RAND

Final Report on Assessment Instruments for a Prospective Payment System

Joan L. Buchanan, Patricia Andres, Stephen M. Haley, Susan M. Paddock, David C. Young, Alan Zaslavsky

Prepared for the
Centers for Medicare and Medicaid Services

RAND Health

The research described in this report was sponsored by the Centers for Medicare and Medicaid Services (formerly the Health Care Financing Administration). The research was conducted through a subcontract from RAND to Harvard University and represents a collaborative effort involving faculty from the department of Health Care Policy at Harvard Medical School, Sargent College of Health and Rehabilitation Sciences at Boston University and RAND Health.

Library of Congress Cataloging-in-Publication Data

Final report on assessment instruments for prospective payment system / Joan L. Buchanan ...
[et al.].
p. cm.
"MR-1501."
Includes bibliographical references.
ISBN 0-8330-3221-6
1. Hospitals—Rehabilitation services—Prospective payment. I. Buchanan, Joan, 1947–

RA971.32 .F564 2002
362.1'786—dc21

2002031821

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

A profile of RAND Health, abstracts of its publications, and ordering information can be found on the RAND Health home page at www.rand.org/health.

Published 2004 by the RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Summary

The Balanced Budget Act of 1997 mandated the implementation of three prospective payment systems for post-acute care providers—one for nursing homes, another for home health agencies, and a third for inpatient rehabilitation facilities. Prospective payment systems pay providers a predetermined fixed price (per day, per episode, or per case) that depends on patient resource needs (often a disease profile or reason for admission) but is independent of the amount of services actually provided. Since the payment is independent of service provision, such systems are thought to create an incentive for efficient, cost conscious care. Although the populations being treated in each post-acute setting have many similarities, the new payment systems have little in common. Each is based on different case mix measures from different assessment tools and, further, each uses different levels of aggregation for payment. The new rehabilitation PPS uses the rehabilitation impairment category (a broad grouping of those admitted for similar rehabilitation reasons), patient age, and functional and cognitive status to classify patients and a single payment is made for the admission. The initial design work for this PPS was based on a functional assessment tool, called the Functional Independence Measure (FIM) and a patient classification system called the Functional Independence Measure-Function Related Groups (FIM-FRGs). RAND researchers refined, completed, and updated that classification work and designed the payment system (see Carter et al., 2002a, 2002b, 2002c). As time passed, policymakers increasingly realized their need for cross setting comparisons of the populations being cared for, the treatments being given, and the outcomes. A new assessment tool, similar to that used in the nursing home industry, the Minimum Data Set—Post-Acute Care (MDS-PAC), was developed to replace the FIM in the rehabilitation PPS. This study was undertaken to evaluate the implications of that substitution.

The MDS-PAC is a comprehensive data collection tool with over 300 items including sections on sociodemographic information, pre-admission history, advance directives, cognitive patterns, communication patterns, mood and behavior patterns, functional status, bladder/bowel management, diagnoses, medical complexities, pain status, oral/nutritional status, procedures/services, functional prognosis, and resources for discharge. Data collectors are instructed to interview the patient and family members and to talk to all caregivers over all shifts for the first 72 hours of care as well as to consult the patient's chart. Functional status assessments allow for one or two exceptions where more care is

needed. The MDS-PAC explicitly recognizes that an activity may not have occurred.

In contrast, the typical FIM form contains a short list of items asking for sociodemographic information, an item asking for the impairment group (reason for the rehabilitation admission) and its underlying etiologic diagnosis, and 18 FIM motor and cognitive items scored at both admission and discharge. The instrument must be scored sometime in the first 72 hours after admission (and within 72 hours before discharge) but is generally scored for the most recent 24-hour period. Scoring on the 18 FIM items is usually evaluated by therapists within their areas of expertise. All items must be scored. Any patient who cannot safely perform an activity is automatically scored as totally dependent.

