The Resident Assessment Instrument–Mental Health (RAI–MH): Inter-Rater Reliability and Convergent Validity

John P. Hirdes, PhD
Trevor F. Smith, PhD
Terry Rabinowitz, MD
Keita Yamauchi, MD, PhD
Edgardo Pérez, MD
Nancy Curtin Telegdi, RN, MA
Peter Prendergast, MD
John N. Morris, PhD
Naoki Ikegami, MD, PhD
Charles D. Phillips, PhD, MPH
Brant E. Fries, PhD
On behalf of the RAI-MH Group

Abstract

An important challenge facing behavioral health services is the lack of good quality, clinically relevant data at the individual level. The article describes a multinational research effort to develop a comprehensive, multidisciplinary mental health assessment system for use with adults in facilities providing acute, long-stay, forensic, and geriatric services. The Resident Assessment Instrument–Mental Health (RAI-MH) comprehensively assesses psychiatric, social, environmental, and medical issues at intake, emphasizing patient functioning. Data from the RAI-MH are intended to support care planning, quality improvement, outcome measurement, and case mix–based payment systems. The article provides the first set of evidence on the reliability and validity of the RAI-MH.

Address correspondence to John P. Hirdes, PhD, Scientific Director, Homewood Research Institute, and Professor, Department of Health Studies and Gerontology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1; e-mail: hirdes@healthy.uwaterloo.ca.

Trevor F. Smith, PhD, is Assistant Professor, Department of Health Studies, University of Waterloo.

Terry Rabinowitz, MD, is Director, Psychiatric Consultation Service, Fletcher Allan Health Care.

Keita Yamauchi, MD, PhD, is Associate Professor, Faculty of Nursing and Medical Care, Keio University.

Edgardo Pérez, MD, is Chief Executive Officer and Chief of Medical Staff, Homewood Health Centre.

Nancy Curtin Telegdi, RN, MA, is Clinical Trainer, Homewood Research Institute.

Peter Prendergast, MD, is Director of Professional Affairs/Psychiatrist-in-Chief, Ministry of Health.

John N. Morris, PhD, is Alfred A. and Gilda Slifka Chair in Social Gerontological Research and Co-Director, HRCA Research and Training Institute, Hebrew Rehabilitation Center for the Aged.

Naoki Ikegami, MD, PhD, is Professor, Department of Health Policy and Management, Keio University School of Medicine.

Charles D. Phillips, PhD, MPH, is Professor, Health Policy and Management, and Director, Health Services Research Program, School of Rural Public Health,Texas A&M University System Health Center.

Brant E. Fries, PhD, is Professor, Institute of Gerontology, University of Michigan, and Chief, Health Systems Research, Ann Arbor VA GRECC.

Journal of Behavioral Health Services & Research, 2002, 29(4), 419–432.

Introduction

Mental health care is changing rapidly, the population is diverse, and resource demands vary between patients. Answering key policy questions depends on the availability of good-quality, individual-level data.

Existing funding systems for psychiatry tend to use models that employ uniform rates allocated on a per patient basis. Consequently, facilities that provide services to more resource-intensive patients are underfunded and those targeting a lighter care population would be relatively more well off. This creates financial disincentives to the admission and retention of patients requiring difficult and expensive care.

Other sectors of the health care system have begun to implement case mix–based systems where payment is driven, at least in part, by the distribution of patient needs. For example, the United States, Iceland, and one Canadian province have begun to use resource utilization groups (RUG-III)[1] to support funding of long-term care facilities such as nursing homes and chronic hospitals. Initial research on case mix systems for psychiatry yielded models with modest levels of explained variance.[2,3] This is a particular problem for episodic models, but more recent research[4] used a per diem model to explain about 33% of the variance in resource utilization in Japanese psychiatry facilities. Hence, there is a growing expectation that a case mix approach to funding inpatient psychiatric could and should be developed. The result would be a more equitable funding system that is driven by the distribution of patient needs rather than by facility or provider characteristics.

As a first step in creating such a system, a comprehensive, standardized assessment instrument is needed. This article describes a new instrument, which can serve multiple purposes beyond classification of patients for resource allocation, the Resident Assessment Instrument–Mental Health (RAI-MH).

Quality and accountability of mental health services also have begun to move to the forefront of the policy agenda. Government and regulatory agencies are interested in mechanisms to monitor the quality of care provided by psychiatric facilities to evaluate, at least in part, the effectiveness of resource allocation. Given the limitations of patient satisfaction surveys, chart audits, and other widely used approaches to obtaining evidence on service quality,[5] other sectors have moved toward the implementation of performance indicators based on the process and outcomes of care.[6] Such an approach may be helpful to mental health quality managers who could use this information to organize priorities for quality management activities. Consumer report cards also have been popularized as a means of providing information to the general public in a way that will increase the sense of choice and empowerment to consumers of health services, including mental health programs, although there have been cautions raised as to their appropriateness.[7]

Evidence-based practice and policy development demand the use of valid and reliable information to support decision making. In psychiatry, evaluation of the cost-effectiveness of interventions is an ongoing concern. The evidence on nonpharmaceutical interventions, for example, is relatively limited with regard to outcomes, costs, and benefits.

