Capturing the Patient’s View of Change as a Clinical Outcome Measure

David Fischer, MD
Anita L. Stewart, PhD
Daniel A. Bloch, PhD
Kate Lorig, DrPH
Diana Laurent, MPH
Halsted Holman, MD

E specially in the past 2 decades, the growing prevalence of chronic disease has spurred the development of methods to measure a patient’s health state and its changes resulting from disease or therapy. Such methods have been necessary because the clear outcomes of recovery or death, historically applied to acute disease, no longer suffice. Chronic disease unfolds over time with an undulating course, and available treatments have varying consequences. Thus, midcourse corrections in a patient’s management are the rule and accurate assessments of changes in the patient’s health state are necessary to guide those corrections.

Many variables can contribute to understanding what is happening to a patient: biological markers, physical and emotional symptoms, technological images, observed behaviors or functioning, and patient perceptions. Because there is often a discrepancy between seemingly objective biological or imaging data and the patient’s symptoms or functioning (eg, serological data and the patient’s clinical state in rheumatic disease, lumbosacral x-rays, and low back symptoms), instruments have been designed to assess the patient’s health state in terms of comfort, physical and emotional symptoms or function, and activities of living. To make the data comparable across clinical trials, instruments have used a question format that, when repeated, yields serial data similar to that used in physical measurement: patients are asked to complete scales at different times and the difference between 2 points in time is considered to represent the change. In contrast to clinical trials, clinical practice most commonly uses a different evaluation method, namely, a retrospective appraisal by the patient and physician of what happened through a single, integrative assessment such as perceived direction and magnitude of change or degree of satisfaction with the changes. In general, it has been assumed that change inferred from serial measurements is more accurate than the patient’s retrospective perceptions of the

Context Measurement of change in patients’ health status is central to both clinical trials and clinical practice. Trials commonly use serial measurements by the patients at 2 points in time while clinicians use the patient’s retrospective assessment of change made at 1 point in time. How well these measures correlate is not known.

Objective To compare the 2 methods in measurement of changes in pain and disability.

Design Longitudinal survey of patients starting new therapy for chronic arthritis in 1994 and 1995. Surveys were completed at baseline (before intervention) and at 6 weeks and 4 months.

Setting Community health education program and university medical and orthopedic services.

Subjects A total of 202 patients undertaking self-management education (n = 140), therapy with prednisone or methotrexate (n = 34), or arthroplasty of the knee or hip (n = 28).

Main Outcome Measures Concordance between serial (visual analog scale for pain and Health Assessment Questionnaire for disability) and retrospective (7-point Likert scale) measures, sensitivities of these measures, and their correlation with patients’ satisfaction with the change (7-point Likert scale).

Results When change was small (education group), serial measures correlated poorly with retrospective assessments (eg, r = 0.13-0.21 at 6 weeks). With greater change, correlations improved (eg, r = 0.45-0.71 at 6 weeks). Average agreement between all pairs of assessments was 29%. Significant lack of concordance was confirmed in all 12 comparisons by McNemar tests (P = 0.02 to <.001) and by t tests (P = .03 to <.001). Retrospective measures were more sensitive to change than serial measures and correlated more strongly with patients’ satisfaction with change.

Conclusion The 2 methods for measuring health status change did not give concordant results. Including patient retrospective assessments in clinical trials might increase the comprehensiveness of information gained and its accord with clinical practice.

JAMA. 1999;282:1157-1162
www.jama.com

©1999 American Medical Association. All rights reserved.
change and the latter has been rarely measured in clinical trials.

Our study asked whether serial measures and the patient's retrospective assessment agreed when measuring the same change in a patient's health state. The serial method assessed the subject's level of pain and disability at 2 points in time and determined change by subtraction. The retrospective method, applied at the second point in time and often called a transition question, asked the patient's perception of the magnitude of the change that had occurred. Agreement between the 2 would imply that the methods provide the same information about the change while disagreement would suggest that different information is being obtained. We then explored the relative sensitivity of the 2 measures to change in health state and the relationship of each of the 2 measures to the patient's satisfaction with the change.

