American Journal of Epidemiology

Vol. 147, No. 8
Printed in U.S.A.

Short-term Consistency in Self-reported Physical Functioning among Elderly Women

The Women's Health and Aging Study

Paul J. Rathouz,[1] Judith D. Kasper,[2] Scott L. Zeger,[3] Luigi Ferrucci,[4] Karen Bandeen-Roche,[3] Diana L. Miglioretti,[3] and Linda P. Fried[5]

The assessment of physical functioning and disability is integral to population-based and clinical research carried out among elderly people. Typically, functional status is measured through self-reported responses to questions of the form "Do you have difficulty [doing a specific task]?" Knowledge of the reliability and validity of these self-report measures is key to the interpretation of many research efforts, but data on these measurement parameters are sparse. This paper addresses this deficiency through analyses of data from the Weekly Substudy of the Women's Health and Aging Study, a cohort of Baltimore-area women aged ≥65 years with moderate to severe physical disability. Self-reported data on 20 activities, obtained weekly over a 6-month period in 1993 or 1994, were analyzed to investigate how time intervals between assessments and a subject's age and baseline level of disability influenced the consistency of self-reports of disability at both the population level and the individual level. The prevalence of self-reported difficulty increased with baseline disability and, to a lesser extent, with age group. Consistency for all items was very high over short time intervals, but it decreased substantially with increasing intervals between responses (although associations between responses remained significant at 24 weeks). Consistency did not vary with age or baseline disability. Graphic techniques and statistical methods for use with repeated binary data are also illustrated. *Am J Epidemiol* 1998;147:764–73.

activities of daily living; aged; aging; disability evaluation; epidemiologic methods; questionnaires; women

The assessment of physical functioning and disability is an integral component of population-based and clinical research (1), studies of medical effectiveness, and efforts to assess quality of care (2, 3). Despite an important shift from a medical model of disability to a functional model (1), neither the natural history of disability nor the performance over time of commonly used measures of physical functioning is well understood. Several recent studies have indicated that functional status among disabled elderly people both improves and deteriorates when assessed at intervals of 2 years or longer (4, 5). However, it is possible that these changes reflect inconsistency in responses to assessment instruments, or poor validity of the instruments themselves (6). This is of special concern in population-based studies, wherein many instruments for measuring functional status rely on self-reports.

Measuring functional status and disability through the self-reports of individuals is the standard approach in epidemiologic and other population-based studies. While many commonly used indicators of functioning had their origin in observational assessments made by professionals in rehabilitation (7) and work (8) settings, self-report questions have largely supplanted this earlier approach. Self-report measures are easily adapted to the types of questionnaires used in population-based surveys; they do not require the involvement of clinical professionals; and they provide information on the subject's perspective, which has been suggested to have high predictive validity for later morbidity, mortality, and service use (9–12). Not surprisingly, use of these measures has become wide-

Received for publication December 2, 1996, and in final form August 15, 1997.

Abbreviations: ADL, activities of daily living; CI, confidence interval; LOR, log odds ratio.

[1] Department of Health Studies, School of Medicine, University of Chicago, Chicago, IL.
[2] Department of Health Policy and Management, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, MD.
[3] Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, MD.
[4] Geriatrics Department, "I Fraticini," National Research Institute (INRCA), Florence, Italy.
[5] Departments of Medicine and Epidemiology, The Johns Hopkins Medical Institutions, Baltimore, MD.
Reprint requests to Dr. Judith Kasper, Department of Health Policy and Management, Johns Hopkins School of Hygiene and Public Health, 639 North Broadway, Baltimore, MD 21205.

spread, although standardization of questions designed to assess functioning and disability has not been achieved (e.g., having "difficulty" performing activities without assistance (13) versus "needing help" with activities (14)). Self-report measures of physical disability have potential drawbacks, however. It is possible that, over time, subjective judgments change independently of changes in physical function. Unfortunately, very little is known about the magnitude of this variability, since most studies of the course of disability have had relatively long assessment intervals (e.g., from 6 months to 24 months (15)). Mathiowetz and Lair, examining self-reports of limitations in activities of daily living (ADL) given 1 year apart in a large national survey, argued that changes in self-reports of ADL status over time "may have a significant error component" (6, p. 260). In fact, although knowledge of the reliability and validity of self-report measures of functioning is key to the interpretation of many research efforts in older adults, data on these measurement parameters remain sparse.

In this report, using data from the Weekly Substudy of the Women's Health and Aging Study, we address these issues by examining how time intervals between assessments, as well as selected characteristics of individuals, such as age, influence the consistency of self-reports of disability. Using self-reported data on physical functioning collected weekly over a 6-month period, we assessed the validity of self-report measures by searching for trends in the prevalence of reported functional difficulty by age or severity of disability. We hypothesized that the prevalence of reported difficulty over time would be greater among older women and those who reported more severe disability at baseline. We assessed test-retest reliability (16) at the group level by examining systematic time trends in reported difficulty over the 6-month period, and at the single-subject level by examining the "consistency" of self-report measures—i.e., the degree to which a subject's responses in one week were similar to her responses in a later week. In particular, we investigated 1) whether consistency varies by age or baseline disability and, if so, in what way, and 2) how consistency varies by length of the interval between responses. In addition to addressing the issues of validity and reliability, this paper illustrates the use of a statistical method for analysis of repeated binary data.

