An Individualized Approach to Outcome Measurement in Geriatric Rehabilitation

Paul Stolee,¹,²,³ Karen Stadnyk,³ Anita M. Myers,⁴ and Kenneth Rockwood³

¹Southwestern Ontario Regional Geriatric Program, Canada.
²Department of Epidemiology and Biostatistics, University of Western Ontario, Canada.
³Division of Geriatric Medicine, Dalhousie University, Halifax, Nova Scotia, Canada.
⁴Department of Health Studies and Gerontology, University of Waterloo, Canada.

Background. The heterogeneity of health problems experienced by frail elderly patients makes it difficult to use a single standard measure to evaluate multiple outcomes of geriatric rehabilitation. Commonly, several measures are used, but an alternative is to use an individualized measure such as Goal Attainment Scaling (GAS). This study investigated the reliability, validity, and responsiveness of GAS as an outcome measure in geriatric rehabilitation.

Methods. We studied 173 consecutive admissions (mean age 81; 77% female; mean length of stay 33 days) to a geriatric rehabilitation unit. Assessment instruments were completed at admission and discharge. Individualized treatment goals were identified for each patient by using GAS; standardized measures included self-rated health, a global clinical assessment, the Barthel Index, the OARS IADL scale, the Folstein Mini-Mental State Examination (MMSE), and the Nottingham Health Profile (NHP).

Results. Mobility, future care arrangements, and functional impairment were the most commonly identified GAS goal areas. The interrater reliability of the GAS discharge score was 0.93. The GAS discharge score correlated strongly ($r ≥ 0.50$) with the standardized measures, except for self-rated health, the MMSE, and the NHP ($r ≥ 0.31$). GAS was more responsive to change than any of the standardized measures. The GAS score was used to derive receiver operating characteristic curves for other measures; this can provide insight into the interpretation of clinically important outcomes.

Conclusions. GAS appears to be a feasible, reliable, valid, and responsive approach to outcome measurement in geriatric rehabilitation.

The complex, multidimensional nature of health problems in frail older adults means that geriatric services that specialize in their care require comprehensive, interdisciplinary approaches (1) directed at a wide range of health goals and outcomes (2). This can present challenges both for clinical work and for the evaluation of geriatric services, especially in the selection of outcome measures. Patient care typically includes many goals, and not all goals will be relevant for every patient. For example, pain control applies only if pain is present, and improvement of function is unlikely to be a goal in palliative care. The complexity of different clinically important goals might be addressed, in part, through the use of a variety of health status measures and outcome indicators, but such an approach can be time consuming and has other practical limitations. For example, for randomized trials, sample size and data analysis considerations can limit the number of outcome measures that will be meaningful.

Many evaluations of geriatric services have shown equivocal results (3). Although a meta-analysis suggested that some types of comprehensive geriatric assessment (CGA) programs can improve survival and function in the elderly (4), not all forms or settings of CGA were effective. Inadequate or inappropriate outcome measures may contribute to these results (5,6).

Strong arguments have been made for patient-focused approaches that reflect the opinions, preferences, and concerns of individual patients (7,8). An individualized approach may be particularly relevant for frail older adults. Williams has noted that, given the immense individual differences among older people, "...when we consider what quality of life means to an older person and what features of quality of care may contribute to that quality of life, we must arrive at highly individualized conclusions" (9).

Goal Attainment Scaling (GAS) may have merit as a responsive and clinically meaningful individualized measure for use with older people. GAS is a measurement approach developed in the 1960s for use in evaluating human service and mental health programs (10). It accommodates multiple, individualized patient goals, yet it also provides a numerical score allowing for comparisons across patients. GAS allows for relative degrees of goal attainment to be captured, on the basis of goals relevant to individual clients.