A fourth key consideration is the need for a tool to help integrate health information across sectors of the health care system. Patients with mental health problems are increasingly being served in diverse care settings, ranging from psychiatric hospitals (or units in acute hospital) to group homes, community mental health settings, nursing homes, and home care programs. From a patient perspective, the ability to integrate information across sectors can reduce assessment burden and increase the continuity of care. From the perspective of health care organizations, the integration of information across sectors can allow effective communication with other service providers and holds the potential for implementing care plans more responsive to the needs of new patients. It also provides the opportunity to ensure that care plans formulated in one's own setting will have continuity after discharge to other providers. From the perspective of government, the allocation

of health resources is becoming increasingly focused on populations rather than sectors, and this demands the use of comparable data across care settings.

Existing information systems for mental health are plagued by a variety of problems. Data for many key variables of interest to psychiatry tend to be incomplete or absent. For example, a study of community mental health agencies in the greater Toronto area showed that gender was the sole variable gathered by more than 50% of community mental health agencies.[8] Although most psychiatric service providers conduct assessment as part of routine practice, there is little standardization of assessments across settings, and the lack of comparable data remains a problem. Of particular concern is the widespread use of intake assessments developed in-house without any systematic psychometric evaluation. New patients are routinely assessed using multiple, internally developed forms that have never been evaluated for reliability or validity. Finally, mental health service providers tend to gather information that is unique to them, and there is no ability to integrate with other care sectors.

Development of the RAI-MH

In December 1996 the Ontario Joint Policy and Planning Committee (JPPC; a partnership of the Ontario Ministry of Health and Long Term Care and the Ontario Hospital Association) and an international research consortium known as inter*RAI*[9] began to collaborate in response to the policy challenges mentioned earlier and develop the RAI-MH. The initial focus of the development effort was to support the creation of a case mix–based funding system for inpatient psychiatry, but the research mandate soon was expanded to include development of an assessment system for care planning, quality improvement, outcome measurement, and case mix. The aim of developing an assessment system that could meet the needs of multiple audiences was to reduce the sense of administrative burden on clinical staff and to increase their buy-in through the inclusion of clinically relevant applications. It also was recognized that the quality of data could be increased by both clear operationalization of the assessment and by the tension among different applications. For example, there may be financial incentives to report clinical characteristics associated with increased resource intensity (eg, risk of self-harm). These would be counterbalanced by the tendency to not exaggerate outcome measures or indicators of process of care that could indicate potential quality problems in benchmarking activities comparing facilities (eg, on their prevalence of aggressive behavior disturbance). Hence, the use of the data for quality improvement could reduce the tendency to game the case mix applications of the data and vice versa. Finally, by serving multiple applications for multiple audiences one can increase the cost-effectiveness of new assessments when redundant data collection activities to serve these purposes are eliminated.

The RAI-MH development project[10,11] is an international effort involving clinicians and researchers from Canada, the United States, the United Kingdom, The Netherlands, Norway, and Japan. A research team based in Ontario led the effort, which was conducted under the auspices of the JPPC Psychiatric Working Group (PWG). The PWG comprises a mixture of mental health stakeholders (including consumer perspectives through a representative of the Psychiatric Patient Advocate Office) and has responsibility for development of policy recommendations to the provincial government regarding the implementation of the RAI-MH in Ontario. inter*RAI* is a consortium of more than 40 researchers and clinicians from 21 countries in North America, Europe, The Pacific Rim, and the Middle East. inter*RAI* has substantial experience in creating comprehensive assessment instruments for nursing homes, home care, acute care, assisted living, palliative care, and rehabilitation settings.[12–14] Its Resident Assessment Instrument 2.0 (RAI 2.0) has been mandated in long-term care settings in the United States, Canada, and Iceland, and other inter*RAI* instruments

are being identified for government-sponsored implementation. inter*RAI* members also have conducted an extensive program of international comparisons based on these efforts.[15] The RAI has been used successfully to study the needs of chronically mentally ill individuals in nursing home settings.[16]

The charge to the RAI-MH team was to create an instrument compatible with previously developed RAI instruments, but designed to meet the unique needs of adults in inpatient settings including acute, long-term, forensic, and geriatric psychiatry patients. Like other RAI instruments,[14] the RAI-MH was designed to include "trigger" items that indicate the presence or imminent risk of problems that affect the patient's ability to function independently. These trigger items are associated with clinical algorithms included in mental health assessment protocols (MHAPs) that flag patients with a potential problem in need of further clinical review. Each MHAP is accompanied by a statement of the purpose of identifying the clinical problem, specifications of trigger algorithms used to flag patients with the potential problem, definitions of the issues of interest, a brief background review of current knowledge related to the problem, and questions that may be asked as part of a more detailed clinical review. Moreover, they also suggest interventions that may be used if the presence of the problem is confirmed. Therefore, the RAI-MH incorporates the philosophy of evidence-based practice through the use of experts to summarize the scientific literature in each problem area and to identify potential responses to be considered. It should be emphasized, however, that the RAI-MH does not provide automated care plans. That is, the RAI-MH aims to organize information to support decision making by clinicians, but it does not replace their clinical judgment.