MATERIALS AND METHODS

Subjects

The Weekly Substudy sample consisted of 108 women drawn from the 1,002 women in the Women's Health and Aging Study, a population-based prospective study of women aged ≥65 years who were moderately to severely disabled and living in the community. The Women's Health and Aging Study sample was based on an age-stratified random sample drawn from Medicare enrollment files that listed all female beneficiaries in 12 contiguous zip codes in Baltimore City and Baltimore County, Maryland (15). A screening interview was used to determine disability in four domains (upper extremity movement, mobility and exercise tolerance, performance of tasks indicative of higher-level functioning, such as shopping or preparing meals, and basic self-care tasks (17)), and women with baseline disability in two or more domains were recruited into the study.

The 108 participants in the Weekly Substudy were selected randomly and in roughly equal numbers from each of nine age × disability level strata (the age groups 65–74, 75–84, and ≥85 years, and disability in two, three, or four domains). Participants were interviewed at home each week for 24 weeks on the same day and at the same time. Half of the women were interviewed between July and December of 1993 and half between January and June of 1994.

Information collected each week included self-reported physical functioning, results from a limited set of performance-based functional measures (18), and any health-related events that had occurred during the prior week. Self-reports of difficulty with physical activities or tasks were obtained via questions such as, "By yourself—that is, without help from another person or special equipment—do you have any difficulty walking across a small room?" and, if the response was affirmative, "How much difficulty do you have walking across a small room?". The responses provided an ordinal measure of task difficulty, using the categories "none," "a little," "some," "a lot," or "unable to do." Self-reports of difficulty with 19 other activities or tasks were obtained in an analogous manner. All 20 self-report items and the prevalence of reported difficulty for each item are listed in table 1.

Statistical methods and models

In most of the analyses for which results are presented in this report, ordinal difficulty responses were dichotomized into *any* reported difficulty versus *no* reported difficulty. We refer to the proportion of positive reports of difficulty on a particular task as the *prevalence* of difficulty in the sample of 108 women. Prevalences of difficulty by subgroup and over time were initially examined using graphic smoothing techniques (19, 20) to plot the proportion of positive responses versus study week for subgroups defined by age and baseline disability. Research questions and

TABLE 1. Domains of disability in which difficulty performing certain tasks was assessed, and prevalences of "any" reported difficulty, WHAS* Weekly Substudy, Baltimore, Maryland, 1993–1994

Disability domain†	Task	Abbreviation	Prevalence (%)
Upper extremity	Grasping or handling	Grip	27
	Lifting something as heavy as 10 lb‡	Lift	52
	Raising arms up over head	Arms	23
	Turning a key in a lock	Turn key	8
Mobility/exercise tolerance	Crouching or kneeling	Kneel	68
	Doing heavy housework	Hv. hwk.	78
	Getting into and out of a bed or chair	Bed	37
	Walking 1/4 mile§	Walk (1/4)	78
	Walking up 10 steps without resting	Steps	62
	Walking across a small room	Walk (room)	34
Higher functioning	Doing light housework	Lt. hwk.	27
	Managing money	Money	22
	Preparing meals	Meals	28
	Shopping for personal items	Shop	48
	Taking medications	Medicine	8
	Using the telephone	Phone	8
Self-care	Bathing or showering	Bathe	43
	Dressing	Dress	25
	Eating	Eat	14
	Using the toilet	Toilet	23

* WHAS, Women's Health and Aging Study.
† Defined at baseline from the screening interview, using self-reported difficulty with at least one of a subset of tasks: upper extremity = arms, grip, lift; mobility/exercise tolerance = steps, bed, hv. hwk., walk (1/4); higher functioning = phone, lt. hwk., meals, shop; self-care = bathe, dress, eat, toilet.
‡ 10 pounds = 4.5 kg.
§ 1/4 mile = 0.4 km.

hypotheses about the prevalence were then more systematically addressed through logistic regression models (21) for the binary responses. Formally, we have

$$\text{logit}\{E(Y_{it}|x_i, t)\} = f_1(t) + x_i'\beta, \quad (1)$$

where Y_{it} is the binary observation on the ith subject at time t, x_i is a generic vector of covariate values and indicators such as age group (three groups) or disability level (number of disability domains at baseline), and $f_1(t)$ is a smooth function of study time, t. While this model describes the average prevalence as a function of age, disability level, and study time, it does not imply that responses from a given subject are independent. In fact, they may be strongly associated.

