Published in final edited form as:

J Am Geriatr Soc. 2000 November ; 48(11): 1478–1482.

Qualifying Change: A Method for Defining Clinically Meaningful Outcomes of Change Score Computation

Rochelle E. Tractenberg, PhD^{*}, Shelia Jin, MD, MPH^{*}, Marian Patterson, PhD[†], Lon S. Schneider, MD[‡], Anthony Gamst, PhD^{*}, Ronald G. Thomas, PhD^{*}, and Leon J. Thal, MD^{*}

^{*}Alzheimer's Disease Cooperative Study, Department of Neurosciences, University of California, San Diego

[†]Department of Neurology, Case Western Reserve University, Cleveland, Ohio

[‡]University of Southern California Keck School of Medicine, Los Angeles, California

Abstract

OBJECTIVES—To identify clinically meaningful change in longitudinal assessment.

DESIGN—A novel approach that qualifies item-level change over time by the degree to which it is clinically meaningful.

SETTING—The classification method was tested by applying it to changes over 12 months in the frequency ratings of the items of a behavioral assessment instrument that is used commonly in clinical trials with Alzheimer's disease (AD) patients.

PARTICIPANTS—Responses from a cohort of 235 well characterized, community-dwelling subjects with AD were analyzed by this method.

MEASUREMENTS—The approach allowed us to describe the proportions of items that emerged, ceased, worsened, and improved between the baseline and 12-month visits.

RESULTS AND CONCLUSIONS—One-year change in the behavioral symptoms of persons with AD was used to exemplify the methodology. This approach can be used in other populations and with other measurements and was designed for analyses of clinical trial data. This method uses item-level changes to generate global impressions of clinically meaningful change; it also facilitates the definition of change that can be used in the clinical setting.

Keywords

follow-up studies; outcome assessment (health care); health status indicators; clinical trials; methods; data interpretation; statistical

The assessment of change over time is an important health issue: it is a way of testing new medical treatments and of investigating the natural courses of aging and disease. Of considerable import and interest in clinical research is the definition of “clinically

meaningful” change. The definition of clinically meaningful change is challenged on at least two fronts, one statistical and the other clinical.

Statistically, there are two important problems associated with measuring change. First, there are limits to the amount of change possible when scores or ratings at the first of two timepoints in question are very high or very low. This differential sensitivity to change due to extreme scores is a characteristic of many assessment instruments, and it makes interpretation of some change scores impossible. Second is the issue of the reliability of the raw change score. Several authors¹⁻³ have argued that the errors associated with measurements at each timepoint greatly complicate the computation, interpretation, and subsequent testing of change scores.

Clinical impediments to defining meaningful change arise because (1) there is little information available on the performance of control populations on many instruments and (2) there is little information on the longitudinal use of many instruments and the types of changes to be expected (although see Green et al.⁴).

An example of a domain in which measuring change over time is important but difficult is behavioral problems arising during the course of Alzheimer’s disease (AD). There are many different tests being used to study various aspects of behavioral symptoms in AD; however, the relationship between changes on any two of these tests is unknown. Complicating interpretation of change on behavioral measures is the finding of waxing and waning⁴ or recurrence⁵ of behavioral symptoms, rather than sustained change, in persons with AD.

Rasmusson and colleagues⁶ reviewed concisely several mechanisms for defining cognitive changes in AD. In commenting on the issue of predicting decline, the assumption is that decline will be the only change. This assumption is valid over the very long term, but a substantial proportion of AD patients may remain stable or show cognitive improvement over a 12-month period. The distribution of cognitive change in the short term can also cover the spectrum from improvement through no change to worsening.⁷ Similarly, no-change, worsening, and improving have all been observed with different behavioral measures in studies of persons with AD.^{4,8,9} There are clearly many contingencies and difficulties in defining change in behavioral symptoms, as well as in other domains.

Although degenerative disease inevitably results in decline, truly sustained change in AD, especially behavioral change, may be observed only over very long periods of study. Therefore, change scores observed in clinical trials may not be unidirectional. It would be helpful to have a simple way of characterizing change over the short or long term.

Another impediment to defining change is that on many instruments, change in the total score has no fixed meaning. Currently, no norms exist for the behavior assessment instruments used widely in research of patients with AD, such as the Behavior Rating Scale for Dementia (BRSD)¹⁰ and the Cohen-Mansfield Agitation Inventory (CMAI),¹¹ although standardization data are available for the BRSD from a large sample of individuals with AD.^{12,13} Statistical means to assess change in behavioral symptom frequency do exist,³ but mapping clinical meaning on to a statistically significant change over a study period is not straightforward.^{14,15}

Only the actual quantification of change is required to determine differences in change across treatment groups, but this approach may lead to the identification of a treatment as effective when the actual change, while statistically significant, is clinically meaningless.¹⁶ Therefore, this study sought to qualify change in an effort to differentiate clinically meaningful change from fluctuation and, at the same time, identify stability (no change) caused by the absence of symptoms from that caused by the continued presence of symptoms.