Table 1 provides a list of the MHAPs included in the RAI-MH. The MHAPs are intended to deal with a broad range of patient needs, strengths, and preferences, with the aim of supporting optimal functioning. While some deal specifically with the impact of psychiatric, medical, and/or functional problems, other MHAPs address psychosocial issues. The concepts of recovery, rehabilitation, and empowerment underlie the guidelines for responses to triggered MHAPs. For example, MHAPs for vocational rehabilitation, support systems, economic status, and discharge resources aim specifically to support the patient functioning as independently as possible in the community. Additional discussion on the development of the MHAPs is provided elsewhere.[11]

Other applications of data from the RAI-MH include outcome measurement (eg, scales related to cognitive performance, depression, anxiety, mania, negative symptoms, addictions, quality of life, disability, and extrapyramidal symptoms), quality improvement (about 35 indicators of the outcomes and process of care), and case mix–based funding (through an algorithm developed in subsequent research).

The content of the RAI-MH instrument was developed through a series of steps, including an extensive series of literature reviews; consultations with front-line clinicians and experts; crosswalks of data elements from other RAI instruments and preexisting mandated administrative forms; expert working group sessions; surveys of front-line staff; debriefing sessions after reliability testing; focus groups; and nursing retreats.[11]

There is an ongoing commitment to psychometric evaluation of the RAI-MH, particularly through cross-national comparisons. As part of the development effort, reliability and validity of the instrument were evaluated to support refinement to the current version, which is now deemed ready for implementation. The article describes the results of the first inter-rater reliability trial performed using a preliminary version of the RAI-MH during the development effort. This version contained the bulk of the items retained in the final Version 1.0 of the RAI-MH, since most modifications after that involved deletion or simplification of items.

Validity is not a singular concept. Therefore, it is important to establish different kinds of validity. The development process established face and content validity through the previously mentioned efforts to obtain feedback from a broad range of stakeholders through a variety of communication

Table 1
List of mental health assessment protocols (MHAPs)

MHAPs related to violence/criminal activity
Violence
Self-harm
Abuse by others
Criminal activity
MHAPs related to self-care
Self-care
MHAPs related to role performance
Social function
Interpersonal conflict
Vocational rehabilitation
MHAPs related to social resources
Support systems
Economic status
MHAPs related to psychiatric oversight
Adherence
Psychotropic drug review
Physical restraints and seclusion
Chemical restraint
Revolving door
Discharge resources
MHAPs related to substance use
Addictive behaviors
MHAPs related to health and functional problems
Nutrition
Dehydration
Polydipsia
Skin and foot conditions
Oral health
Pain
Bladder/bowel functioning
Cognition
Communication disorders
Behavior disturbance
Decisional integrity

methods. The article reports on results intended to illustrate evidence gained regarding the convergent validity of the RAI-MH through the examination of associations among some of the key variables of interest. Previous research with other RAI instruments has established the criterion validity of a number of the included outcome scales, such as the Cognitive Performance Scale (CPS),[17] activities of daily living (ADLs),[18] and the Depression Rating Scale (DRS).[19] Further efforts are currently being pursued to establish the criterion validity of these and other scales in psychiatric settings and to establish the predictive validity of algorithms such as the MHAPs with respect to future outcomes of interest. Therefore, this article is intended to report on the first step of a continuing program of psychometric and substantive research with the RAI-MH in Canada and abroad.

Methods

The version of the RAI-MH tested for inter-rater reliability was the product of 18 months of development resulting in seven incremental draft versions of the instrument (another five iterations were completed to refine the instrument before Version 1.0 was finalized). Independent assessors twice assessed a sample of 261 psychiatric patients in acute, long-term, geriatric, and forensic mental health beds in 14 Ontario hospitals. The sample included a mixture of free-standing psychiatric facilities and psychiatric units attached to acute hospitals. The study was based on a convenience sample rather than a random sample, since representativeness of distributions was not the primary purpose of the reliability study. The aim was to use an approach that would reduce the overall level of burden on staff, since the reliability testing was rather onerous. Nonetheless, the methodology may have resulted in a partial bias toward somewhat easier to manage patients for whom it was possible to get consent readily.

Nurses, social workers, and/or psychiatrists carried out all the assessments. Master's-prepared nurses who were members of the research team trained the assessors. A typical training session took place over a 2-day period and included the following: (1) discussion of the RAI assessment instruments and inter*RAI*'s partnership with the JPPC; (2) item-by-item reviews of the RAI-MH to train staff in the intent, definition, assessment process, and coding of individual items; (3) completion of a practice assessment on a patient familiar to the staff person; (4) group discussion of the practice assessments to resolve any areas of confusion; and (5) preliminary introduction to the use of the MHAPs.