Questions of the consistency of repeated responses were addressed by examining the *association* between responses given by the same individual at two different time points, where association was quantified via the odds ratio. For example, data from the two binary reports of difficulty walking across a small room for weeks 1 and 2 yielded an odds ratio of 12.4, indicating that the odds of reporting difficulty at week 2 for a subject reporting difficulty at week 1 were 12 times higher than those for a subject reporting no difficulty at week 1. To study the week-to-week consistency of responses, we set up an association regression model. Here, the log odds ratio is assumed to be a linear function of a second covariate set, z_i, which may contain some of the same covariates as x_i. We have

$$\log\{\text{OR}(Y_{it}, Y_{it+\tau}|z_i, x_i)\} = f_2(\tau) + f_3(\gamma) + z_i'\alpha, \quad (2)$$

where f_2 and f_3 are smooth functions, τ is the time lag between two responses, $\gamma = t + \tau/2$ is their time midpoint, and z_i is a vector of subject-level characteristics. Note that in equation 2, we model the association between two responses, adjusting for any association due simply to the fact that observations on the same subject "share" covariate values, x_i. Using such a model, we were able to study both the effect of time interval between two responses on week-to-week consistency and the effect of subject-level characteristics such as age and disability level. A special case of the model shown in equation 2 is $\log\{\text{OR}(Y_{i,t}, Y_{i,t+\tau}|x_i)\} = f_2(\tau)$, the "LORellogram" (log odds ratio) function proposed by Heagerty and Zeger (22). A plot of this function displays the log odds ratio between two binary responses obtained from a single subject at time points t and $t + \tau$ as a function of the lag time (τ)

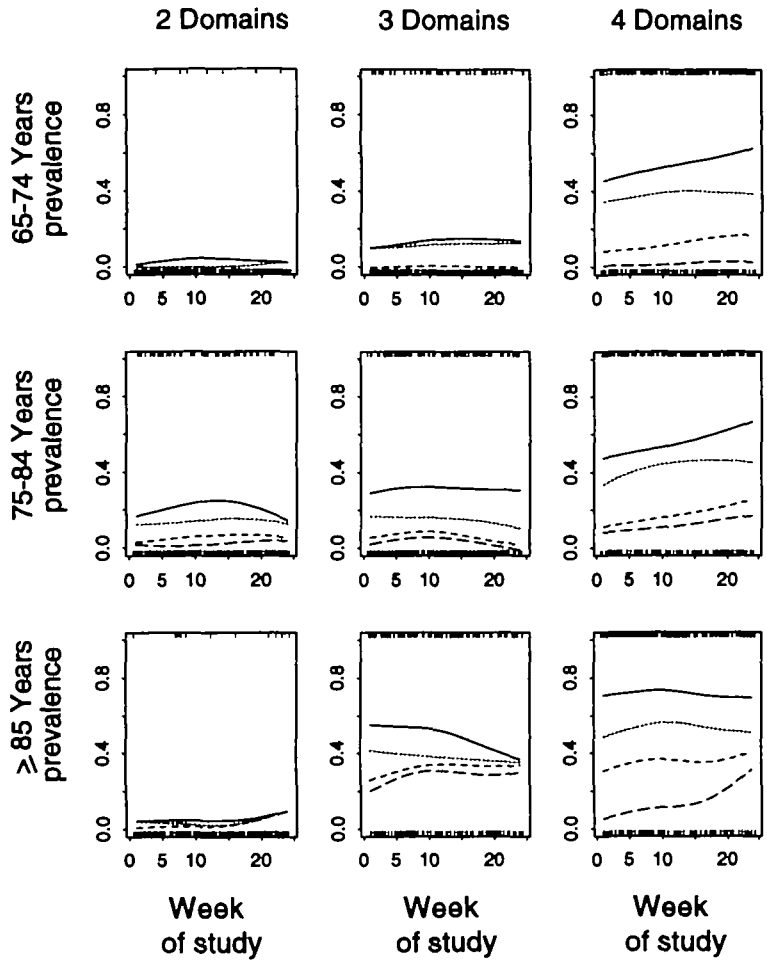

FIGURE 1. Cumulative prevalence of reported difficulty walking across a small room, by age group, number of disability domains, and week of study: Weekly Substudy of the Women's Health and Aging Study, Baltimore, Maryland, 1993–1994. Each hash mark (|) along the top and bottom axes represents a response of no difficulty (bottom) versus any difficulty (top). Points are "jittered" horizontally for readability. The solid curve (—) shows the prevalence of any (versus no) reported difficulty walking across a small room. The next-lowest two curves show the prevalences of reporting at least "some" difficulty (. . .) or at least "a lot" of difficulty (- - -), respectively. The lowest curve (– – –) shows the prevalence of reporting being "unable" to walk across a small room. Thus, the space between the horizontal axis and the lowest curve represents the proportion of subjects who reported being "unable" to perform the task; the space between the bottom curve and the next-highest curve represents the proportion who reporting having "a lot" of difficulty; etc. The space above the solid line represents the proportion of women reporting no difficulty.

separating them, creating a binary data analogue of the autocorrelation function (correllogram) for continuous longitudinal data (23).

To initially examine the association, we used a simple technique for estimating the LORellogram (see Appendix). Then the prevalence and association models were simultaneously fitted using the extension of the generalized estimating equations (24) methodology proposed by Heagerty and Zeger (25). In this method, each regression (equations 1 and 2) takes account of the results of the other to produce valid and efficient inferences about the regression coefficients β and α. Given the large number of functional status items available, we selected "difficulty walking across a small room" to study in detail and to illustrate the methodology. We then extended these analyses to the broader set of 20 self-report items.

RESULTS

Only study subjects who had at least two weekly responses to each of the 20 self-report items—102 of the original 108 subjects—were included in this analysis. For the item "difficulty walking across a small room," 50 percent of the subjects responded at least 23 times out of a possible 24, and 75 percent responded at least 20 times, for an overall response rate of 86 percent (all responses/all possible responses). A similar pattern held for the other self-report items. In addition, 33 percent of the subjects never reported any

difficulty walking across a small room, and about half reported difficulty in fewer than 10 percent of their interviews. Approximately one fourth reported difficulty more than 80 percent of the time, and nine subjects reported being unable to perform the task most of the time.