Although its individualized nature and clinical relevance give GAS face validity, concerns have been raised about the psychometric properties of GAS (11). Preliminary investigations in geriatric care settings suggest that GAS provides clinically relevant information and has a good interrater reliability (intraclass correlation coefficients of 0.87 to 0.91) (5,12). These studies have also begun to examine how GAS correlates with other measures commonly used in geriatric care. In a small study of 15 patients, GAS change scores correlated strongly with both a
global clinical rating ($r = 0.82$) and with change in the Barthel Index ($r = 0.86$) (12). On one hand, the relationship with the Barthel Index seemed reasonable given that all of these patients had mobility or functional impairments. On the other hand, this high correlation called into question the merits of the extra work required to perform GAS. In a second study of 45 patients with a broader cross section of health problems (5), a lower correlation with the Barthel Index was found ($r = 0.59$), and correlations with other standard measures also suggested that GAS may be capturing additional information on improvement in patient outcomes. Results of the second study also suggested that GAS may be more responsive to clinically important change compared with other commonly used standard measures (5).

Although these preliminary investigations of GAS in geriatric care have been promising, a more thorough investigation of GAS with a larger group of geriatric patients was warranted. This study investigated the reliability, validity, and responsiveness of GAS as an outcome measure in geriatric rehabilitation. In addition, we used the GAS data to describe and quantify clinically important outcomes in frail older adults.

**METHODS**

**Design and Subjects**

This prospective, descriptive study was carried out in a 16-bed geriatric rehabilitation unit at a tertiary care facility in Halifax, Nova Scotia, Canada. Subjects were 173 consecutively admitted elderly patients (mean age, 81 ± 7; range, 61–96; 77% female; mean length of stay 33 days). Data were collected over a period of 21 months. The study protocol was approved by the institutional ethics review board. Assessment measures such as those included in this study are routinely given to all patients admitted to the unit, and informed consent was not required.

**Goal Attainment Scaling**

In the GAS approach (10), patients are assessed and individualized goals are identified on a GAS follow-up guide. On the geriatric rehabilitation unit, patients were assessed by members of an interdisciplinary team (including a geriatrician, physiotherapist, occupational therapist, social worker, dietitian, and nursing staff), who then developed the GAS follow-up guides on a consensus basis during the team’s regular case conference. Although patients, their family members, and other caregivers were not formally involved in establishing the follow-up guides, their concerns, preferences, and expectations were identified in the discipline-specific assessments and taken into account in goal setting. An example follow-up guide is presented in Figure 1. The example shows rehabilitation goals for a 78-year-old woman admitted to a geriatric rehabilitation unit following a hip fracture (several details have been changed to protect anonymity). The individualized goals are scaled based on the expected outcome (the 0 category) and plausible outcomes that are better (+1 or +2) or worse (−1 or −2) than expected. These outcomes should be as objectively verifiable as possible. The goals may be weighted to express relative clinical importance, although this step is optional and common practice is to use equally weighted goals.

At the end of the follow-up period, the status of the patient is assessed against the goal attainment levels and an overall goal attainment score is calculated as follows (10):

$$\text{Goal Attainment Score} = 50 + \frac{10 \sum (w_i x_i)}{[.7 \sum w_i^2 + .3 \sum w_i]}$$

where $w_i$ is the weight assigned to the $i$th goal, and $x_i$ is the numerical value (−2 to +2) of the attainment level of the $i$th goal. This score is an average goal attainment score adjusted for the relative weighting assigned to the goals, the number of goals, and the expected intercorrelation among the goal scales. The intercorrelation is represented in the formula by the generally assumed value of .3 for $p$ (.7 = 1 − $p$). As goals in this study were unweighted, $w_i = 1$. The baseline levels for each of the GAS goals are included as scale descriptors on each of the individualized goal scales (usually at the −1 or −2 levels). A baseline GAS score can thus be calculated by using these numerical values in the GAS equation. The follow-up period may vary for each individual goal. For example, it may be expected that one patient’s mobility goal will take 5 weeks to accomplish, whereas a goal in activities of daily living will take 3 weeks. GAS can accommodate this range of clinical expectations. Often, it is expected that most goals for an individual patient will be attained by an anticipated discharge date.