The reliability assessments were completed using the most conservative approach possible in order to replicate the day-to-day experience of the field. That is, the assessments were done completely independently, so that assessors were blind to each other's findings, having been explicitly instructed not to discuss cases until after the trial was completed. Complete assessments were done within 24 hours of each other for acute patients, but within a 7-day time span for long-term geriatric and forensic patients. For the latter cases, it was assumed that the rate of clinical change would be slower than in acute patients, thereby allowing for an interval between the dual assessments that would be less burdensome for clinical staff. Assessors were trained to use a variety of information sources including direct observation of patients; interviews with family, friends, or other formal service providers; chart review; and use of other assessment records. They were instructed to exercise their best clinical judgment in order to record observations based on their evaluation of the most reliable and valid information source.

All assessments were recorded in paper form and sent to the project team for transcribing and analysis. For purposes of this study, clinical findings were not reported back because no prior formal evaluation of the psychometric properties of the RAI-MH had been done. Therefore, it was felt that at that stage the data only should be used for research purposes and not care planning.

Assessors were asked to track the time used to complete the assessment and to fill out a debriefing form on their experience in doing the assessment. This information was used to evaluate how close the instrument was to meeting its target completion time of 60 to 75 minutes. Debriefing forms were used to create further revisions to new versions of the instrument. The average time to complete assessments in the trial was 80 minutes.

Analysis

The reliability of the RAI-MH was evaluated using a number of methods. Individual items were assessed for inter-rater reliability based on weighted kappa coefficients using Fleiss-Cohen weights[20] and percentage agreement. Kappa values of 0.40 reflect acceptable reliability; values of 0.70 reflect excellent reliability. In addition, some subscales were evaluated using Cronbach's alpha to measure internal consistency based on parallel items. The evaluation of validity reported here is based on

Table 2
Sample characteristics for inter-rater reliability study

	Percentage (number)
Gender	
Male	56 (146)
Female	44 (115)
Type of patient	
Acute	38 (99)
Long term	28 (73)
Forensic	19 (50)
Geriatric	12 (31)
Missing	3 (8)
Number of prior lifetime admissions	
0	22 (57)
1–3	28 (73)
4–6	20 (52)
7+	30 (78)

patterns of associations in data that can demonstrate the presence of convergent validity. This article reports on some of the comparisons that illustrate the approach used to evaluate this aspect of validity.

Results

Table 2 provides a basic summary of the characteristics of the patients for whom dual assessments were completed in the reliability study. Among the 261 patients assessed, 56% were male and the average age was 45.7 years (standard deviation [SD] = 17.4 years). The largest proportion of patients came from acute care settings (38%), followed by long term (28%), forensic (19%), and geriatric (12%) psychiatry based on an item reported in the RAI-MH. There also was a mixture of patients with respect to their degree of prior involvement with the mental health system. While 22% had no prior lifetime admissions, 30% had been admitted seven or more times in their lives. Therefore, the reliability sample comprised a mixture of the different types of individuals typically encountered in inpatient psychiatry.

Table 3 gives the average kappa (for binary data) or weighted kappa (for ordinal data) values and the average percentage of agreement between raters for the areas of the draft version of the instrument that were retained for Version 1.0 of the RAI-MH. It should be noted that kappa coefficients may be highly unstable with variables that have a low prevalence rate (eg, the rate of setting fires is below 1% during the observation period used by the RAI-MH). In those cases it may be more relevant to consider the percentage agreement between raters.

Almost all domain areas that were retained for Version 1.0 of the instrument obtained average kappa values in excess of the 0.40 cutoff for acceptable reliability. The only area below that level consisted of the items on delirium, but these were retained based on previous evidence that demonstrated the utility of these items in other settings. Behavior symptom frequency had an average weighted kappa value of 0.44. The stronger items in this set were retained (eg, resisting care and physically abusive behavior had kappas of .61 and .84, respectively), but other items were modified. All behavior items were redesigned to simplify the coding for frequency of occurrence. However, it should be noted that even before redesign there was an average of almost 92% agreement on these items. The items with the highest levels of reliability (based on kappa values) were mental health service history, physician

Table 3
Average inter-rater reliability for items retained* in Resident Assessment Instrument–Mental Health (RAI–MH) version 1.0