The planned payment system organizes patients into rehabilitation impairment categories based on the therapeutic reason for admission and then uses the FIM motor scale (sum of the 13 motor item scores), the FIM cognitive scale (sum of the five cognitive item scores), and patient age to classify cases into case mix groups (CMGs) for payment. The age, motor, and cognitive scale values that define each payment cell within a rehabilitation impairment category were defined using classification and regression tree analysis. The CMGs used in this report are available in the Notice of Proposed Rule Making, *Federal Register*, November 3, 2000. These have been further refined and the definitions for the final CMGs can be found in Carter et al. (2002a, 2002b).

To use the MDS-PAC in the new payment system, we needed a method to create a FIM-like motor score and a FIM-like cognitive score. Since the basic FIM concepts were embodied in both instruments, we began with a translation that took several items from the MDS-PAC and converted them into 18 FIM-like items. By summing the 13 "pseudo-FIM" motor items from the MDS-PAC, a motor scale was created. Similarly, the five pseudo-FIM cognitive items were created and summed to form a cognitive scale.

The goal of this project was to compare two instruments, the MDS-PAC and the FIM, to provide insight into whether the planned substitution of the MDS-PAC for the FIM in the proposed inpatient rehabilitation hospital prospective payment system would adversely affect system performance, patients, or hospitals.

Study Design and Implementation

The study design called for two types of data collection: (1) institutionally based teams of rehabilitation therapists and nurses collected FIM and MDS-PAC data on all Medicare admissions within a 10-week study time frame, and (2) study-employed data collection teams, also nurses and rehabilitation therapists, traveled to each hospital during the 10-week data collection phase to re-score FIM and MDS-PAC data on a subset of patients. The latter were referred to as calibration teams. The data provided by the institutionally based teams were used for our primary analyses that examined how well the translation of the MDS-PAC into FIM-like items worked and the payment comparisons. The data collected by the calibration teams were used to examine scoring reliability and to see if institutions were scoring to the same set of norms.

All FIM-certified institutions were invited to participate in the study. Potential participants were asked to send one or more teams to a two-day training session to learn how to score the MDS-PAC and were told that training costs would be paid by the study. Institutions were told that they would receive $35 per completed case (MDS-PAC and FIM) up to $4,000. Within a week, the study received over 180 volunteer responses. To facilitate training and limit calibration team travel, all responding facilities were mapped and hospitals in geographic clusters were linked to together. We then created an expected caseload for each cluster using data on the number of Medicare admissions reported during the previous month for each facility in the cluster. This process allowed us to select clusters that geographically spanned the country and had adequate caseload. Consequently, we were able to manage the travel and workload scheduling for the calibration teams and to manage the training of institutionally based data collectors. Six broad regions were selected with 53 hospitals. Three of the selected hospitals could not meet our schedule and were dropped from the study.

FIM and MDS-PAC data were collected on over 3,200 Medicare cases on hand-written forms from the 50 participating rehabilitation units and hospitals. The facilities ranged in size from 13 to 150 beds. Sixteen percent of rehabilitation hospitals were rural and 28 percent were freestanding facilities. Data collectors were teams of clinicians (physical therapists, occupational therapists, speech language pathologists, and nurses) from each site who attended a two-day MDS-PAC training session and successfully completed a certification exam before the start of the study.

Three calibration teams re-rated over 200 of these cases using both the MDS-PAC and the FIM giving us estimates of inter-team scoring reliability. The calibration teams each included a nurse and two therapists at the beginning of the study. Two nurses were lost to the study early in the data collection phase. Before beginning data collection, the calibration teams were formally trained and certified on both the FIM and the MDS-PAC. Then they spent three weeks working intensively together in four rehabilitation hospitals in the greater Boston area. During the 10-week data collection phase, one or more calibration teams visited all study hospitals re-scoring three to eight cases in each hospital.

