Interpreting Changes in Quality-of-Life Score in *N* of 1 Randomized Trials

Roman Jaeschke MD, Gordon H. Guyatt MD, Jana Keller BSc, and Joel Singer PhD

Department of Clinical Epidemiology and Biostatistics (G.H.G.; J.K.; J.S.), Department of Medicine (R.J.; J.K.), Department of Family Medicine (J.S.), McMaster University, Hamilton, Ontario, Canada

ABSTRACT: To provide additional evidence regarding the plausible range of the differences in health-related quality of life (HRQL) questionnaire scores within which the minimal important difference (MID) falls, we reviewed the results of 32 randomized controlled trials in individual subjects (*N* of 1 RCTs) with chronic diseases. These trials had been conducted to establish whether a patient was obtaining more good than harm from a medication. Each *N* of 1 RCT included a series of pairs of treatment periods, one period on active drug, and one on placebo or alternative drug. We examined the relationship between small (MID), medium, and large differences between periods within pairs, as indicated by Global Ratings and differences between these same periods according to HRQL questionnaires. The results showed a mean difference of 0.29 points per question in HRQL questionnaire scores corresponded to the MID. Differences of approximately 0.66 points per question corresponded to a moderate difference as ranked by the Global Rating; difference of about 1.09 points per question represented marked difference.

KEY WORDS: *Measurement, health-related quality of life*

INTRODUCTION AND REVIEW OF PREVIOUS WORK

The clinician wanting to include health-related quality of life (HRQL) measures in assessing the effectiveness of therapy faces several challenges. These include deciding which aspects of HRQL should be measured, which measurement instruments should be used, and last but not least, how to interpret the results and communicate them in a meaningful fashion to other clinicians.

The effect of any treatment should be expressed in terms of both statistical significance and clinical relevance. Meeting criteria of statistical significance carries no guarantee that the differences observed are large enough to mandate treatment or that the outcomes that result are important. When discrete events are the measures of outcome (mortality or adverse, well-defined events), the effect of a treatment may be translated into number of lives saved, number

Address reprint requests to: Dr. Gordon Guyatt, Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre, Room 2C12, 1200 Main Street West, Hamilton, Ontario, Canada, L8N 3Z5.

Received July 11, 1990; revised March 25, 1991.

0197-2456/91/$3.50

Controlled Clinical Trials 12:226S–233S (1991)

655 Avenue of the Americas, New York, New York 10010

of strokes avoided, cost of hospitalization saved, etc. In these circumstances the significance of the treatment effects is similar for the physicians, patients, patients' families, and society.

For most HRQL measures the interpretation is much more difficult. Measuring HRQL involves analysis of patients' subjective assessment of their level of physical dysfunction or psychologic discomfort. If one finds a mean change of 0.4 cm on a 10-cm visual analogue scale measuring tiredness, does this constitute a large difference, or a clinically trivial difference [1]? Is a mean difference of 0.4 points per question on a 5-point Likert scale measuring the severity of the same symptom worth continuation of treatment [1]?

Translating changes in a HRQL instrument score into clinically meaningful terms is clearly crucial in the interpretation of study results. This is true for quantifying both the minimal important difference as well as larger effect sizes. The minimal important difference (MID) can be defined as the smallest difference in score in the domain of interest that patients perceive as a change and that would mandate, in the absence of troublesome side effects and excessive cost, modification in the patient's management. While the clinician would participate in the decision regarding modification of management, the definition otherwise focuses on the patient's experience. This follows from a conceptual or philosophical perspective that sees quality of life, including HRQL, as part of an individual's subjective experience. A person's capabilities (for example, whether or not it is possible to climb stairs) are very likely to bear directly on his or her HRQL. Measurement of capabilities, or HRQL estimates by others, may, in some cases, be used as surrogate measures of HRQL. The ultimate definition, however, should still focus on the experience of the individual.

We have recently reported our observations from three studies examining the relationships between patients' perception of changes in symptom severity and changes on a HRQL questionnaire [2]. In these studies of patients with chronic airflow limitation or with congestive heart failure, the primary outcome was a HRQL questionnaire measuring severity of dyspnea, fatigue and emotional function. Response options were presented using seven point scales. At the time each questionnaire was administered the patients were asked about their perception of overall change in dyspnea, fatigue, and emotional function. We found that a change in questionnaire score of approximately 0.5 points per question corresponds to the MID. The data suggested that changes of approximately 0.9 and 1.2 points per question corresponded to moderate and large changes in the overall perception of symptom severity [2]. One of the limitations of our report was that the data were from studies using a single questionnaire and examining similar populations.