Prevalence of difficulty by study time, age, and disability level

Figure 1 presents the observed prevalence of reported difficulty in walking across a small room (at several levels of difficulty), stratified by age, disability, and study time (week of study). There was a clearly increasing trend in reported difficulty with baseline disability level and, in the more disabled groups, with age. The within-stratum time trends for walking across a small room were weak and inconsistent. In addition, trends in prevalence over time (averaged over age group and disability level) were very flat for all 20 tasks considered (results not shown).

Table 2 presents the effects of age and baseline disability on the prevalence of reported difficulty in walking across a small room, assuming no time trend. Model 1 shows that age had an effect on the prevalence of difficulty only in the two more disabled groups; however, the four interaction terms were not statistically significant ($\chi^2 = 4.55$, 4 df). Assuming a common age effect across disability groups (model 2), both increasing age and increasing disability level were associated with increasing prevalence of difficulty. The effect of baseline disability was quite strong, with an estimated odds ratio of 20.4 (95 percent confidence interval (CI) 8.1–51.3) for reporting difficulty in the four-domain group versus the two-domain group. The age effect was weaker, with an odds ratio of 3.8 (95 percent CI 1.5–9.2) for the oldest group compared with the youngest.

Statistically significant effects of age and baseline disability for all 20 items are shown in table 3. A pattern similar to the results for walking across a small room was observed for increasing level of disability (two, three, or four domains); that is, subjects with more baseline disability reported more difficulty on the individual tasks. Results were weaker for the effect of age. Older age (75–84 years or ≥85 years compared with 65–74 years) was associated with more difficulty for many items. However, being in the oldest (≥85 years) age group versus the middle age group was associated with more difficulty for only five items.

Consistency of reported difficulty by time interval, age, and disability level

The associations between pairs of responses obtained from the same subject, as a function of time lag and baseline disability group, are displayed in figure 2 (based on model 1 in table 4). The association was very strong for short time lags, indicating high consistency, and it was strongly dependent on the number of weeks separating the responses ($\chi^2 = 18.6$, 4 df). The estimated odds ratio for a time lag of 1 week in the two-domain group was 21.7 (95 percent CI 9.6–49), as compared with 3.4 (95 percent CI 1.0–11.2) for a time lag of 23 weeks; the associations were similar across all three disability groups.

Table 4 shows the estimated effects of time lag τ, age

TABLE 2. Odds ratios for self-reported difficulty walking across a small room, as estimated under association model 2, WHAS* Weekly Substudy, Baltimore, Maryland, 1993–1994

	Model			
	1		2	
	OR*	95% CI*,†	OR	95% CI†
No. of disability domains				
2	1.0		1.0	
3	4.7	0.8–27.8	4.3	1.7–10.8
4	36.1	10.4–125.1	20.4	8.1–51.3
Age group (years)				
65–74	1.0		1.0	
75–84	6.6	1.8–25.0	2.7	1.1–6.7
≥85	2.7	0.6–12.8	3.8	1.5–9.2
No. of domains × age				
3 × 75–84	0.4	0.1–3.8		
3 × ≥85	2.8	0.2–32.3		
4 × 75–84	0.2	0.0–1.7		
4 × ≥85	1.1	0.1–7.9		

* WHAS, Women's Health and Aging Study; OR, odds ratio; CI, confidence interval.
† Confidence intervals were constructed using empirical variance estimates.

TABLE 3. Estimated odds ratios* for self-reported difficulty performing various tasks, by baseline disability level and age, WHAS† Weekly Substudy, Baltimore, Maryland, 1993–1994

Item	No. of disability domains‡		Age group§ (years)	
	3	4	75–84	≥85
Grip	1.0	3.2		
Lift	1.4	4.4	2.9	1.3
Arms	1.7	5.6	1.4	0.5
Turn key	0.8	4.4		
Kneel	1.9	4.5		
Hv. hwk.	3.1	12.4	3.8	1.2
Bed	1.5	5.3		
Walk (1/4)	2.0	11.7	3.2	2.7
Steps	1.6	5.1	2.6	2.4
Walk (room)	4.4	20.8	2.6	3.8
Lt. hwk.	1.8	9.5	2.9	1.3
Money	3.7	2.9	3.2	5.4
Meals	3.4	11.7		
Shop	3.2	7.6	1.5	3.7
Medicine			1.3	3.9
Phone			0.4	1.8
Bathe	2.0	8.3	3.2	2.8
Dress	1.6	7.1		
Eat	0.5	5.8		
Toilet	2.2	11.2	2.7	3.4

* Entries presented were found to be significant factors using Wald-type test statistics (2 df) constructed with empirical covariance estimates.
† WHAS, Women's Health and Aging Study.
‡ Reference category: two domains.
§ Reference category: ages 65–74 years.