Study Findings

Translating the MDS-PAC into FIM-Like Items

To classify patients into case mix groups for payment using the MDS-PAC, we needed to create motor and cognitive scales similar to those in the FIM. The FIM motor scale includes 13 items that cover self-care (eating, bathing, grooming, dressing, and toileting), mobility (transfers, locomotion, and stairs), and sphincter control. The FIM cognitive scale has five items (comprehension, expression, social interaction, problem solving, and memory). Each item in these scales is scored from 1 = total assistance to 7 = complete independence.

Like the FIM, the MDS-PAC also includes functional status items covering self-care, mobility, and sphincter control. In the MDS-PAC, these are scored in reverse order with 0 = complete independence and 6 = total assistance. The MDS-PAC uses two questions for each item; one to cover patient self-performance and the other to indicate the level of assistance provided by others. In the FIM, these concepts are combined into a single rating. The MDS-PAC does not have items with obvious parallels to the FIM cognitive items. For the FIM cognitive scale, we used an empirically derived translations of MDS-PAC items into the pseudo-FIM cognitive items that were developed by Dr. John Morris. For the FIM motor scale, we revised his proposed translation of items.

The revised motor scale translation (1) re-aligned the response category mappings often by incorporating information from other parts of the MDS-PAC, (2) incorporated physical assistance more completely into the scoring, and (3) substituted items where this improved performance. Specifically, the revised translation tried to distinguish the concept of modified independence from total independence (the top two categories in the FIM scoring), collapsed setup and supervision into the next level, incorporated the physical assistance items, and tried to correct several other item-specific scoring inconsistencies. The revised translation also substituted the “walk in facility” for the “locomotion” item, since FIM instructions indicate that the locomotion item should be scored for current capability but uses the mode of locomotion expected at discharge and over 85 percent of cases walk at discharge.

Although relatively short, the FIM actually has a fairly complex set of scoring rules, some of which differed explicitly from those in the PAC, and others merely could not be replicated. Among the more obvious differences are (1) the difference in the assessment periods—the MDS-PAC looks back at the first three days after admission and the FIM looks back over 24 hours any time during the first three days; (2) for patients who appear to be independent, the absence of information on the MDS-PAC about whether the task is completed safely and in a reasonable amount of time; (3) the absence of information in the MDS-PAC on one person assistance with the torso or multiple limbs; (4) different definitions of the need for total assistance; and (5) differences in the task definitions and the treatment of medication use for bowel and bladder management.

Evaluating the Translation

We used factor analysis to assess whether the revised translation improved the conceptual agreement between the pseudo-FIM and FIM concepts and found that, in fact, it did. Neither the raw items nor those from the original translation loaded onto the same factors as the corresponding FIM items, but items from the revised translation did.

The revised translation reduced the mean difference in motor scores between the FIM and the MDS-PAC by 50 percent from the original Morris translation. Despite the improvement, we found that the agreement between the instruments for institutionally based scoring teams (as measured by weighted kappa statistics) was only moderate. Absolute agreement (as assessed by simple kappas) was worse, ranging from poor to moderate. However, when the calibration teams scored patients using both instruments, we found notably higher levels of agreement.

We anticipated that differences in the assessment periods between the instruments contributed to the mean difference in motor scores and found, in fact, that they did. Patients whose motor exams were completed on days 1 and 2 had significantly larger differences than those completed on day 3, with day 2 showing the largest difference. Other factors that influenced the difference were the size of the team scoring the MDS-PAC (three-person teams had smaller differences than one-person teams and those with four or more persons after controlling for other variables) and whether the patient was in for lower extremity joint replacement (RIC 8). After controlling explicitly for the variables that we could, we found that a random effect for hospitals was highly significant. The latter implies that hospitals were systematic in their scoring differences and this was not explained by any of the independent variables. This suggests that more training is needed to adequately standardize the assessment process.

Scoring Reliability

Some of the translation difficulties could be attributable to poor scoring reliability within one or both instruments. A well-designed instrument should yield the same or nearly the same scores for a given patient when administered by different teams or individuals. To assess the reliability of the FIM and the PAC, we compared data re-scored by the calibration teams with that collected by the institutional teams.