Bereiter¹ points out that the unreliability of change scores comes from the respective unreliabilities of test scores at times 1 and 2 (echoed by Overall and Woodward²). To increase the reliability of a change score, Bereiter proposed focusing analysis on the item level. The model described here follows from this proposal by computing change on a per-item basis and then qualifying this change to facilitate the decision as to whether the observed change is clinically meaningful.

In order to define clinically meaningful change, with the example of a behavioral measurement for persons with AD, we defined the types of changes we expect to carry the most clinical meaning based on the possible changes we can observe between two visits. We classified changes depending on whether the symptom never appeared, emerged, disappeared, increased or decreased in frequency, or remained at a constant level over 1 year. To test this approach, we applied it to changes in the reported frequencies of the specific behaviors assessed in the BRSD over 12 months in a group of well characterized AD patients living in the community. The BRSD total and subscores, as well as each item, have well established validity and reliability.^{9,10,12,13}

METHODS

Subjects

This patient population has been described elsewhere.¹⁷ Subjects were 235 community-dwelling individuals with NINCDS-ADRDA-based diagnoses of probable AD¹⁸ who were participants in an investigation of new instruments for use in AD clinical trials. The Mini-Mental State Examination (NIMSE)¹⁹ was administered to all participants at a screening visit. Informed consent was obtained for all participants from caregivers. Subjects' mean (SD) age was 72.30 (9.01) years, mean educational level was 13.15 (2.88) years, mean MMSE score was 12.80 (7.93), and 61.0% were women.

Materials

The BRSD was administered by a clinician or technician to a caregiver as a 48-item instrument but was recoded as 46 items according to current scoring rules.¹³ Our analyses focused on the 37 items with frequency ratings of 0 to 4, excluding items 9, 10, 12, 14, 15, 17, 26, 32, and 46, which are rated as 0/1 and do not describe specific behaviors (see Table 1 for list of items).

Design and Procedure

Changes in ratings for the 37 frequency-rated BRSD items were calculated as (BL-12m). These change scores were recoded as described below. The observed change types were then tabulated.

Statistical Methods

SPSS 8.1 for Windows 98 was used to compute change scores and to tabulate the results.

Frequency ratings of BRSD items range from 0 to 4: 0 = not in the past month; 1 = 1–2 days in the past month; 2 = 3–8 days in the past month (up to twice per week); 3 = 9–15 days in the past month (up to half the days in the past month); and 4 = 16 or more days in the past month. We defined “not in the past month” as “not present”, and 1–2 days in the past month as “minimally frequent”. Symptoms that occurred at least 3 days in the past month were defined as “moderately frequent”. Based on the difference between the rating given for each item at baseline and at 12 months, we classified symptom status over 1 year as one of seven mutually exclusive and exhaustive possibilities:

1. Absent/Intermittent: not present or minimally present at both visits.
2. Waned: decreased in frequency from moderate to lower, but still moderate, over two visits.
3. Abated: decreased in frequency from moderate to minimal over two visits.
4. Ceased: at least moderately frequent at baseline and not present at the second visit.
5. Emergent: not occurring at baseline and rated as moderate at second visit.
6. Persistent: at least moderate frequency consistent at both baseline and second visit.
7. Intensified: at least minimally frequent at baseline and at least moderately frequent at second visit.

Only four of these categories (abated, ceased, emergent, and intensified) represent clinically meaningful changes (improvement/worsening) in the status of a symptom. Items with change scores categorized as “absent” or “persistent” (both represent no change) could help to define syndromes or symptom-complexes, but as there is no change observed, these could not be considered “clinically meaningful change”. However, symptoms that remain absent over the study period may be important in defining overall change.

Levy et al.⁵ described the presence of a symptom as a score of at least mild (greater than either of the ratings “none” and “very mild”). We adopted a similar approach by classifying symptoms as “present” if they were rated to be more frequent than “not in the past month” (“none”) or “minimally frequent” (“very mild”), i.e., rated as occurring more than 2 days in the past month. Change categorized as “waned” is not considered to represent meaningful change, although this may be a “marker” for the onset of a more substantial decrease in frequency. Although we classified very minimal improvement as “waning”, and did not consider this to be clinically meaningful, we classified any (including very minimal) worsening as “intensified”. This was done in order to emphasize a sensitivity to worsening of symptomatology.