Section	Number of items	Average kappa	Average percentage agreement
Income source	7	0.56	91.3
Advance directives	2	0.58	88.0
Residential history	11	0.52	87.5
Mental health service history	5	0.78	74.6
Physician, emergency visits	2	0.70	58.0
Behavior symptom frequency	8	0.44	91.7
Self-injury	4	0.66	83.0
Violence toward others	4	0.56	78.8
Delirium	4	0.39	72.5
Self-care			
ADL performance	7	0.48†	90.5
IADL capacity	6	0.67	72.5
Role functioning	4	0.59	80.0
Vocational rehabilitation	4	0.52	79.5
Social activities/isolation	2	0.45	68.0
Health condition and medical symptoms	15	0.48	84.9
Pain	2	0.54	72.0
Falls	2	0.55	86.5
Traumatic life events	12	0.51	86.7
Abuse by others	2	0.57	86.0
Alcohol and tobacco use	3	0.76	92.3
Substance use	7	0.61	95.0
Addiction history	3	0.59	80.3
Weight change	3	0.49	80.7
Polydipsia	1	0.40	84.0
Oral/dental status	4	0.47	85.3
Disease diagnoses	16	0.71	92.7
Restraints/seclusion	5	0.66	95.7
Treatments after admission	8	0.48	85.8
Types of medications	13	0.75	91.7

ADL, activities of daily living; IADL, instrumental ADLs
*Results reported here include items retained with no revisions as well as items with minor modifications.
†When the sample is restricted to geriatric psychiatry patients ($n = 34$) the average kappa for the ADL items is .83 and the average percentage agreement is 81.1.

or emergency visits, self-injury, role functioning, IADL capacity, alcohol and substance use, disease diagnoses, restraints or seclusion, and types of medications.

One of the most problematic areas in the reliability trial was the section on codes from the *Diagnostic and Statistical Manual* (DSM). There was a substantial degree of incomplete information because psychiatrists, who tended to prefer to wait until discharge, had not yet made psychiatric diagnoses. As a result, on intake Version 1.0 of the RAI-MH requests only general information on provisional diagnosis and leaves the specification of DSM codes until discharge.

Table 4
Internal consistency of selected outcome measures in the RAI–MH

	Number of items	Cronbach's alpha
ADL Long Form	7	0.95
IADL Summary	6	0.92
Depression Rating Scale	7	0.77

ADL, activities of daily living; IADL, instrumental activities of daily living

The ADL items showed interesting differences when patient subtypes are considered. When one examines ADL impairment in the total adult psychiatric population, the average weighted kappa is .48 with an average of 91% agreement. This kappa value is markedly lower than has been reported for these same items in nursing home settings.[21] However, if one considers only geriatric psychiatry patients (where the prevalence of ADL impairment is higher than in other psychiatric patients), the average weighted kappa rises to .83 with 81% agreement in these scores.

Table 4 provides evidence on the internal consistency of selected outcome measures based on Cronbach's alpha. The ADL and instrumental ADL (IADL) scales each yielded alpha scores in excess of 0.90, indicating excellent reliability. The DRS demonstrated acceptable reliability with an alpha score of 0.77. It should be noted that in provincial data on all Ontario chronic hospital patients ($n > 30{,}000$), the DRS also achieves alpha values in excess of 0.90 (results available on request).

Tables 5 to 7 and Figure 1 show results that demonstrate convergent validity for components of the RAI-MH. First, there are clear relationships of age with cognitive impairment and disability (see Table 5). Multiple comparisons using analysis of variance (ANOVA) showed that patients age 65 years and over were significantly more cognitively impaired (demonstrated by higher CPS scores) and more disabled (demonstrated by higher ADL scores) than their younger and middle-aged counterparts who did not differ significantly from each other. The ANOVA F test values were 8.4 ($p < .0001$) and 31.9 ($p < .0001$) for the CPS and ADL comparisons, respectively.

Table 6 demonstrates a clear association between items on suicidality and the DRS. Patients who had suicide attempts in the previous 12 months and those who had suicidal ideation in the last 30 days had depression scores that were significantly higher than those not showing these indicators

Table 5
Relationship of patient age with Cognitive Performance Scale (CPS) and ADL Long Form scores

	CPS		ADL	
Age	***n****	**Mean (SE)**	***n****	**Mean (SE)**
Less than 45	151	0.80 (0.11)	151	0.21 (0.01)
45–64	61	0.92 (0.18)	59	0.34 (0.23)
65 and over	45	1.78 (0.25)	41	4.32 (1.09)

SE, standard error; ADL, activities of daily living

*Missing values for some ADL items reduce the numbers of patients for whom the ADL Long Form could be computed. The analysis of variance (ANOVA) F test values were 8.4 ($p < .0001$) and 31.9 ($p < .0001$) for the CPS and ADL comparisons, respectively.

