the one-way analysis of variance approach for both subject groups. Using two-way analysis of variance, both intra-class correlation coefficient values were gratifyingly close to the usual recommendation of 0.70 (Nunnally and Bernstein, 1994). These results are encouraging, as the raters had no prior psychiatric experience and had minimal training. In addition, overall inter-rater reliability was found to be adequate in two diverse groups of psychiatric inpatients.

On the other hand, some severity items had better intra-class correlations in the general hospital group (items 1, 8, 11, 12), whereas other items had better correlation in the civilly committed group (items 6, 9). Some of these differences are attributable either to missing data or to insufficient variation among patients to attain adequate inter-rater reliability.

The severity measure also had satisfactory validity within the psychiatric inpatient population. As hypothesized, civilly committed subjects had substantially greater psychiatric severity than general hospital subjects even after adjustment for covariates. The effect size for the difference between groups exceeded three and would be described by Kraemer et al. (2003) as 'much larger than typical'. The implication here is that the Flynn and Henisz instrument can detect dif-

Table 2. Inter-rater reliability of the severity measure

Item	Item weight	General hospital		Intra-class correlations		Civilly committed	
		One-way	Two-way	95% confidence interval (Searle's method)	One-way	Two-way	95% confidence interval (Searle's method)
1 Suicidal preoccupation, in fantasy or thoughts of patient?	2	0.81	0.81	0.58, 0.92	0.34	0.54	-0.10, 0.67
2 Suicidal attempts or active preparations to harm self (e.g., buying a gun, etc.)?	4	0.92	0.93	0.82, 0.97	0.82	0.86	0.61, 0.92
3 Threatened to hurt someone else physically? (Limit to verbal threats)	2	1.00	1.00		0.78	0.83	0.53, 0.91
4 Aggressive outbursts occurred toward people?	4	0.90	0.89	0.76, 0.96	0.75	0.74	0.47, 0.89
5 Aggressive outbursts occurred toward animals or objects?	2	1.00	1.00		0.89	0.88	0.74, 0.95
6 Antisocial behavior occurred?	1	0.00	0.00	-0.43, 0.43	0.74	0.76	0.46, 0.89
7 Impairment of reality assessment, judgment, logical thinking, and planning?	1	0.70	0.74	0.39, 0.87	0.42	0.71	-0.01, 0.72
8 Condition deteriorating rapidly or failing to improve despite supportive measures?	1	0.92	0.92	0.81, 0.97	0.28	0.48	-0.16, 0.64
9 Physical or neurological conditions or a psychotic state?	2	0.44	0.44	0.01, 0.73	0.74	0.73	0.46, 0.89
10 Noxious situation exists among patient's family or disordered state creates difficulties for family?	1	-0.15	-0.12	-0.54, 0.30	0.08	0.38	-0.36, 0.49
11 Emotional contacts severely limited or behavior pathologically ingrained?	1	0.55	0.61	0.17, 0.79	0.32	0.30	-0.12, 0.66
12 Evaluation requires the 24-hour observation or treatment of drug or alcohol dependence?	4	0.66	0.70	0.33, 0.85	-0.09	0.00	-0.49, 0.36
Weighted average		0.62	0.65	0.26, 0.83	0.58	0.63	0.20, 0.81

ferences in severity between groups of inpatients who are expected (on clinical grounds) to be dissimilar.

It is interesting that less than two-thirds of the general hospital patients had severity scores above the admission cutoff. This result may reflect use of psychiatric hospitalization to address homelessness (which occurs all too often in the Oregon Medicaid population). As expected, all the civilly committed patients had severity scores well above the admission cutoff point.

On the other hand, there were several limitations to this study. For example, diagnoses were provided by attending psychiatrists and were not standardized. Some subjects also had insufficient data for severity rating.

Nonetheless, the Flynn and Henisz measure has adequate reliability and validity when used as a retrospective gauge of severity for psychiatric inpatients. Unlike many scales of its kind, this instrument can be completed by minimally qualified, non-medical staff after brief training. Additional research will be needed to establish the measure's use as a pre-admission screening instrument (Flynn and Henisz, 1975) or as a post-treatment assessment of pharmacotherapy outcomes (Shores-Wilson et al., 2002).