When we looked at the impairment group item that was the same on both instruments, we found high levels of disagreement between the institutional teams and the calibration teams. We did not compare the impairment groups directly, but rather we employed a weaker test, comparing the RICs that they mapped into and found that 27–29 percent of the time they were invalid or mapped into different RICs. This finding indicated that additional rules or instructions governing RIC selection were needed for both instruments.

When we compared the scoring reliabilities on the FIM and pseudo-FIM items from the FIM and the MDS-PAC, we found that for the motor items, the FIM had modestly higher kappas and levels of absolute agreement than the PAC. However, regardless of which instrument was used, scoring reliabilities on the weighted kappas were generally only moderate (simple kappas showed poor agreement on 8 out of 18 FIM items and 14 out of 18 MDS-PAC items), a concern for measures intended for use in a payment system. Further, our reliability measures for the FIM motor scale, the cognitive scale, and 11 of 13 motor items were less than those reported in a meta analysis of 11 studies in the literature (see Ottenbacher et al., 1996). The inter-team scoring reliabilities in this study fell below the mean, median, and lower confidence limits on the means that they reported for the motor scale, the cognitive scale, and 11 of the 13 motor items. For three of the five cognitive items, our inter-team scoring reliabilities fell between the reported means and medians. For two of the 13 motor items and two of the five cognitive items, our inter-team reliabilities exceeded those reported in the meta analysis. The meta analysis does not provide information on how actual FIM assessments were performed in the 11 studies. Our calibration teams were observers and information gatherers who did not actually do any physical assessment. At times, they were trying to gather information that was as much as three days old. These procedural differences may have contributed to lower scoring reliabilities. However, one could also argue that their greater dependence on information from treating clinicians makes their individual judgment less important and should have increased agreement.

Patient Classification Agreement and Implications for Payment

Next, we mapped each case into a CMG first using the FIM motor and cognitive scale scores and then using the pseudo-FIM motor and cognitive scale scores. The FIM scales and the pseudo-FIM scales from the MDS-PAC mapped into the same CMG 53 percent of the time. Several different approaches to improve the match between the mappings were subsequently tried. Ultimately, the best effort improved the level of agreement to 57 percent by using a regression mapping of pseudo-FIM items onto the FIM scores and by dropping one facility. The facility that we dropped had a mean difference in motor scores between the two instruments of 14 points (compared to an overall mean difference of 2.4). Further, that facility's team was only team to initially fail our certification exam.

To help understand whether agreement was better for some types of cases, we looked at agreement by RIC, the first tier within the payment system. CMG agreement within RICs was best for a few small RICs (which have only a few payment cells), and it was generally much lower among the larger RICs. Although this level of CMG agreement between instruments (53 to 57 percent) is low for use in a payment system, we found that scoring error within an instrument was high and led to equally poor levels of agreement, 50 percent for the FIM and 55 percent for the MDS-PAC (when the CMGs that result from calibration team responses are compared to institutional team responses on the same instrument).

Despite the poor levels of classification agreement, mean payment differences between the two instruments were small, averaging –$46, and not significantly different from zero. At the facility level, mean per case differences increased somewhat to $82. Despite good overall agreement, we found that more than 20 percent of the facilities would experience revenue differences of 10 percent or more. This remained true when we restricted our sample to hospitals with at least 50 cases. Our multivariate analysis of payment differences showed significant differences across hospitals but these were not systematically associated with patient or hospital characteristics.

Administrative Burden

By far the biggest difference between the instruments was their length. An important limitation of this study was that we did not examine the benefits of the expanded conceptual base provided by the MDS-PAC. We did, however, look at the costs in terms of the administrative burden.