**Assessment Measures**

Assessment measures included two standardized measures of physical functioning completed by an occupational therapist or physiotherapist. These were the Barthel Index (13) as modified by Granger and colleagues (14) and the OARS Index of Instrumental Activities of Daily Living (IADL) (15). The Standardized Folstein Mini-Mental State Examination (MMSE) (16,17), a self-rated health question (For your age, would you say that your health is: excellent, good, fair, poor, very poor?; scaled from 1 to 5), and a quality of life measure, the Nottingham Health Profile (NHP) (18), were completed by a trained research assistant. A geriatrician completed a global clinical assessment (for 61 patients) of the patient’s overall outcome on discharge (1, very poor, to 7, exceeds expectations). Baseline and follow-up assessments were usually completed within 48 hours of admission and discharge, respectively. The interrater reliability of the secondary instruments was not tested.

**Reliability**

To assess interrater reliability, 61 GAS guides were scored on follow-up by the multidisciplinary team in their case conference, and independently by a nurse involved in the patient’s care but not present at the team conference.

**Validity**

Content validity was examined through content analysis of identified goals areas. Construct validity was assessed through correlations with other study measures. GAS should be expected to correlate strongly with standardized measures that cover clinically relevant domains that are similar to the goal areas identified in the GAS follow-up guides. As a test of GAS as an evaluative instrument (19), the GAS change score should correlate strongly with other measures that show change in response to intervention, and less strongly with measures that show little change. Correlations of the discharge or follow-up
scores for GAS and the other measures were used to assess GAS as a discriminative measure (19).

Responsiveness was analyzed by using several methods, including effect size, relative efficiency, and analysis of variance. Two effect-size statistics were calculated; in the first statistic, the difference in before and after mean scores was divided by the baseline standard deviation (20). A second effect-size statistic (the standardized response mean) was calculated by using the standard deviation of the change score as the denominator (21). Effect sizes exceeding 0.80 are large, effect sizes of 0.50 to 0.79 are medium, and those between 0.20 and 0.49 are small (20,22). Relative efficiency is the squared ratio of paired-\( t \) statistics, in which each measure is compared against an arbitrary standard (23). For this study, the Barthel Index was used as the standard. As described by Liang and colleagues (23), the selection of the comparison "standard" is arbitrary and will not affect the interpretation of the results. Nonetheless, the selection of the Barthel Index seemed appropriate, as this is a well-known measure in widespread use in geriatric rehabilitation. Finally, responsiveness was calculated by using a repeated measures analysis of variance (ANOVA) as proposed by Norman (24), in which

\[
\text{responsiveness} = \frac{\text{variance due to change}}{\text{variance due to change} + \text{error variance}}
\]

For each responsiveness method, higher numbers indicate a greater sensitivity to change.

Deyo and Centor (25) have drawn an analogy between the analysis of responsiveness and the assessment of a diagnostic

<table>
<thead>
<tr>
<th>ATTEINMENT LEVELS</th>
<th>GOAL AREAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility</td>
<td>ADLs</td>
</tr>
<tr>
<td>Much less than expected -2</td>
<td>2 person assist with transfers and walking.</td>
</tr>
<tr>
<td>Somewhat less than expected -1</td>
<td>1 person assist with transfers &amp; walking. Walking 10' with rolling walker.</td>
</tr>
<tr>
<td>Expected level (Program Goal) 0</td>
<td>Independent transfers. Independent dressing. Minimal assist bathing.</td>
</tr>
<tr>
<td>Somewhat better than expected +1</td>
<td>Walking with cane short distance (50').</td>
</tr>
<tr>
<td>Much better than expected +2</td>
<td>Walking with cane longer distances (&gt; 50').</td>
</tr>
<tr>
<td>Comments:</td>
<td></td>
</tr>
</tbody>
</table>

 ✓ level on admission.
 * level on discharge.