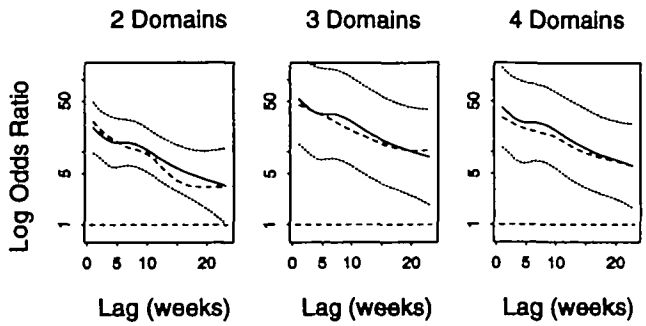

FIGURE 2. Fitted LORellograms (—), pointwise 95 percent confidence intervals (. . .), and empirical Mantel-Haenszel LORellograms (– – –) for reporting having "any" difficulty walking across a small room, by number of disability domains and time interval between reports (association model 2): Weekly Substudy of the Women's Health and Aging Study, Baltimore, Maryland, 1993–1994. The vertical axis shows the odds ratio on the logarithmic scale.

group, disability level, and time midpoint γ on the degree of association between pairs of observations obtained from the same subject. These estimates were adjusted for the changing prevalence across age groups and disability levels observed in the prevalence model. Before continuing, we must briefly consider the interpretation of the parameters in association model 1. Since the unit of analysis is a pair of observations obtained from the same subject, the intercept in equation 2 is the log odds ratio between two responses in the reference group of pairs of responses. In model 1, the reference group consists of pairs of observations made 23 weeks apart on subjects in the two-domain disability group (odds ratio = 3.4, 95 percent CI 1.0–11.2), indicating that a subject reporting difficulty in the first week will have 3.4 times the odds of reporting difficulty during the last week in comparison with a subject reporting no difficulty during the first week. The other coefficients are differences between log odds ratios. Exponentiating them produces ratios of odds ratios for the association between two pairs of points, relative to two points in the reference group. Continuing with model 1, the estimated odds ratio for the association between the first and last responses given by subjects in the three-domain disability group is 2.5 times that for the two-domain group (95 percent CI 0.5–11.8). Specifically, the odds ratio between the first and last responses obtained from subjects in the three-domain group is given by $3.4 \times 2.5 = 8.5$.

The association was slightly higher between responses obtained from subjects in the three- and four-domain disability groups, but these differences were not statistically significant (model 1; $\chi^2 = 1.80$, 2 df). Responses from subjects in the 75–84 and ≥85 age groups were somewhat more consistent than responses from subjects in the 65–74 age group, but again this effect was not statistically significant (model 2; $\chi^2 = 2.85$, 2 df).

Consistency of reported difficulty with 20 tasks, by time interval

In similar analyses of the association between two responses, LORellogram models (without effects for age or disability level) were fitted to the data for each of the 20 task difficulty items in table 1, adjusting for varying prevalence by baseline age × disability stratum. Results similar to those for walking across a small room are displayed in figure 3 and table 5. Considering the entire set of 20 items, the test-retest consistency over a time lag of 1 week was very high; the minimum odds ratio relating two consecutive responses was 31, with a lower 95 percent confidence limit of 16. In addition, there was a strong negative effect of time lag on the degree of consistency. Nevertheless, over a time lag of 11 weeks, the association was still high; the odds ratios were all at least 10, and the lower limits of the 95 percent confidence intervals were at or above 5. At a time lag of 23 weeks, the odds ratios were all at least 4, and the lower limits of the 95 percent confidence intervals were all at least 1.

TABLE 4. Fitted ratios of odds ratios relating two self-reports obtained from a given subject on difficulty walking across a small room, WHAS* Weekly Substudy, Baltimore, Maryland, 1993–1994

	Model					
	1		2		3	
	ROR*,†	95% CI*,‡	ROR†	95% CI	ROR†	95% CI
Intercept (odds ratio)§	(3.4)	(1.0–11.2)	(3.7)	(0.8–16.7)	(6.5)	(2.3–18.9)
Time lag¶ (weeks)						
23	1.0		1.0		1.0	
11	2.7	1.1–7.0	2.6	1.0–7.0	3.2	1.1–9.8
1	6.4	2.1–19.2	6.1	2.0–18.9	9.8	2.0–47.6
Disability domains (no.)						
2	1.0					
3	2.5	0.5–11.8				
4	1.9	0.5–7.7				
Age group (years)						
65–74			1.0			
75–84			4.8	0.7–31.8		
≥85			1.9	0.2–15.0		
Time midpoint** (weeks)						
12.5					1.0	
6.5					0.5	0.2–1.1
1.5					0.2	0.0–1.2

* WHAS, Women's Health and Aging Study; ROR, ratio of odds ratios; CI, confidence interval.
† The denominator odds ratio quantifies the association between two responses given in the reference group, and the numerator odds ratio quantifies the association between two responses given in the comparison group.
‡ Confidence intervals were calculated using robust estimates of variance.
§ Fitted odds ratio for the association at the reference values of the model. In model 1, for example, observations made at weeks 1 and 24 (lag = 23 weeks, midpoint = 12.5 weeks) had an odds ratio of 3.4.
¶ Effects of time interval between responses (τ), modeled as a natural spline function (4 df).
** Effects of time midpoint (γ), modeled as a natural spline function (2 df).