**METHODS**

We surveyed samples of 3 groups of volunteers undergoing interventions for arthritis: (1) participants in the Arthritis Self-Management Program (ASMP), (2) patients starting prednisone or methotrexate therapy for inflammatory arthritis, and (3) patients undergoing joint replacement of the knee or hip. These groups were chosen as representative of low-, medium-, and high-intensity interventions expected to yield small, moderate, and large changes in health. The study was approved by the Stanford Administrative Panel on Human Subjects in Medical Research.

The ASMP consists of 6 weekly, 2-hour sessions of patient education to foster self-management skills. Details of the program are published elsewhere. In aggregate, patients with chronic arthritis participating in the ASMP experience a 15% to 20% reduction in pain, a significant increase in their perceived self-efficacy to cope with the consequences of chronic arthritis, and a substantial decline in use of medical services. These effects have been found to persist for at least 4 years. The participants receiving methotrexate or prednisone (the medication group), who had not taken these medications previously, experienced inflammatory synovitis related to rheumatoid arthritis or a seronegative spondyloarthropathy. The patients undergoing hip or knee arthroplasty, who were recruited from an orthopedic surgery clinic, had a variety of diagnoses but most commonly osteoarthritis and were undergoing the surgery for pain relief or to improve function. Patients undergoing these interventions who could read English were encouraged to participate.

In all groups the initial survey questionnaires were completed by mail prior to the intervention. In the ASMP group, the follow-up surveys were sent to participants 6 weeks and 4 months after initiation of the course. In the medical and surgical groups, the follow-up questionnaires were sent at 6 weeks and 4 months after the intervention date (either initiation of medication or surgery).

In 1994 and 1995, 262 patients were recruited to the study. Enrollment consisted of completing the first questionnaire. If subsequent questionnaires were not returned, the patients were reminded by telephone. Since our interest was the comparison of retrospective assessment of change with the serial changes on instruments, patients served as their own controls and those who did not complete at least 2 of 3 questionnaires (23% of initial sample) were not included in the study. Completers and noncompleters were compared on baseline attributes of age, sex, education level, pain, and disability. The only significant difference was that, in the medication group, the completers had more education than the noncompleters.

Pain was measured using a 0- to 10-cm visual analog scale. Disability was measured using the Health Assessment Questionnaire, which was transformed to a 0 to 10 scale. Both were assessed at baseline and at the follow-up intervals when patients were asked to score their pain during the past 2 weeks. The change in the scores from baseline to follow-up constituted the serially measured change.

Patients were also asked at each follow-up about their perception of the magnitude of change in pain and physical limitations (disability) since the beginning of the self-management program or the initiation of treatment. This retrospective assessment was formulated on a 7-point Likert scale anchored on the left by "very much worse" and on the right by "very much better," with the middle point labeled "no change." Following each of these questions, patients were asked about their satisfaction with the change using a 7-point Likert scale ranging from "not at all satisfied" to "extremely satisfied."

To assess the agreement between the amounts of change in pain and disability as measured serially and retrospectively, Spearman rank correlations were computed for each patient's pair of serial and retrospective measures of change at 6 weeks and at 4 months.

To evaluate the sensitivity to change of the 2 measures, we used the concept of efficiency defined by Anderson and Chernoff. For serial measurement, each group of patients had pretreatment and posttreatment measures. For calculation, we used $E = d/SDd$, in which $E$ is efficiency, $d$ is the mean change in the measure for the group, and $SDd$ is the SD of the change measures. For the retrospective assessments that only have a posttreatment value, we considered efficiency to be the mean absolute difference from no change divided by the SD of the difference. Efficiency is not dependent on the sample size. Higher efficiency of a measure means more power to detect a shift from baseline and thus better sensitivity to change.

To assess in greater detail how well the 2 measures of change correlate, contingency tables were constructed of change scores for each of the 12 follow-up data points (eg, for each treatment group, the change in pain and disability at 6 weeks and at 4 months). By McNemar analysis, we determined whether one measure would yield significantly higher change values than the other. By $t$ test analysis, we determined the probability that the num-
number of individuals whose change assessments differed by 2 or more positions on the contingency table could have arisen by chance. A difference of 2 or more positions on the contingency table axes has substantial clinical meaning; it implies a clinical difference comparable with that between “unchanged” and “much better or worse,” or between “somewhat worse” and “somewhat better.”