Acknowledgements

This study was supported in part by the Stanley Foundation, the Oregon Health and Science University Medical Research Foundation, National Institutes of Health Short Term Research Training: Students in Health Professions grant number NIHT35HL07890, the National Institute on Alcohol Abuse and Alcoholism grant number K02 AA00281. Suzann Khorramzadeh, Rebecca Millius, and D Ann Penalosa MD provided invaluable assistance.

References

- Bartko JJ. Measurement and reliability: statistical thinking considerations. *Schizophrenia Bulletin* 1991; 17: 483–9.
- Costello AJ, Dulcan MK, Kalas R. A checklist of hospitalization criteria for use with children. *Hospital and Community Psychiatry* 1991; 42: 823–8.
- Durbin J, Goering P, Pink G, Murray M. Classifying psychiatric inpatients: seeking better measures. *Medical Care* 1999; 37: 415–23.
- Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3 edn. New York: Wiley Interscience, 2003.
- Flynn HR, Henisz JE. Criteria for psychiatric hospitalization: experience with a checklist for chart review. *American Journal of Psychiatry* 1975; 132: 847–50.
- Glazer WM, Gray GV. Psychometric properties of a decision-support tool for the era of managed care. *Journal of Mental Health Administration* 1996; 23: 226–33.
- Henisz JE, Etkin K, Levine MS. Criteria for psychiatric hospitalization: a checklist approach. *Behavioral Research Methods and Instruments* 1981; 13: 629–36.
- Kraemer HC, Morgan GA, Leech NL, Gliner JA, Vaske JJ, Harmon RJ. Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry* 2003; 42: 1524–9.
- Lansing AE, Lyons JS, Martens LC, O'Mahoney MT, Miller SI, Oblosky A. The treatment of dangerous patients in managed care: psychiatric hospital utilization and outcome. *General Hospital Psychiatry* 1997; 19: 112–18.
- Lyons JS, Colletta J, Devens M, Finkel SI. Validity of the Severity of Psychiatric Illness rating scale in a sample of inpatients on a psychogeriatric unit. *International Psychogeriatrics* 1995; 7: 407–16.
- Lyons JS, O'Mahoney MT, Doheny KM, Dworkin LN, Miller SI. The prediction of short-stay psychiatric inpatients. *Administration and Policy in Mental Health* 1995; 23: 17–25.
- Lyons JS, Stutesman J, Neme J, Vessey JT, O'Mahoney MT, Camper HJ. Predicting psychiatric emergency admissions and hospital outcome. *Medical Care* 1997; 35: 792–800.
- McFarland BH, Khorramzadeh S, Milhius R, Mahler J. Psychiatric hospital length of stay for university hospital clients before and after managed care. *Administration and Policy in Mental Health* 2002; 29: 191–9.
- Newmark CS, Gentry L, Goodman M. Validation of criteria for psychiatric hospitalization checklist. *Journal of Clinical Psychology* 1977; 33: 156–9.
- Nunnally JC, Bernstein IH. *Psychometric Theory*. 3 edn. New York: McGraw-Hill, 1994.
- Roy-Byrne P, Russo J, Rabin L, Fuller K, Jaffe C, Ries R, Dagadakis C, Avery D. A brief medical necessity scale for mental disorders: reliability, validity, and clinical utility. *Journal of Behavioral Health Services and Research* 1998; 25: 412–24.
- Searle SR. *Linear Models*. New York: Wiley, 1971.
- Shores-Wilson K, Biggs MM, Miller AL, Carmody TJ, Chiles JA, Rush AJ, Crismon ML, Toprac MG, Witte BP, Webster JC. Itemized clinician ratings versus global ratings of symptom severity in patients with schizophrenia. *International Journal of Methods in Psychiatric Research* 2002; 11: 45–53.
- Snedecor GW, Cochran WG. *Statistical Methods*. 8 edn. Ames, IA: Iowa State University Press, 1989.
- Whittington HG. *Psychiatry in the American Community*. New York: International Universities Press, 1966.

Correspondence: Bentson H. McFarland, Department of Psychiatry CR-139, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, Portland OR 97239, USA.

Telephone (+1) 503 245-6550.

Fax (+1) 503 494-1209.

Email: mcfarlab@ohsu.edu.