Consistency of responses by study time

Model 3 in table 4 shows a positive but not statistically significant effect of study time on the association between responses, suggesting that responses may be more consistent later in the study period ($\chi^2 = 3.48$, 2 df). For example, holding time lag constant at 1 week, the odds ratio for two responses given in weeks 12 and 13 (midpoint = 12.5 weeks) was estimated to be 64 (95 percent CI 17–238). This odds ratio was obtained by multiplying the intercept term by the effect for "lag = 1 week" ($6.5 \times 9.8 = 64$). The estimated odds ratio for two responses given in weeks 1 and 2 (midpoint = 1.5 weeks) was 12.5 ($6.4 \times 9.8 \times 0.2 = 12.5$; 95 percent CI 3.9–39). Thus, while the association over weeks 1 and 2 was still high, two responses given in the middle of the study period were 5.2 (95 percent CI 0.8–33) times more consistent than those given earlier in the study.

DISCUSSION

In this paper, we examined the consistency of self-report measures of physical functioning among disabled elderly women, using a unique data set with weekly reports obtained over a 6-month period. Women rated the most disabled at baseline (those having difficulty in all four domains) reported more difficulty for each of the 20 individual items over the ensuing 6 months, while the least disabled women (those having difficulty in two domains at baseline) had the lowest prevalence of difficulty. Older age, within each disability group, was also associated with increasing prevalence of reported difficulty. Since the prevalence varied as expected with age and reported severity of baseline disability, these findings support the internal validity of these items as measures of disability. Furthermore, for none of the 20 items did the prevalence of reported difficulty vary by study time (from week 1 to week 24). While those results are not presented here, they indicate strong test-retest reliability when the results are used to assess the population prevalence of disability.

At the individual level, the consistency of reporting difficulty in functioning was quite high with a time lag of 1 week (odds ratios ranged between 31 and 112), and it decreased substantially for longer time lags. Responses remained significantly and positively associated, however, at time lags of up to 23 weeks. Based on our analysis of difficulty in walking across a small