To explore the relevance of the 2 change measures to the patient, levels of satisfaction with the change were correlated with the amount of change assessed by serial and retrospective measures. The group undergoing the least intensive intervention (ASMP) was deemed most likely to have stable health. Therefore, a group of these patients was recruited to repeat the follow-up questionnaire after an interval of 1 week in a test of the reproducibility of the answers.

RESULTS

The baseline characteristics of the groups are shown in Table 1. As expected, there were lower (better) average baseline pain and disability scores for patients primarily with osteoarthritis who were not having surgery (the education group), and higher scores (worse) for patients with inflammatory arthritis who required potent medications or patients who needed arthroplasty. The mean baseline value in each instance lies near the center of the scale and accords with known average scores for patients with osteoarthritis and rheumatoid arthritis. There was no clustering of scores at either scale extreme (eg, floor and ceiling effects) that would interfere with recording of meaningful change.

The changes in pain and disability by serial and retrospective measures are shown in Table 2 and Table 3. Percentage of change is calculated from baseline or the no change point to maximum possible improvement or worsening. The means by both measures at 6 weeks and 4 months form a similar pattern. That is, the education group generally had the smallest change in both measures followed by the medication group and the surgery group. For example, judged by serial measures, the medication group improved in disability at 6 weeks by 30% and by 32% at 4 months. While at 6 weeks postoperatively, the arthroplasty group showed an increase in disability presumably reflecting incomplete recuperation from surgery, by 4 months their disability scores had improved 40% over baseline. Mean retrospective percentage changes were consistently higher than those of serial measures.

Table 4 shows the correlations between the retrospective and serial changes at 6 weeks and 4 months. The correlations are low in the education group and higher but substantially less than 1 for the higher intensity intervention groups.

Table 5 shows results of the sensitivity calculations; retrospective perceptions of change were more sensitive than serial change for all categories. The calculations were made from 4-month data after the surgical patients had stabilized. The sensitivity of retrospective change judgment was approximately double that of serial change measurement for all 3 groups in both
pain and disability. The presence of some P values of more than .05 can be attributed to small sample size.

Contingency tables were created for all 12 comparisons (3 groups at 6 weeks and 4 months by 2 outcomes) using the amounts of change identified by serial and retrospective measures. TABLE 6 and TABLE 7 are examples of these comparisons. The axes represent the serial and retrospective measures. The numbers on the axes show the actual instrument values transposed to a 0 to 5 sequence. The middle position on each axis represents little or no change. The results of the contingency table analyses were similar for all 12 data collections. The on-diagonal agreement, which means agreement between the measures, ranged in the 12 cases from 0% to 50% with an average of 29%. The McNemar analyses showed that the retrospective assessment gave higher values for change than the serial measure in every instance, with 2-sided values ranging from $P = .02$ to $P < .001$. The t test analyses showed that the number of individuals who differed by 2 or more positions on the axes could not have arisen by chance in any of the 12 data collections, with 2-sided values ranging from $P = .03$ to $P < .001$.

Correlations of patient satisfaction vs serial change and retrospective change scores at 6 weeks are shown in TABLE 8. Patient satisfaction with the amount of change was more strongly correlated with retrospective change than with serial change for all 3 groups. Similar results were obtained at 4 months.

It is possible that patient attributes or the patients’ expectations about treatment effects could significantly influence retrospective change perceptions. Therefore, a subset of the patients in the education group ($n = 51$) were also asked about their expectations for improvement prior to the education program. Age, education level, disease duration, and expectations were only weakly correlated with retrospective change assessments ($r < 0.20$).

A subset of the patients in the education group ($n = 31$) repeated the follow-up questionnaires after 1 week. The test-retest correlations for the disability and pain scores were 0.88 and 0.85. The test-retest correlations for retrospective change in disability and pain were 0.81 and 0.58.