In this paper we present data from 32 double-masked *N* of 1 randomized controlled trials in individual patients (*N* of 1 RCTs) conducted between 1985 and 1988 [3]. The purpose of this analysis was to determine if the conclusions drawn from the previous study would be confirmed in different clinical settings and with different questionnaires. We were also interested in examining the validity of the Global Ratings in the current study (which asked the patients to estimate the magnitude of preference between two treatment periods) and the Global Ratings in the previous study (which asked patients to estimate the magnitude of the difference in how they felt between two periods of time).

METHODS

General Concept of the Study

The *N* of 1 RCTs were designed to examine the efficacy of specific treatments in ameliorating symptoms due to a variety of conditions. The primary outcome measure in each N of 1 RCT was a HRQL questionnaire measuring the severity of symptoms identified by patients as related to their disease and important in their day-to-day life. Response options were presented using 7-point scales. The operational definition of the MID used is the smallest difference that is important enough that patients would choose to continue with the treatment indefinitely. It was postulated that the change in the questionnaire score corresponding to patients' subjective perception of small, moderate, and large benefit would approximate values found in our previous study [2].

Conduct of Individual *N* of 1 RCT

To assess *drug efficacy* in the *N* of 1 RCT, an individualized questionnaire examining the severity of symptoms identified by patients as part of their disease and as being important in their day-to-day life, was constructed. The specific symptoms to be measured were obtained from detailed interviews with patients. Symptoms chosen for detailed study from among those identified by the patient were restricted to those that were thought most likely to be influenced by the medication under study. The questionnaire that was developed on the basis of the interview consisted of four to seven items (symptoms) with the severity of symptoms measured on a 7-point scale. For example, if difficulty falling asleep was a symptom, the patient was asked:

Please indicate how much difficulty you had falling asleep during the previous two or three days, by choosing one of the options from the scale below:

1. Extreme difficulty
2. Very large amount of difficulty
3. Quite a bit of difficulty
4. Moderate difficulty
5. Mild difficulty
6. A little difficulty
7. No difficulty

The trial design was based on pairs of active/placebo, high dose/low dose or first drug/alternate drug combinations, the order of administration within each pair determined by random allocation. Treatment targets were monitored in a double-masked fashion on a regular, predetermined schedule throughout the trial. The difference in the mean score per question between active and placebo treatment for each treatment pair was established. Therapies were alternated until the clinician and patient agreed that they did not need more information to get a definite answer regarding the efficacy of the treatment, or until the patient or clinician decided for any other reason to end the trial.

To assess the *patient's* perception of drug benefit, the following questions were asked after each pair of treatment periods (these questions will be subsequently referred to as "Drug Guess"):

Overall in which of the two periods did you feel better?

1. First period
2. Second period
3. No difference

If the patient expressed a preference on the above question, he/she was then asked:

Would you continue the (puffer, pill, device) indefinitely if it was actually the (puffer, pill, device) that made you feel better?

1. Yes
2. No

When the patient answered yes to the above questions, he or she was asked to provide the magnitude of the drug effect by asking the following question (Global Rating):

If it turns out that you felt better during the period in which you were on the active (puffer, pill, device), we would like you to rate how important the difference between the two periods is to you:

1. Not important
2. Slight importance
3. Some importance
4. Moderate importance, consistent benefit
5. Much importance, consistent benefit
6. Very importance, good deal of benefit
7. Great importance

A Global Rating score of 0 was assigned if the patient indicated that there was no difference between periods, or if the observed difference was not sufficient to make him or her take the drug indefinitely. When the Global Rating score was 0, the corresponding difference in the questionnaire score had a positive sign if favoring the active drug, and negative if favoring placebo. For Global Ratings scores other than 0, the difference score on the HRQL questionnaire was assigned a positive value if it favored the period preferred according to the Global Rating, and assigned a negative value if it did not.

ANALYSIS

We examined the relationship between the patient's subjective assessment of drug efficacy (Global Rating) and differences in the quality-of-life questionnaire score in every pair of every *N* of 1 RCTs for which data were available. Because questionnaires included from four to seven questions, the difference in the quality-of-life questionnaire score was expressed as total difference in score divided by the number of questions in a particular questionnaire. The MID was defined as the difference in the questionnaire score corresponding to answers 1 to 3 on the Global Rating; moderate benefit as differences corresponding to answers 4 or 5; and large benefit as differences corresponding to answers 6 or 7.

Results are presented as mean differences on the HRQL questionnaire score corresponding to the categories small, moderate, or large degrees of importance expressed by patients across all trials. Because some subjects contributed several data points at the same level of Global Rating, we were concerned that if there were a dependency between the observations within a subject, subjects with multiple observations would have a greater influence on the comparisons than those with a single observation. To assess the dependency of repeated observations we have compared the mean-squared error between subjects to the mean-squared error within subjects at fixed levels of Global Ratings. We have also examined the data using only one data point per patient per level of Global Rating.