Figure 1. Sample Goal Attainment Scaling follow-up guide.
test; that is, a measure could be assessed in terms of its sensitivity and specificity in “diagnosing” clinical change. A possible application of receiver operating characteristic (ROC) curves in this analysis is to use the GAS scores as a way to define successful or unsuccessful attainment of change that is clinically important. A GAS score of 50 or better is, by definition, an indication that there has been an overall level of goal attainment at the clinically expected level—if all goals are met at the 0 level, for example, then \( \sum (w_i x_i) = 0 \) and the GAS score = 50 (10). This cut point may seem somewhat arbitrary in that, for example, a patient’s goal attainment that fell just short (a score of 49) would be categorized as “unsuccessful.” This is not unlike cut points applied in other diagnostic tests, however, and there are mathematical and conceptual bases to accept 50 as the dividing line.

Scores from standardized instruments could then be compared with the GAS cut point of <50/50+ as a way to determine the level of improvement that would represent clinically important change. This comparison would also provide additional information on the validity and responsiveness of GAS, in that there would be data from standardized instruments describing clinical improvement that is represented by the successful attainment of GAS goals. ROC curves were constructed for the Barthel Index, OARS IADL, the NHP Physical Mobility subscale, and self-rated health change scores. The sensitivity or true-positive rates are plotted against the false-positive rates for each measure. (Converting specificity to the false-positive rate allows both properties of the measure to be plotted on 0–100 scales.) Methods for construction of ROC curves are described in Sackett, Haynes, and Tugwell (26).

**RESULTS**

There were 170 GAS follow-up guides prepared for study patients (guides for the other three patients in the dataset were incomplete and not usable), and a total of 674 individual goals scales were produced (mean of 4 ± 1.3 goals per patient; range, one to seven). To assess content validity, the GAS goals were grouped into major categories (see Figure 2), of which the most common were mobility, future care, personal care, and bowel and bladder problems. Patients may have had more than one goal in a particular category (for example, mobility could include both ambulation and transferring goals for the same patient). The categorizations were reviewed by clinicians from the geriatric rehabilitation unit, with modifications as appropriate.

The interrater reliability (team conference vs independent nurse assessment; \( N = 61 \)) of the GAS follow-up score was 0.93 (intraclass correlation coefficient). A second analysis used the individual goal scales from the 61 guides as the unit of observation, in place of individual patients. Of 255 individual goal attainment levels (−2 to +2) on follow-up, there was complete agreement on 204 (80%) and an intraclass correlation coefficient of 0.89.

Descriptive and responsiveness statistics are shown in Table 1 for each measure at baseline and follow-up. The ordering of the measures in terms of their responsiveness is to some extent affected by the statistic used; however, each method determined GAS to be the most responsive instrument. Data from the NHP were frequently missing, as many patients found these questions inappropriate for their situation.

Correlations of change and follow-up scores with GAS are shown in Table 2. GAS follow-up scores correlate strongly with

![Figure 2. Content analysis of GAS goals for 170 patients.](https://academic.oup.com/biomedgerontology/article-abstract/54/12/M641/605669)