**Table 4. Correlation Coefficients (Spearman) of Serial Change Scores With Retrospective Change Scores**

<table>
<thead>
<tr>
<th>Patient Group</th>
<th>Correlations of Pain Measurement</th>
<th>Correlations of Disability Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 Weeks</td>
<td>4 Months</td>
</tr>
<tr>
<td>Education</td>
<td>0.21</td>
<td>0.40</td>
</tr>
<tr>
<td>Medication</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>Arthroplasty</td>
<td>0.67</td>
<td>0.51</td>
</tr>
</tbody>
</table>

**Table 5. Sensitivity to Change Calculations**

<table>
<thead>
<tr>
<th>Patient Group</th>
<th>Sensitivity (SE)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Serial Change</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>0.18 (0.08)</td>
</tr>
<tr>
<td>Disability</td>
<td>0.23 (0.08)</td>
</tr>
<tr>
<td>Medication</td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>0.66 (0.21)</td>
</tr>
<tr>
<td>Disability</td>
<td>0.66 (0.17)</td>
</tr>
<tr>
<td>Arthroplasty</td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>1.2 (0.23)</td>
</tr>
<tr>
<td>Disability</td>
<td>1.0 (0.24)</td>
</tr>
</tbody>
</table>

* Sensitivity for serial measurements equals $d$ divided by SD$\hat{d}$ where $d$ is the mean change of the measurement and SD$\hat{d}$ is the SD of the change measurement. For retrospective measurements, sensitivity is the mean difference from no change of the coded change responses divided by the SD of the coded change responses.

**Table 6. Education Group Pain Measurements at 6 Weeks**

<table>
<thead>
<tr>
<th>Serial Measurements†</th>
<th>Retrospective Measurements‡</th>
<th>1 or 2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6 or 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 to 6 (Much worse)</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5 to 2</td>
<td></td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1 to −1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>−2 to −5</td>
<td></td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>−6 to −10 (Much better)</td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

* Percentage of agreement on the 2 measures is 38; $r = 0.20$. McNemar test of the probability that 1 type of measure is significantly higher than the other; $P < .001$. The probability that the percentage of pairs disagreeing by 2 or more categories could have equaled zero by chance was calculated by the t test; $P < .001$. Bolded numbers represent perfect correlation.

† Visual analog scale values of change.

‡ Likert values of change.

---

**Comment**

Instruments must be more than reliable and valid to measure health status and its changes effectively. They must be sensitive to change and the identified change must be relevant to the patient and/or the illness. Sensitivity assessments are fairly clear cut in that they record the consistency with which a measure detects a true change. Relevance determinations are more variable; they may be based on a change in a biological parameter, a physical or emotional function, a patient preference or satisfaction, or some combination of these. The important requirement is that the relevance standard be external to the measured item.

Our study found poor agreement between retrospective assessments and serial assessments. Even when compared in the presence of major clinical
change as represented by 2 positions on the contingency table axes, the 2 measures did not consistently coincide. Furthermore, retrospective assessments were more sensitive than serial assessments and correlated more strongly with patient satisfaction with change.

Our results agree in important respects with other studies of retrospective appraisals. They concur with those of others who found that retrospective assessments of change were larger than change derived from serial measurements, particularly with powerful interventions such as surgery. Also, many other studies comparing different health status measurements have found the patient’s retrospective assessment to be among the most sensitive to change of outcome indicators. While we have not evaluated patient assessment of change in a setting with a placebo group, results from a meta-analysis of placebo-controlled nonsteroidal anti-inflammatory drug trials showed that retrospective assessments of change are among the most sensitive of outcome variables.

If the serial and retrospective measures differ, how should the results and the difference be interpreted? While this study focuses on instrument differences when measuring the same change, interpretation is an important issue. There is a large body of research on understanding change scores, but clarity has not been achieved. A major clinical focus has been on identifying the minimal clinically important change. An average change of 0.5 on a 7-point scale, equivalent to a change of a little bit better or worse, has been found to be important to patients as a group, but the importance to any particular patient remains highly individual. Similar small amounts of change on other scales have appeared to be significant to groups of patients. However, it is not known which of the measurement methods is most accurate or even whether they are measuring precisely the same thing; a difference between instruments could merely reflect measurement error or be due to different perceptions of the meaning of change. Given this uncertainty, it is prudent to include more than 1 type of change measurement in a clinical outcome assessment.