Table 6
Relationship of the Depression Rating Scale (DRS) with indicators of self-injurious behavior

	Mean DRS (SD)	*n*	*t* test	*p* value
Suicide attempt in last 12 months				
No	3.0 (3.0)	198	6.59	.0001
Yes	5.9 (3.2)	63		
Suicidal ideation in last 30 days				
No	2.7 (3.0)	178	7.54	.0001
Yes	5.8 (3.1)	83		
Threat or danger to self/others reason for admission				
No	3.6 (3.2)	228	1.87	.062
Yes	4.7 (3.6)	33		

SD, standard deviation

of suicidality ($t = 6.59$, $p < .0001$ and $t = 7.54$, $p < .0001$, respectively). However, being admitted as a threat or danger to self or others had a much weaker association with the DRS ($t = 1.87$, $p = .062$). This latter item was taken from the Ontario Ministry of Health and Long Term Care's data collection form previously mandated for mental health. This weak association was probably the result of combining the threat of harm to self with the threat of harm to others. For the final version of the RAI-MH, two separate items are used to represent these different concepts.

Table 7 reports on the relationship between nonadherence to medications and a lifetime history of seven or more admissions. There was a clear tendency ($\chi^2 = 5.81$, $df = 1$, $p = .016$) for those with multiple admissions to adhere to their medication regimens less than 80% of the time (those with multiple admissions had almost twice the rate of nonadherence than other patients). This demonstrates the widely held view that medication compliance and revolving door syndrome are likely to be strongly linked.[22,23]

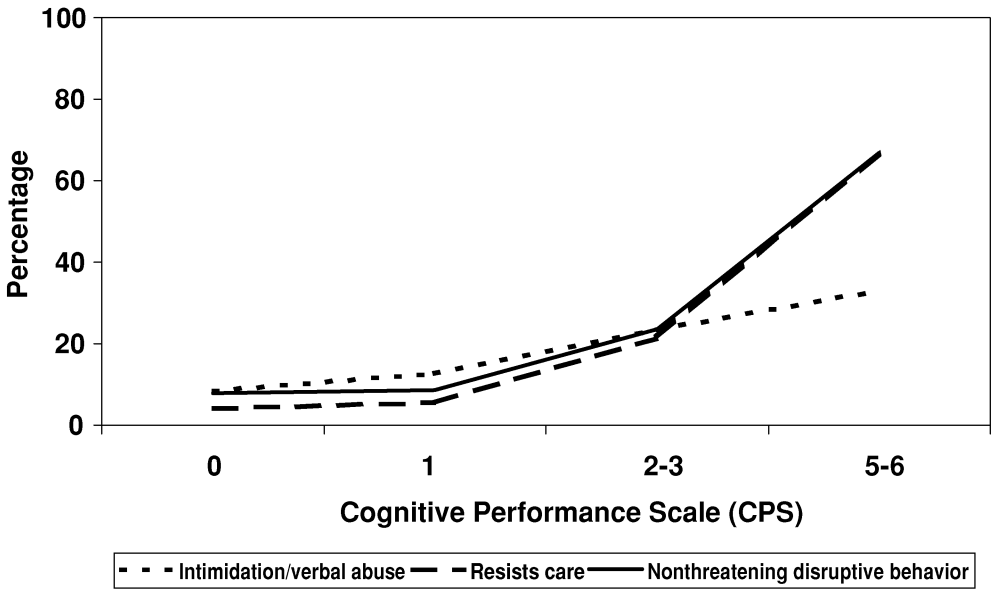
Figure 1 illustrates the relationships of the CPS with three types of behavior disturbance: intimidation/verbal abuse, resisting care, and nonthreatening disruptive behavior. In each instance, a higher level of cognitive impairment was related to a higher prevalence of behavior disturbance. The strongest increase was with resisting care and nonthreatening disruptive behavior, whereas the increase in intimidation/verbal abuse was somewhat less pronounced. The χ^2 values for their relationship with the CPS were 8.2 ($p = .042$) for intimidation/verbal abuse, 27.1 ($p < .0001$) for resisting care and 18.0 ($p < .0001$) for nonthreatening disruptive behavior.

Table 7
Multiple lifetime readmissions and nonadherence to medications prior to current admission

	Adherence to medication regimens less than 80% of the time (*n*)			
	No		Yes	
Seven or more lifetime admissions				
No	84.6%	(165)	15.4%	(30)
Yes	71.2%	(47)	28.8%	(19)

$\chi^2 = 5.81$; $df = 1$; $p = .016$

Figure 1
Prevalence of selected behavior disturbances in Resident Assessment Instrument–Mental Health (RAI–MH) reliability sample



The $\chi 2$ values for their relationship with the Cognitive Performance Scale (CPS) were 8.2 ($p = .042$) for intimidation/verbal abuse, 27.1 ($p < .0001$) for resisting care, and 18.0 ($p < .0001$) for nonthreatening disruptive behavior.

Discussion

In the majority of items retained in Version 1.0 of the RAI-MH, the study results demonstrated acceptable or higher average levels of inter-rater reliability based on kappa coefficients and percentage agreement between raters. Clearly, some domain areas performed better than others. For example, mental health service history, medical diagnoses, physician/emergency services, alcohol/tobacco use, and medication use had average kappa values of 0.70 or more, while measures of delirium, an area known to be difficult to assess even by experts, had an average kappa of 0.39. This is consistent with findings in nursing homes.[24] There was also some variability within domain areas. For example, in the section on potential violence, violence toward others had a kappa of 0.63, while violent ideation scored a kappa value of 0.50. As a consequence, it is not reasonable to state that the RAI-MH *is* or *is not* reliable in global terms. Rather, this issue must be considered on an item-by-item basis.