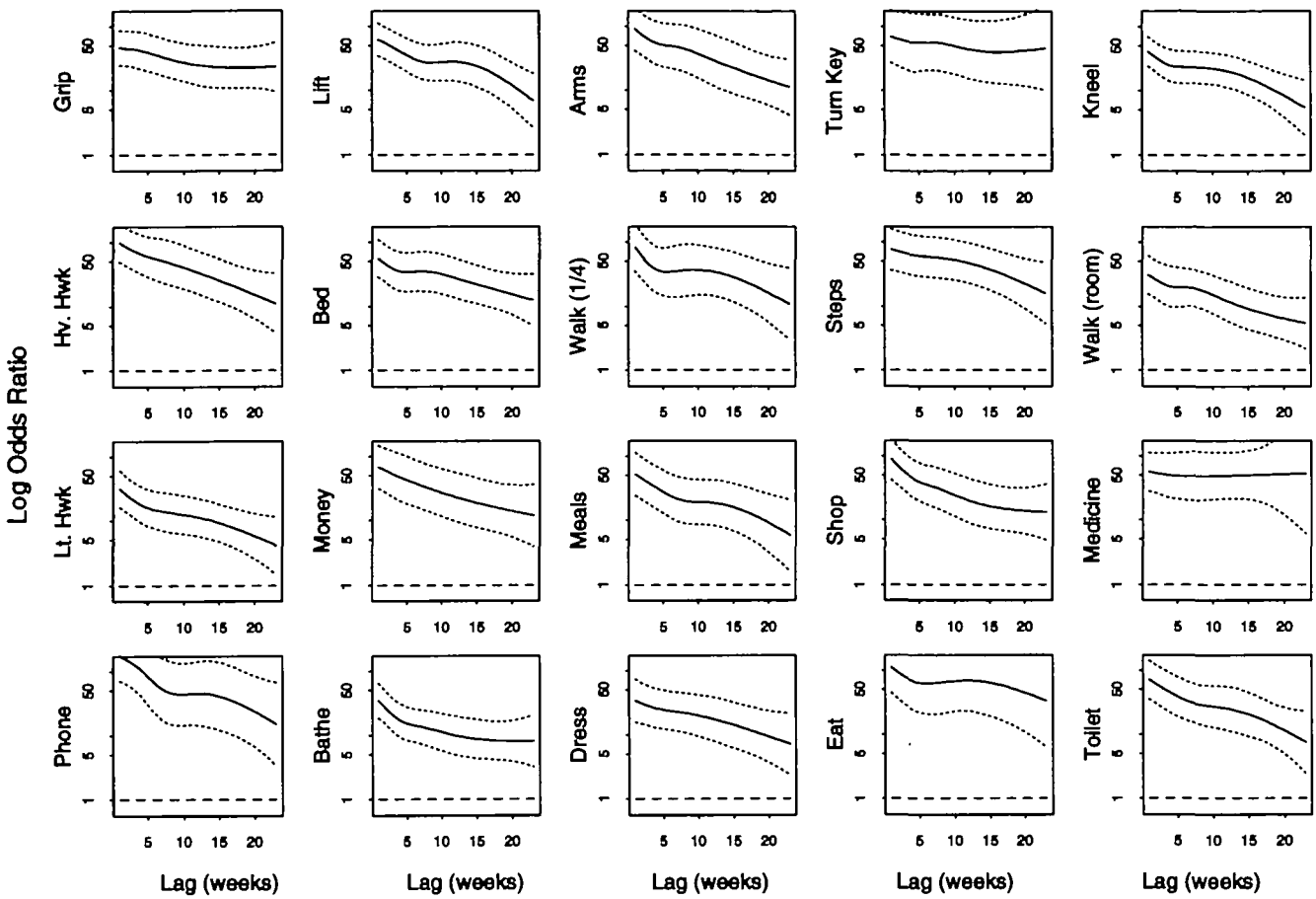

FIGURE 3. Fitted LOReilograms (—) and pointwise 95% confidence intervals (. . .) for reporting having "any" difficulty in performing each of 20 different tasks, by time lag between reports: Weekly Substudy of the Women's Health and Aging Study, Baltimore, Maryland, 1993–1994. All associations were estimated under prevalence model 1, except that for telephone use, which was estimated under prevalence model 2. The vertical axis shows the odds ratio on the logarithmic scale. (Hv. Hwk, heavy housework; Lt. Hwk, light housework; Walk (¼), walking ¼ mile (0.4 km).)

room, there is only weak evidence that the consistency of responses varies with age group or disability level; thus, responses obtained from a very old (over age 85), severely disabled woman would not be expected to differ in consistency from those of a younger, moderately disabled subject. If anything, the data suggest that subjects who are more disabled are slightly more consistent in their responses. This result is important, since self-report measures are often used to compare subjects from different subgroups, and therefore the reliability of such comparisons depends on the reliability of the measures for each subgroup.

Our results provide only weak evidence that week-to-week response consistency is higher later in the study period, after the subjects have experienced the interview more times. The fact that the reliability of self-reported information on functional status could depend on the number of times a participant is assessed is certainly of concern, especially in the context of longitudinal studies. On the one hand, greater consistency among later interviews may indicate that subjects become more attentive to their functional status through repeated questioning, so that later responses are more valid. If this is so, it might be important, for example, to have a "run-in" period of interviewing before assigning subjects to treatment in a clinical trial using outcomes based on reported functioning. On the other hand, if greater consistency is a sign of respondent fatigue—i.e., simply recalling and repeating one's previous responses—it may be important to limit both the frequency of contact and the number of contacts in order to maintain the validity of the measures. Our study provides only weak and nonsignificant evidence for such an effect (i.e., higher consistency among responses given in later interviews); however, this remains an issue that should be addressed in studies with frequent subject assessment.

The statistical properties of indicators of functional status, especially reliability and internal validity, are critical issues, since they form the basis of assessments

TABLE 5. Estimated odds ratios relating subjects' responses at two different times with intervals of 1, 11, and 23 weeks between assessments, in descending order of strength of association at an 11-week interval: WHAS* Weekly Substudy, Baltimore, Maryland, 1993–1994

Item	1-week lag		11-week lag		23-week lag	
	OR*,†	95% CI*	OR†	95% CI	OR†	95% CI
Eat	112	45–280	69	23–202	33	6–178
Steps	77	37–162	51	25–104	16	5–48
Medicine	56	28–113	49	21–112	53	6–455
Turn key	70	27–177	47	17–131	46	10–207
Phone	176	71–438	45	15–138	15	3–67
Hv. hwk.	98	49–195	37	17–80	11	4–33
Walk (1/4)	83	35–195	36	15–85	11	3–39
Arms	91	41–199	34	14–83	11	4–31
Bed	56	29–109	30	15–60	13	5–32
Lift	62	34–112	28	14–55	7	3–18
Toilet	71	35–143	26	12–56	7	2–23
Grip	46	24–87	26	13–52	23	9–56
Money	67	31–146	25	11–59	12	4–36
Shop	89	43–184	22	10–48	13	5–36
Kneel	40	24–69	21	12–37	5	2–14
Meals	52	24–113	19	8–42	6	2–21
Dress	34	16–74	19	9–40	7	2–22
Walk (room)	31	16–61	14	7–28	5	2–14
Lt. hwk.	32	16–61	12	6–23	4	1–11
Bathe	35	19–64	10	5–21	8	3–20

* WHAS, Women's Health and Aging Study; OR, odds ratio; CI, confidence interval.
† Adjusted for age × disability level.

of disability in the elderly. In this study, we found substantial evidence for the internal validity and test-retest reliability of 20 self-report measures of functioning using weekly assessments in a group of elderly, disabled women. Consistency of reporting did decline, however, with time between assessments. Since the opportunity for true change in functioning due to a broad range of factors (e.g., incident disease, changes in living arrangements, increased severity of existing conditions) increases with the time interval between assessments, this finding is not surprising. Whether the decline in the consistency of individuals' reporting between assessments at longer intervals can be attributed to real change in health and functioning is beyond the scope of this analysis, but it represents a plausible hypothesis on the basis of these results.

This study provides insight into the appropriate time intervals for assessment of self-reported functioning in studies of elderly people. Measures made at intervals of less than 12 weeks provide largely redundant information, given the very high association between measures over this time interval. In contrast, measures made at intervals of 24 weeks or longer begin to show substantial within-subject variability, suggesting that the magnitude of change in functional status occurring over these intervals is detectable by self-report, and hence may represent new information on the natural history of disability in an individual.