Downloaded from https://academic.oup.com/biomedgerontology/article-abstract/54/12/M641/605669 by guest on 12 February 2018
Table 1. Descriptive Statistics and Responsiveness of the Study Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Baseline (N, Mean ± SD)</th>
<th>Follow-up (N, Mean ± SD)</th>
<th>Difference (N, Mean ± SD)</th>
<th>t Value (p)</th>
<th>ES*</th>
<th>SRM†</th>
<th>RE‡</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barthel Index (0–100)</td>
<td>50.23 (17.48)</td>
<td>70.09 (23.97)</td>
<td>19.86 (20.39)</td>
<td>1.14 (0.001)</td>
<td>0.97</td>
<td>1.00</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>OARS IADL (0–14)</td>
<td>5.83 (3.12)</td>
<td>10.57 (2.44)</td>
<td>1.70 (1.00)</td>
<td>0.97 (0.001)</td>
<td>0.80</td>
<td>0.69</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>MMSE (0–30)</td>
<td>24.15 (5.20)</td>
<td>23.95 (3.36)</td>
<td>-0.20 (0.474)</td>
<td>0.97 (—)</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>NHP§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy (0–100)</td>
<td>38.48 (36.08)</td>
<td>28.65 (29.90)</td>
<td>9.83 (4.09)</td>
<td>0.27 (0.001)</td>
<td>0.33</td>
<td>0.10</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>Pain (0–100)</td>
<td>20.95 (20.09)</td>
<td>13.73 (20.31)</td>
<td>7.22 (9.00)</td>
<td>0.26 (0.001)</td>
<td>0.36</td>
<td>0.10</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Emotional reactions (0–100)</td>
<td>20.35 (25.08)</td>
<td>14.60 (20.87)</td>
<td>5.75 (16.20)</td>
<td>0.23 (0.001)</td>
<td>0.36</td>
<td>0.11</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Sleep (0–100)</td>
<td>32.81 (32.01)</td>
<td>24.00 (24.80)</td>
<td>8.81 (1.22)</td>
<td>0.28 (0.001)</td>
<td>0.36</td>
<td>0.11</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Social isolation (0–100)</td>
<td>18.84 (25.92)</td>
<td>14.47 (21.66)</td>
<td>4.38 (2.38)</td>
<td>0.17 (0.001)</td>
<td>0.20</td>
<td>0.04</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Physical mobility (0–100)</td>
<td>54.45 (26.11)</td>
<td>44.22 (19.80)</td>
<td>10.24 (5.45)</td>
<td>0.39 (0.001)</td>
<td>0.52</td>
<td>0.18</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Self-Rated Health (1–5)</td>
<td>2.44 (0.93)</td>
<td>2.26 (0.85)</td>
<td>0.18 (0.001)</td>
<td>0.21 (0.001)</td>
<td>0.20</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>GASII (18–82)</td>
<td>47.86 (8.29)</td>
<td>47.86 (8.56)</td>
<td>0.00 (0.001)</td>
<td>0.17 (0.001)</td>
<td>0.13</td>
<td>0.31</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

*Effect size: L, large effect size (≥0.80); M, medium effect size (0.50–0.79); S, small effect size (0.20–0.49); -, negligible effect size.
†Standardized response mean.
‡Relative efficiency.
§For the NHP, lower scores indicate improvement.
‖The range is calculated on the basis of seven goals (the most for any patient), each scored as maximally positive (+2) or maximally negative (–2).

The global clinical rating and with the two measures that cover domains that are most similar to the GAS goals—the Barthel Index and the OARS IADL scale. The Barthel Index and the OARS IADL scale were also the measures found to be the most responsive to change (aside from GAS) and correlated strongly with the GAS change scores.

The ROC curves for four measures are overlaid in Figure 3. For each ROC curve, the point closest to the upper left-hand corner indicates the measure’s cut point that best discriminates successful from unsuccessful goal attainment. This equates to the cut point that maximizes the total sensitivity and specificity of each measure. These points correspond to 11+ points of improvement on the Barthel Index and 3+ points of improvement on the OARS IADL scale. Next to GAS, these measures were the most responsive, and they were the best able to "diagnose" goal attainment. The curves for the NHP Physical Mobility scale and self-rated health approximate diagonal lines from the bottom left-hand to upper right-hand corners, suggesting that these are practically useless tests in this setting for discriminating among patients who have or have not successfully attained their clinical goals (25).