RESULTS

The results indicated that the modal symptom status was “absent/intermittent”, that is, we most frequently observed symptoms occurring 0 to 2 days per month at both baseline and at the 12-month visit. We also saw both waxing and waning in the frequency-rated BRSD items that we classified. Item 30 (repetitiveness) was an exception in that this symptom was persistent in 35.9% of the sample, and absent in 23.2%. These figures each represent “no change over 12 months” in the largest proportion of the group, and the classification methodology described here allows us to identify these as distinct types of stability.

Although the purpose of this paper is to describe the method, rather than the 1-year change on the BRSD, the proportions of symptom absence and the types of changes we defined to be clinically-meaningful (cessation, emergence, improvement, and intensifying) are presented in Table 1.

Aside from the large proportions of the sample for which these behaviors were absent over 12 months, we saw clinically-meaningful change in 10% or more of the sample on 13 of 37 items (1, 2, 3, 5, 11, 13, 19, 20, 21, 24, 25, 31, and 41). The different types of change per item can be seen in Table 1. For example, the proportions of the group in whom symptoms ceased and intensified, or ceased and emerged, were very similar in items 1, 2, 13, 24, and 25. We can see that item 3 intensified in 20% of the sample, more than twice the proportion in whom it ceased (8.0%). Similarly, frequency on item 5 intensified in twice as many (11.9%) as the number in whom it ceased (5.5%), and item 41 emerged in nearly twice as many (11.2%) as it ceased (7.6%). Conversely, item 31 ceased at a rate (12.1%) nearly twice that for either emergence or intensifying.

General Discussion

By qualifying rather than quantifying the types of changes in symptoms observed in these AD patients, we hoped to provide a basis for characterizing observed changes as meaningful improvement, meaningful worsening, or as fluctuation only (i.e., not clinically relevant change). As defined here, improvement is clearly distinguishable from worsening on an item-by-item level. It is clear that the behavioral symptoms in this group are changing in different ways, and symptoms seem to be more likely to cease than they are to abate. Even without discussing the specifics of the symptoms themselves, classification of 1-year change over time indicates that behavioral disturbance in AD may be highly individualized. These distinctions can be used in experimental, observational, or clinical settings for assessments where responses are categorical.

Matthews et al.²⁰ propose a straightforward, statistically valid approach to assessing medical data over time, but it requires a “suitable summary of the response in an individual, such as a rate of change or an area under a curve” (p. 230). The qualitative categories described here can be used to construct such summary quantitative analytic parameters. An example of this would be to count the number of items that have ceased, abated, or waned, (no. improved) and compare this figure with the number of items that have emerged or intensified (no. worsened). If no. improved is less than no. worsened, then label the subject worse; if no.

improved is more than no. worsened, then the subject is improved; and if the two figures are equal, the subject has “not changed”. The total number of items which were absent over the study period can also be figured into these simple summaries.

To increase sensitivity to multiple dimensions of a domain under assessment, one way to define improvement is to label a subject “improved” if the number of items on which there was improvement is greater than the larger of the number of items on which there was either worsening or no change (present-stable):

“improved” {#improved > [max(#worsened, #unchanged)]} | (#improved > #absent)

In this example, the label is contingent on whether the number of items on which there was improvement is greater than the number of items which were absent. Someone with seven improved, six worsened, four unchanged, and six absent items would qualify under this definition as “improved” ($7 > [\max(6, 4)]$ & $7 > 6$).

A simpler alternative would be to classify individuals as “improved” or “worsened” by comparing the numbers of symptoms that emerged and ceased. In this example, any individual for whom no. ceased was greater than no. emerged could be called improved; or someone for whom the ratio of these two figures is greater than some criterion could be called improved. These definitions of clinically meaningful change (improvement) can be tested for validity, easily interpreted, and, of course, they can also be modified. Also, rather than applying these categories to an entire measure, clinically meaningful change on subgroups of items or factors within a measure can be assessed individually.

Ultimately, the definitions of “improved” and “worsened” can be constructed from the observed types of change, depending on how conservative the analysis is intended to be. Furthermore, the definitions of “not present”, “minimally,” and “moderately” frequent can be changed according to treatment, instrument and protocol parameters.