ACKNOWLEDGMENTS

This research was supported by contract NO-1AG-1–2112 from the National Institute of Aging and grant 2P30-MH-38725–11 from the National Institute of Mental Health. The work was performed while Dr. Rathouz was a Ph.D. candidate in the Department of Biostatistics at the Johns Hopkins University School of Hygiene and Public Health.

REFERENCES

1. Guralnik JM. Understanding the relationship between disease and disability. (Editorial). J Am Geriatr Soc 1994;42:1128–9.
2. Ware JE Jr. The status of health assessment 1994. Annu Rev Public Health 1995;16:327–54.
3. Ware JE Jr. Measuring functioning, well-being and other generic health concepts. In: Osoba D, ed. The effect of cancer on quality of life. Boca Raton, FL: CRC Press, 1991:17–23.
4. Freedman VA, Soldo BJ, eds. Trends in disability at older ages: summary of a workshop. Washington, DC: National Academy Press, National Academy of Sciences, 1994.
5. Manton KG, Corder LS, Stallard E. Estimates of change in chronic disability and institutional incidence and prevalence rates in the U.S. elderly population from the 1982, 1984, and 1989 National Long Term Care Survey. J Gerontol 1993;48:S153–66.
6. Mathiowetz NA, Lair TJ. Getting better? Change or error in the measurement of functional limitations. J Econ Soc Meas 1994;20:237–62.
7. Katz S, Ford AB, Moskowitz RW, et al. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA 1963;185:914–19.
8. Nagi SZ. An epidemiology of disability among adults in the United States. Milbank Mem Fund Q Health Soc 1976;54: 439–67.

9. Idler EL, Kasl S. Health perceptions and survival: Do global evaluations of health status really predict mortality? J Gerontol 1991;46:S55–65.
10. Idler EL, Kasl SV, Lemke JH. Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982–1986. Am J Epidemiol 1990;131:91–103.
11. Kaplan GA, Camacho T. Perceived health and mortality: a nine-year follow-up of the Human Population Laboratory cohort. Am J Epidemiol 1983;117:292–304.
12. Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. Am J Public Health 1982;72:800–8.
13. Kovar MG. Aging in the eighties: preliminary data from the supplement on aging to the National Health Interview Survey, United States, January–June 1984. Hyattsville, MD: National Center for Health Statistics, 1986. (Advance data from vital and health statistics, no. 115) (DHHS publication no. (PHS) 86–1250).
14. Manton KG, Stallard E, Corder L. Changes in morbidity and chronic disability in the U.S. elderly population: evidence from the 1982, 1984, and 1989 National Long Term Care Surveys. J Gerontol B Psychol Sci Soc Sci 1995;50:S194–204.
15. Guralnik JM, Fried LP, Simonsick EM, et al. Screening the community-dwelling population for disability. In: Guralnik JM, Fried LP, Simonsick EM, et al, eds. The Women's Health and Aging Study: health and social characteristics of older women with disability. Bethesda, MD: National Institute of Aging, 1995:9–18. (NIH publication no. 95–4009).
16. Streiner DL, Norman GR. Health measurement scales. New York, NY: Oxford University Press, 1995.
17. Fried LP, Ettinger WH, Lind B, et al. Physical disability in older adults: a physiological approach. Cardiovascular Health Study Research Group. J Clin Epidemiol 1994;47:747–60.
18. Ferrucci L, Guralnik JM, Salive ME, et al. Effect of age and severity of disability on short-term variation in walking speed: The Women's Health and Aging Study. J Clin Epidemiol 1996;49:1089–96.
19. Hastie TJ, Tibshirani RJ. Generalized additive models. London, England: Chapman and Hall Ltd, 1990.
20. MathSoft, Inc. S-plus, version 3.3. Seattle, WA: MathSoft, Inc, 1995.
21. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London, England: Chapman and Hall Ltd, 1989.
22. Heagerty PJ, Zeger SL. LORellogram: a regression approach to exploring dependence in longitudinal categorical responses. J Am Stat Assoc (in press).
23. Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. New York, NY: Oxford University Press, 1994.
24. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–30.
25. Heagerty PJ, Zeger SL. Marginal regression models for clustered ordinal measurements. J Am Stat Assoc 1996;91:1024–36.
26. Davis LJ. Generalization of the Mantel-Haenszel estimator to non-constant odds ratios. Biometrics 1985;41:487–95.

APPENDIX

To construct a simple estimate of the LORellogram, we draw on the idea behind the Mantel-Haenszel estimator for the adjusted odds ratio. The subjects are first stratified by covariate class. Then, for each class, i, and each pair of time points, (t, t'), a 2×2 table is constructed, containing components $a_{itt'}$, ... $d_{itt'}$ and total $n_{itt'}$, where these are the conventional 2×2 table notations. For each table, $a_{itt'}d_{itt'}/n_{itt'}$ and $b_{itt'}c_{itt'}/n_{itt'}$ are computed. The ad/n terms and the bc/n are then fitted via separate smoothing splines as functions of $\tau = |t - t'|$ (19). The logarithm of the quotient of these two smooth functions can then be plotted as a function of τ to obtain the LORellogram. Davis (26) has proposed a more formal technique for regressing the log odds ratio on a set of covariates using a Mantel-Haenszel type of estimator; our procedure is more exploratory in nature, but it has the advantage of being very simple to implement.