Certain reservations concerning our results deserve discussion. First, in the follow-up questionnaires, the question about satisfaction with change followed the retrospective assessment of change, and the latter could have biased the satisfaction estimate. However, all estimates of satisfaction with changes in clinical practice or trials contain an intrinsic component of retrospective appraisal; interaction between perception of change and satisfaction cannot be avoided. Second, the scales measuring serial and retrospective changes were not the same and therefore the axes of the contingency tables cannot be precisely compared. Nonetheless, comparability was improved by transposing the scales used in the contingency tables to a 5-point form and by assigning the same clinical meaning to the 5 points on each axis. Third, retrospective assessments of change as an outcome measure in clinical studies have been avoided because of concern that patients cannot remember their baseline conditions, and that assessments may not accurately reflect the benefits or harms that occurred. Our data do not address these concerns directly. However, the attributes of the retrospective assessments obtained by others and by us are similar, indicating that the retrospective assessments are not capricious or random but rather are detecting a particular perception of outcome. Further, retrospective assessments have face validity and, as judged by their higher correlation with satisfaction, they appear to have greater convergent validity and greater relevance than serial measures. Thus, in this setting, retrospective assessment had reasonable psychometric properties. Finally, retrospective measures are sensitive to change. Therefore, they cannot be rejected as legitimate outcome appraisals.

It is interesting that the poorest correlation between serial and retrospective appraisals of change was found when change was small or modest. Particularly with chronic disease, small or modest changes occurring over time are the most common response to treatment, and appraisal of those changes is crucial to the management program. Is the finding of discordance between the 2 measures unexpected? Not

### Table 7. Arthroplasty Group Disability Measurements at 4 Months

<table>
<thead>
<tr>
<th>Serial Measurements†</th>
<th>Retrospective Measurements‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or 2 (Much Worse)</td>
<td></td>
</tr>
<tr>
<td>10 to 6 (Much worse)</td>
<td>0</td>
</tr>
<tr>
<td>5 to 2</td>
<td>0</td>
</tr>
<tr>
<td>1 to −1</td>
<td>0</td>
</tr>
<tr>
<td>−2 to −5</td>
<td>0</td>
</tr>
<tr>
<td>−6 to −10 (Much better)</td>
<td>0</td>
</tr>
</tbody>
</table>

*Percentage of agreement on the 2 measurements is 7; \( r = 0.38 \). McNemar test of the probability that 1 type of measurement is significantly higher than the other; \( P = .001 \). The probability that the percentage of pairs disagreeing by 2 or more axis categories could have equaled zero by chance was calculated by the \( t \) test; \( P = .001 \). Bolded numbers represent perfect correlation.

†Likert values of change.

‡Visual analog scale values of change.

©1999 American Medical Association. All rights reserved.
fundamentally because clinicians have long known that the results of randomized clinical trials, based on serial measures, are often difficult to apply successfully to individual patients. As methodologists have sought to gauge health outcomes, they have increasingly identified the importance of patient views. In rheumatology, which deals with prototypic chronic diseases, the issue has been recognized for some time and patient retrospective assessments of change have been used in some instruments.

In 1 study, the retrospective estimates of change were found to correlate much higher with physical and biological indicators of change in disease state than did serial measurements. Why should the measures differ? One can speculate that each serial measure report is sharply focused on the precisely defined variable at a moment in time whereas the retrospective measure captures to some extent the patient's general experience of a change in symptom or health state over time. In the latter case, the symptom magnitude and some of its consequences (eg, the amount of pain together with its effects on physical function and pleasure) mingle in the assessment. That would move the assessment away from temporal precision toward a more composite appraisal over time, but it would also move it toward greater relevance to the patient.

The issue of relevance raises an important question: what are the clinical implications of obtaining the patient's assessment? Taking the patient's views into account is associated with greater satisfaction with care, with better compliance with treatment programs, and with maintenance of continuous relationships in health care. These are ample reasons for including the patient's retrospective assessment of outcome in clinical studies.

In conclusion, the retrospective assessments appear to provide information that is different from serial change data, are more sensitive to change, and are more highly correlated with patient satisfaction. They may ultimately be found to be independent outcome measures. However, the results of this study argue not for replacing serial measures but rather for supplementing them with the patient's retrospective appraisal whenever the study results are likely to be applied in clinical practice.

Funding/Support: The research reported herein was supported by grant AR20610 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Md.

REFERENCES


©1999 American Medical Association. All rights reserved.