The categories presented here are applicable to other instruments used in the assessment of behavior, both in AD and in other populations. Examples of instruments used to assess behavioral symptomatology in AD are the CMAI and the Brief Psychiatric Rating Scale (BPRS).²¹ These categories can also be applied to measures of severity (rather than frequency), caregiver burden, and/or distress. This approach is also generalizable for the analysis of change for any categorical clinical data because it separates items with no change (change score = 0) into three different types: absent at both visits (rated 0 - rated 0), intermittent (1-1), and stable (2-2, 3-3, etc.). Such differentiation may facilitate selection of, or improve, missing data imputation schemes such as “last observation carried forward” by differentiating stable-absent from stable-present symptoms. This method also differentiates between emerging symptoms and present-and-worsening symptoms, which may facilitate outcome analyses; various interventions may act differently in terms of preventing the emergence of new symptoms, ameliorating the worsening, or bringing about the cessation, of symptoms that are already present. This can be useful for describing outcomes in clinical trials, in observational studies, and in clinical practice.

In addition to facilitating the analysis of change, this approach can also elucidate underlying mechanisms in terms of symptoms (such as disturbed behaviors) that are observed at a

constant frequency over time (stable-present); these may be refractory to treatment if observed in the context of a clinical trial. In observational studies, stable-present symptoms may represent syndromes that could be studied further, whereas individuals with stable-absent symptoms may be protected or differ in some systematic way from the rest of the population. The method outlined here is flexible in that it allows both qualitative and quantitative analysis at the item level, or it can be used globally. It can also be used for domains other than behavior and in populations other than AD.

We are in the process of utilizing this approach to investigate the utility of emergent behaviors as an outcome measure in a behavioral-disturbance clinical trial, and we are in the process of determining its applicability for cognitive measures. We will also use this approach to study short-term changes and stability in behavioral symptoms on the BRSD in this patient population.

Acknowledgments

Work was supported by Grant AG 10483 from the National Institute on Aging.

REFERENCES

1. Bereiter, C. Some persisting dilemmas in the measurement of change. In: Harris, CW., editor. *Problems in Measuring Change*. Madison, WI: The University of Wisconsin Press; 1963. p. 3-20.
2. Overall JE, Woodward JA. Unreliability of difference scores: A paradox for measurement of change. *Psychol Bull.* 1975; 82:85–86.
3. Streiner, DL.; Norman, GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 2nd Ed.. Oxford, UK: Oxford University Press; 1995.
4. Green CR, Marin DB, Mobs RC, et al. The impact of behavioral impairment on functional ability in Alzheimer's Disease. *Int J Geriatr Psychiatry.* 1999; 14:307–316. [PubMed: 10340193]
5. Levy ML, Cummings JL, Fairbanks LA, et al. Longitudinal assessment of symptoms of depression, agitation, and psychosis in 181 patients with Alzheimer's Disease. *Am J Psychiatry.* 1996; 153:1438–1443. [PubMed: 8890677]
6. Rasmusson DX, Carson KA, Brookmeyer R, et al. Predicting rate of cognitive decline in probable Alzheimer's Disease. *Brain Cogn.* 1996; 31:133–147. [PubMed: 8811989]
7. Schneider LS, Olin JT, Doody RS, et al. Validity and reliability of the Alzheimer's Disease Cooperative Study—Clinical Global Impression of Change. *Alzheimer Dis Assoc Disord.* 1997; 11:S22–S32. [PubMed: 9236949]
8. Koss E, Weiner M, Ernesto C, et al. Assessing patterns of agitation in Alzheimer's disease patients with the Cohen-Mansfield Agitation Inventory. *Alzheimer Dis Assoc Disord.* 1997; 11:S45–S50. [PubMed: 9236952]
9. Patterson MB, Mack JL, Mackell JA, et al. A longitudinal study of behavioral pathology across five levels of dementia severity in Alzheimer's Disease: The CERAD Behavior Rating Scale for Dementia. *Alzheimer Dis Assoc Disord.* 1997; 11:S40–S44. [PubMed: 9236951]
10. Tariot PN, Mack JL, Patterson MP, et al. The CERAD Behavioral Rating Scale for Dementia. *Am J Psychiatry.* 1995; 52:1349–1357. [PubMed: 7653692]
11. Cohen-Mansfield, J. *Instruction Manual for the Cohen-Mansfield Agitation Inventory (CMAI)*. Rockville, MD: The Research Institute of the Hebrew Home of Greater Washington; 1991.
12. Mack, JL.; Patterson, MB. *Manual: CERAD Behavior Rating Scale for Dementia*. 2nd Ed.. Durham, NC: Consortium to Establish a Registry for Alzheimer's Disease; 1996.
13. Mack JL, Patterson MB, Tariot PN. The Behavior Rating Scale for Dementia (BRSD): Development of test scales and presentation of data for 555 individuals with Alzheimer's disease. *J Geriatr Psychiatry Neurol.* 1999; 12:211–223.