There are many reasons why tests of inter-rater reliability for specific items may result in low kappa values that suggest potential low reliability. First, any tables with highly skewed distributions will tend to yield kappa values that are more volatile because of low prevalence rates.[25] For example, an item on the alterability of nonthreatening disruptive behaviors had a kappa value of 0.03, but there was 91% agreement between the independent raters. Second, for some items there is a high rate of true clinical change between the times the first and second ratings are completed (eg, fever).

Third, some items involve conditions that are inherently difficult to detect; one would expect a lower level of reliability than for some more obvious conditions. For example, items related to

delirium discussed previously can be expected to have a lower inter-rater reliability level than items related to continence.

Fourth, there may have been a difficulty with misinterpretation of assessment instructions or the instructions themselves may have been unclear. For example, the version of the RAI-MH that was tested included an item to evaluate literacy based on interpretation of a written paragraph. Some nurses allowed patients to keep this paragraph as instructed, while others did not. This misunderstanding of the protocol may have reduced the estimated reliability in this item. In another case, an item on the occurrence of serious accidents or illness in the past 12 months led to confusion because some assessors were uncertain of whether the illness was restricted to physical illnesses or included mental illnesses.

Fifth, some assessment items were poorly designed, with overlapping response categories. For example, some anticonvulsant medications also are used as mood stabilizers; the assessors were sometimes uncertain as to how to code these drugs. Sixth, a few items relied on the presence of an appropriate informant to provide the perspective of family members of patients. These informants may have been available for one, but not the other, rater, thus leading to lower item reliability than might be experienced in the field. Finally, some items included in the draft instrument were simply poorly designed and could not be revised in a way that warranted retaining them in the final version of the instrument.

Once the reliability results were obtained for the draft instrument, the draft RAI-MH was reviewed on an item-by-item basis to determine what solution would be implemented in response to evidence of low reliability. Where the problem lay clearly with the instrument, the solutions used included rewording of items, rewording of instructions, improvement in training protocols, and provision of additional information to clarify the meaning of specific concepts. In some cases, items were deleted outright because it was felt that the item could not be improved and was not central to the purposes of the RAI-MH. In a limited number of other cases, decisions about specific items were deferred to later versions of the instrument where it would be possible to reevaluate results based on a bigger sample size or the availability of data from more clinically stable populations.

Finally, a small number of items with low kappa values were retained because they were felt to be of sufficient clinical importance that they should be kept regardless of weak performance on inter-rater reliability. That being said, the evidence reported in Table 2 suggests that the large bulk of the RAI-MH content performed well in terms of inter-rater reliability. It also should be noted that some items achieved reasonable inter-rater reliability, but were ultimately dropped from the instrument because of problems of acceptance from staff or a sense of creating undue burden. One example was a section on preadmission service utilization patterns. Reasonable reliabilities were achieved, but clinical feedback from front-line staff suggested that these items were viewed as especially burdensome.

The initial analyses also provide reasonable evidence of convergent validity at least for the components of the RAI-MH examined here. This work will continue as part of an ongoing commitment to research on this instrument in Canada and abroad. For at least some key domain areas, the expected patterns of associations between selected variables were found in the studied populations. These results therefore provide the necessary assurance that the aspects of the instrument studied in the convergent validity test performed in the expected manner.

Implications for Behavioral Health Services

There are a number of important developments in the road ahead for the RAI-MH. In Ontario, voluntary implementation of the instrument in inpatient psychiatric settings began in fall 1999, and the Ministry of Health and Long Term Care has confirmed that it expects to mandate use of this instrument by 2003. More tests of reliability and validity are underway, with a particular emphasis on studying criterion and predictive validity. inter*RAI* has begun an international program of research to allow

for cross-national and cross-cultural evaluation of the RAI-MH, beginning with implementation in Spain and studies underway in the United States. Researchers in non–English language countries including Japan, Iceland, Germany, and Sweden have undertaken translation efforts.

Work also has been done to develop and test new applications, including case mix measurement and funding systems and placement decision support (eg, to plan discharges from inpatient psychiatry settings to long-term care facilities or group homes). The US Center for Medicare and Medicaid Services (CMS; formerly the Health Care Financing Administration [HCFA]) has funded a 5-year project (led by Brant E. Fries) to develop a prospective payment system for psychiatric care in the United States. There is a clear opportunity to link the Canadian and US case mix studies in a way that builds on findings from the two jurisdictions. Finally, inter*RAI* has begun work on the development of an outpatient version of the instrument.