**DISCUSSION**

This study examined the reliability, validity, and responsiveness of GAS, and, having demonstrated suitable measurement properties, attempted to use GAS to derive insights about what constitutes clinically important change in several standard measures. In particular, the interrater reliability of the GAS follow-up score was found to be excellent, and the interrater reliability of ratings of individual goal attainment levels was somewhat lower, but still acceptable (27). These results may be attributable in part to the use of clear, objective, and measurable indicators for the goal scales, which facilitated an independent identification of the patients’ attainment levels. A limitation of
this study is that we have not investigated the interrater reliability of the development of the GAS follow-up guide. This is an area where additional research would be valuable. In a previous study (12) we compared GAS guides developed separately by two geriatricians and found that 82% of GAS goal areas were identified independently. In the current study, however, the GAS follow-up guides were developed on a consensus basis by the interdisciplinary team, based on assessment information and in consultation with patients and their caregivers. The most appropriate reliability test in this context would be to compare results of two independent versions of this process; this was not feasible in our setting. An option would be to have the second GAS guide developed by a single blinded clinician, but this would be a comparison of two different models of clinical decision making.

Overall, the construct validity correlations (of both change scores and follow-up scores) appear reasonable, given the types of patients admitted and their presenting problems (e.g., a high prevalence of functional and mobility impairments). GAS correlates strongly with other measures that show change, and it also discriminates between lower and higher functional or quality of life status. A caveat arises from the surprisingly low correlations with the NHP. This may cast doubt on GAS as a potential quality of life measure, or on the appropriateness of the NHP as a quality of life measure in this context. It may also be an indication that GAS is measuring a distinct construct. The answer may depend on whether the domains covered in the GAS goals can be considered as quality of life goals. In any case, the domains identified in the content analysis appear relevant to those that would be identified in a comprehensive geriatric assessment approach (28) with these patients.

Inasmuch as there is currently no single standard approach to assess responsiveness (29), it seems important to use a variety of methods and to consider the responsiveness of measures in specific clinical contexts. GAS was found to have high responsiveness, however measured, in comparison with standardized measures, suggesting that it may be useful as an evaluative tool for measuring outcomes of geriatric interventions.

The cut points derived from the ROC curve analysis could be considered as the amount of change in each measure that could be considered a clinically important improvement. These could also be related to clinical descriptors. For example, a 15-point improvement on the Barthel Index could represent improvement from dependence to independence in transferring in and out of a chair; a 10-point difference corresponds to gaining independence in bladder control.

These data give an indication of what level of patient improvement, as measured by standardized instruments, is generally experienced when a GAS score of 50 or better is reached, that is, when there is overall goal attainment. This provides information on what the responsiveness of GAS means in clinical terms in this context, and it provides further data on the validity of GAS. The data can also help to alleviate skepticism that the achievement of individualized goals may be clinically trivial.
INDIVIDUALIZED OUTCOME MEASUREMENT

GAS can be a valuable research and clinical tool for use in specialized geriatric services for frail older adults. Future investigations of GAS in geriatric care could assess its generalizability to other geriatric care services, and the potential for formal involvement of patients and caregivers in goal setting.

ACKNOWLEDGMENTS

This research was supported by a grant from Glaxo Canada, by a National Health PhD Fellowship to Paul Stolee, and by a National Health Scholar Award to Kenneth Rockwood.

This paper was presented at the annual Scientific Meeting of The Gerontological Society of America, November, 1997.

Address correspondence to Dr. Paul Stolee, Southwestern Ontario Regional Geriatric Program, c/o St. Joseph’s Health Centre—Parkwood Site, 801 Commissioners Road East, London, Ontario, N6C 5J1, Canada. E-mail: p.stolee@parkwood.london.on.ca

This research was supported by a grant from Glaxo Canada, by a National Health PhD Fellowship to Paul Stolee, and by a National Health Scholar Award to Kenneth Rockwood.

REFERENCES


Received August 31, 1998
Accepted May 31, 1999