Not unexpectedly, the administrative burden of the MDS-PAC overall was greater than that of the FIM. The magnitude of the difference was large, 147 minutes on average for institutional teams to complete the MDS-PAC compared to 25 minutes to complete the FIM, a sixfold difference. We found a clear learning curve effect during the study (average completion time for the first two weeks of the study of 184 minutes fell to 120 minutes for weeks 7 and 8), which could continue to reduce times beyond those reported here. The size of the data collection team also influenced data completion times significantly; the larger the team the longer the time. By the end of the study, one-person teams had times that were consistent with those reported in the November 3, 2000, Notice of Proposed Rule Making (85–90 minutes). Administration took longer for patients with lower motor function and for those with poor ability to communicate. Urban hospitals had lower times and there was notable variation across regions. The latter may be reflecting facility level differences that we did not control for.

In summary, our study's most important findings are (1) scoring reliabilities, while generally higher on the FIM than the PAC, were not as high as we would hope to see in an instrument intended for payment; (2) the best translation and mappings of the MDS-PAC into CMGs (created from FIM data) agreed with the FIM only 53–57 percent of the time; (3) despite this poor agreement, overall payment differences between the instruments were small; (4) however, 20 percent of the hospitals could see revenue differences of 10 percent or more depending upon which instrument is used; (5) all our multivariate analyses show strong random effects for hospitals with few other significant variables suggesting that additional training could help standardize responses and remove hospital-specific differences; and (6) the administrative burden associated with the MDS-PAC, 120 minutes compared to 23 minutes for the FIM at the end of the study, was substantial.

Instrument Specific Study Recommendations

If the MDS-PAC is selected as the basis of the instrument and the CMGs developed from the FIM are used, then we recommend the following:

- Add the list of impairment codes to the form and improve the guidance given for selecting the proper impairment code.
- Consider adding a scoring category between maximal assistance and total dependence that captures patients completing less than 25 percent of subtasks or change the definition of total dependence.

- Change or supplement the ADL Assist Codes—either add one-person torso and multiple limb or change one limb weight-bearing to one person.
- Revise the scoring to capture the distinction between independence and modified independence and collapse the setup and supervision categories.
- Identify wheelchair-dependent cases.
- Drop Metamucil® from the medication list.
- Continue to use medications to help distinguish complete independence from modified independence but drop medications from the appliance support list.
- Develop additional training materials to further standardize scoring.

In addition, the heavy administrative burden associated with this instrument is of concern. This suggests limiting the number of administrations and possibly limiting implementation to only those items that are relevant for rehabilitation. Items that are currently included on the MDS-PAC so that patient comparability across settings can be assessed might be deferred until the instrument is introduced in multiple settings.

If the FIM is selected, then we recommend enhancing the instrument by making explicit items that are implicitly being evaluated in the FIM scoring process. FIM scoring is deceptively complex and this should improve inter-rater reliabilities. For example, persons were misscored more than half the time when they were independent in eating but had chewing problems and/or swallowing problems that led to the use of modified diets. Similarly, in the locomotion item, FIM scores were not consistent with walking distances explicitly reported in the PAC. Thus, for the FIM, we would recommend the following:

- Standardize the assessment period.
- Add the list of impairment codes to the form and improve the guidance given for selecting the proper impairment code.
- Add explicit scoring aides to improve reliability including
  - — Distance walked or traveled in a wheelchair,
  - — Diet modification and chewing problems, and
  - — Instructions to score locomotion item using expected mode at discharge.
- Separate and record both bowel continence and bowel management assistance.
- Separate and record both bladder continence and bladder management assistance.

- When scoring items such as transfer tub/shower where options are not equivalent, specify rules for which option is to be used and then record which option is being used.

Finally, we suggest that if this option is selected, consideration be given to creating a flexible add/drop section that allows for experimentation and the introduction of new items in the future.

Postscript

Policymakers elected to use a FIM-like instrument called the Patient Assessment Instrument (PAI). Study recommendations for instrument refinement, additional training, and scoring guidance were followed. A section for possible additional items has been added to the PAI and additional research is under way to evaluate the content and format of additional items.