Acknowledgments

The RAI-MH is a copyrighted instrument that is owned jointly by the Ontario Ministry of Health, the Ontario Hospital Association, and inter*RAI*. The Ontario Mental Health Foundation, Joint Policy and Planning Committee, Providence Centre Foundation, Homewood Foundation, and inter*RAI* provided funding for RAI-MH development. The work of Drs Hirdes and Smith was supported by a grant from Health Canada. The authors gratefully acknowledge the efforts of Mounir Marhaba, Leah Clyburn, Lori Mitchell, Rita Ann Lemick, and Kim Hallman as well as the clinical staff who participated in study completion.

References

1. Fries BE, Schneider DP, Foley JW, et al. Refining a case-mix measure for nursing homes: resource utilization groups (RUG-III). *Medical Care*. 1994;32:668–685.
2. Fries BE, Nerenz D, Ashcraft M, et al. A classification system for long-staying psychiatric patients. *Medical Care*. 1990;28:311–323.
3. Joint Policy and Planning Committee. *Improving Patient Classification Systems in Psychiatry: A Review*. Toronto, Ontario: The Activity Measurement Working Group of the Hospital Funding Committee of the Joint Policy and Planning Committee; 1995.
4. Yamauchi K. Designing a new payment system for psychiatric care: development of a case-mix classification system. *Journal of the Japanese Society on Hospital Administration*. 1997;34:155–200.
5. Hirdes JP, Zimmerman D, Hallman KG, et al. Use of the MDS quality indicators to assess quality of care in institutional settings. *Canadian Journal for Quality in Health Care.* 1998;14:5–11.
6. Zimmerman DR, Karon SL, Arling G, et al. Development and testing of nursing home quality indicators. *Health Care Financing Review*. 1995;16:107–127.
7. Fries BE, Morris J, Skarupski KA. Facility report cards and the ecological fallacy. *Canadian Journal of Quality in Health Care.* 1998;14:18–22.
8. Community Mental Health Working Group. *Central East Region Mental Health Survey: Report on Survey Findings.* Toronto, Ontario: Canadian Mental Health Association; 1996.
9. Steel K, Jónsson PV, DuPasquier JN, et al. Systems of care for frail older persons. *Transactions of the American Clinical and Climatological Association.* 1999;110:30–37.
10. Hirdes JP, Pérez E, Curtin-Telegdi N, et al. *RAI-Mental Health (RAI-MH)©: Training Manual and Resource Guide, Version 1.0*. Toronto, Ontario: Queen's Printer for Ontario; 1999.
11. Hirdes JP, Marhaba M, Smith TF, et al. Development of the Resident Assessment Instrument–Mental Health (RAI-MH). *Hospital Quarterly.* 2001;4:44–51.
12. Hawes C, Morris J, Phillips C, et al. Development of the nursing home Resident Assessment Instrument in the USA. *Age and Aging.* 1997;26:19–26.
13. Morris JN, Fries BE, Steel K, et al. Comprehensive clinical assessment in community setting: applicability of the MDS-HC. *Journal of the American Geriatrics Society.* 1997;45:1017–1024.
14. Hirdes JP, Fries BE, Morris JN, et al. Integrated health information systems based on the RAI/MDS series of assessment instruments. *Healthcare Management Forum.* 1999;12:30–40.
15. Fries BE, Schroll M, Hawes C, et al. Approaching cross-national comparisons of nursing home residents. *Age and Ageing.* 1996;26:13–16.
16. Phillips C, Spry KM. Chronically mentally ill residents of nursing homes. *Canadian Journal on Aging.* 2000;19(suppl 2):1–17.
17. Morris JN, Fries BE, Mehr DR, et al. MDS Cognitive Performance Scale. *Journals of Gerontology*: *Series A Biological Sciences and Medical Sciences*. 1994;49:M174–M182.
18. Morris J, Fries BE, Morris SA. Scaling ADLs within the MDS. *Journals of Gerontology: Series A Biological Sciences and Medical Sciences.* 1999;54:M546–M553.
19. Burrows AB, Morris JN, Simon SE, et al. Development of an MDS-based Depression Rating Scale for use in nursing homes. *Age and Ageing.* 2000;29:165–172.
20. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use.* 2nd ed. Toronto, Ontario: Oxford University Press; 1995.
21. Morris JN, Nonemaker S, Murphy K, et al. A commitment to change: revision of HCFA's RAI. *Journal of the American Geriatrics Society.* 1997;45(8):1011–1016.
22. Green JH. Frequent rehospitalization and noncompliance with treatment. *Hospital and Community Psychiatry.* 1988;39:963–966.
23. Haywood TW, Kravitz HM, Grossman LS, et al. Predicting the "Revolving Door" phenomenon among patients with schizophrenic, schizoaffective, and affective disorders. *American Journal of Psychiatry.* 1995;152:856–861.
24. Sgadari A, Morris J, Fries BE, et al. Efforts to establish the reliability of the Resident Assessment Instrument. *Age and Ageing.* 1997;26:27–30.
25. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology.* 1993;46:423–429.