14. Desmond DW, Taremini TK, Stem Y, Sano M. The determination of clinically meaningful cognitive decline: Development and use of an alternative method. *Arch Clin Neuropsychol*. 1995; 10:535–542. [PubMed: 14588907]
15. Stern RG, Mohs RC, Davidson M, et al. A longitudinal study of Alzheimer's disease: Measurement, rate and predictors of cognitive deterioration. *Am J Psychiatry*. 1994; 151:390–396. [PubMed: 8109647]
16. Jacobson NS, Truax P. Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol*. 1991; 59:12–19. [PubMed: 2002127]
17. Ferris SH, Mackell JA, Mohs R, et al. A multicenter evaluation of new treatment efficacy instruments for Alzheimer's disease clinical trials: Overview and general results. *Alzheimer Dis Assoc Disord*. 1997; 11:S1–S12. [PubMed: 9236947]
18. McKhann G, Drachman D, Folstein M, et al. Clinical Diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*. 1984; 34:939–944. [PubMed: 6610841]
19. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975; 12:189–198. [PubMed: 1202204]
20. Matthews JNS, Altman MJ, Royston P. Analysis of serial measurements in medical research. *BMJ*. 1990; 300:230–235. [PubMed: 2106931]
21. Overall JE, Gorham DR. The Brief Psychiatric Rating Scale. *Psychol Rep*. 1962; 10:799–812.

Table 1
Frequency-Rated BRSD Items over 1 Year: Proportion of Absent or Clinically Meaningful Change

Item	Absent	Cease	Abate	Emerge	Intensify
1. Feelings of anxiety	59.8%	10.6%	4.0%	7.5%	10.1%
2. Physical signs of anxiety	47.3	11.4	2.0	10.9	10.9
3. Sad appearance	50.5	8.0	2.0	6.0	20.0
4. Feelings of hopelessness	73.5	5.6	2.0	4.6	5.6
5. Crying	69.7	5.5	5.5	4.5	11.9
6. Feelings of guilt	92.9	2.6	0.5	1.5	1.0
7. Poor self-esteem	74.5	8.2	3.1	5.6	4.1
8. Feels life is not worth living	90.4	2.5	0.5	3.0	1.5
11. Tiredness	32.3	15.4	4.5	10.4	10.0
13. Trouble falling asleep	63.5	11.2	1.5	11.2	4.6
16. Excessive physical complaints	83.4	7.5	0.5	4.5	2.5
18. Sudden changes in emotion	83.0	8.0	0.0	6.0	0.5
19. Agitation	40.3	11.4	6.5	7.5	15.9
20. Irritability	45.5	8.5	4.0	11.5	14.5
21. Uncooperativeness	53.2	6.0	0.5	15.4	12.9
22. Verbal aggression	85.1	4.5	0.5	5.5	1.0
23. Physical aggression	89.6	2.5	0.0	5.0	1.0
24. Restlessness	39.8	11.9	1.5	15.4	10.0
25. Purposeless behavior	30.0	14.1	3.5	11.6	14.6
27. Wandering	77.5	8.5	0.5	8.0	3.0
28. Trying to leave home	89.4	5.5	0.0	4.0	0.5
29. Socially inappropriate behavior	79.9	5.5	0.5	8.0	3.5
30. Repetitiveness	23.2	15.2	3.5	7.1	10.1
31. Social withdrawal	62.3	12.1	2.5	6.0	6.5
33. Misidentification of people	74.1	9.3	0.0	6.7	6.2
34. Doesn't recognize self in mirror	87.0	2.7	0.0	5.4	4.3
35. Misidentification of things	85.0	2.6	1.0	6.7	1.6
36. Feeling threatened	78.6	6.0	2.0	8.0	1.5

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Item	Absent	Cease	Abate	Emerge	Intensify
37. Belief that spouse is unfaithful	96.4	1.8	0.0	0.6	0.6
38. Belief that one is being abandoned	93.8	2.6	0.0	2.1	1.0
39. Belief that spouse is an imposter	91.8	3.1	0.0	3.6	1.5
40. Belief that TV characters are real	84.4	5.7	0.5	3.6	3.1
41. Belief that people are in house	70.6	7.6	1.5	11.2	5.1
42. Belief that dead person is alive	75.5	6.1	0.5	8.2	4.1
43. Belief that house is not home	79.6	5.8	0.0	5.8	6.3
44. Auditory hallucinations	85.6	5.2	0.5	5.2	2.6
45. Visual hallucinations	83.8	2.0	1.0	7.1	3.5