To compare the variability in each category of response (0, 1 to 3, 4 and 5, 6, and 7) from the previous study [2] to the current one, an F test was conducted. In addition, the number of directional errors (instances in which the Global Rating and HRQL score moved in different directions) were calculated both for the current and prior data and compared using a chi-square test.

RESULTS

The clinical settings of the 32 N of 1 RCTs are presented in Table 1. Patients completed Drug Guess on 110 occasions. In 11 cases they stated that there was no difference in their overall well-being between treatment periods; three stated the difference was not large enough to make them willing to continue the drug indefinitely. In 96 out of 110 treatment pairs (87%) the difference in the subjective health status experienced by the patients was large enough to make them willing to continue the drug indefinitely. In 84 (87.5%) of these cases the HRQL questionnaire score difference corresponded to the period during which the patient felt better according to the Global Rating, as compared to 93.6% in the previous study. The difference in the proportion in concurrence between the two studies (i.e., 87.5% versus 93.6%) is statistically significant (chi-square 3.99, $p=0.046$). The proportion of agreement was higher in the first study despite the fact that the proportion with a Global Rating of 1–3 was greater in the first study (31%) than in the current study (12%). One would expect the risk of disagreement to be higher for Global Ratings of 1–3 (rather than 4–7). Indeed, this proved to be the case: misclassification rates

Table 1 Clinical Settings of N of 1 RCTs

Condition	Drug	No. of Trials
Fibrositis	Amitryptilline	17
Fibrositis	Nitrazepam	2
CAL[a]	Ipratropium	5
CAL[a]	Prednisone	1
CAL[a]	Salbutamol	1
Myasthenia gravis	Pyridostigmine	2
Anxiety	Lorazepam	1
Rheumatoid arthritis	Clonidine	1
Addison's disease	Hydrocortisone	1
Syncope	Amitryptilline	

[a]Chronic airflow limitation.

in the prior study were 11.8%, 5.3%, and 0.0%, respectively, for Global Ratings of 1–3, 4–5, and 6–7.

The extent to which quality-of-life scores varied within a given Global Rating category was also examined. Of the four Global Rating categories ("no difference," "small," "medium," and "large" difference) in three categories there was a trend suggesting a larger variance of differences in HRQL score in the current study. For the "no difference" category, variances were statistically significantly greater in the current study [$F(13,422) = 2.32$, $p < 0.05$].

The mean differences on the quality-of-life questionnaire score corresponding to the different categories of the Global Rating are presented in Table 2. The differences on the quality-of-life questionnaire score increase with the Global Rating of the Drug Guess. A mean difference of 0.29 points per question in HRQL questionnaire score corresponds to the MID. Differences of approximately 0.66 points per question correspond to the moderate difference as ranked by the Global Rating; differences of about 1.09 points per question represent a marked difference. There is a large between-patient variability in the estimates of differences in HRQL questionnaire score corresponding to the Global Ratings. The comparison of between- to within-patient variability indicated that on two levels of Global Rating (0 and 6–7) the former was significantly larger. The results of the analysis using one data point per patient per level of Global Rating are presented in the last two columns of Table 2.

DISCUSSION

Our goal was to provide a clinically meaningful interpretation of changes in questionnaire scores used in the conduct of *N* of 1 RCTs. The initial hypothesis, based on prior work, was that changes in score of approximately 0.5, 0.80–1.0, and over 1.0 would correspond, respectively, to small, medium, and large effect sizes.

One potential source of confirmation concerning the clinical importance of changes in questionnaire scores is the impression of clinicians using these questionnaires in a clinical setting. Prior to our initial work, we had formed the impression that a change of 0.5 points per question approximated the MID. In the current study, clinicians generally interpreted differences of less

Table 2 Changes in the Quality of Life Questionnaire Score Corresponding to Varying Global Rating of Drug Value on Drug Guess

Global Rating Category	Hypothesized[a] Relationship from (2)	N[b]	Mean Score Difference (SE)[c] All Data	Mean Score Difference/SE[c] 1 Data Point
0	0.00	14	0.23 (0.26)	0.27 (0.33)
1–3	0.50	12	0.29 (0.25)	0.21 (0.23)
4–5	0.80–1.00	34	0.66 (0.11)	0.70 (0.19)
6–7	>1.00	52	1.09 (0.13)	0.97 (0.21)

[a]Based on previous study.

[b]Number of observations multiple data points from each patient.

[c]Standard error of the mean.

than 0.5 points per question as trivial. When differences between periods were greater than 1.0, clinicians perceived large and very important treatment effects.

There are two major differences between our prior work and the current study. First, in the initial work, the independent standard to which the questionnaire was related was a Global Rating of change from the previous visit. In the current study, patients were asked first to judge in which of two treatment periods they were better, and then to judge magnitude of the difference. Second, although the domains examined in the previous study were quite different (dyspnea, fatigue, and emotional function) the patients studied all used the same questionnaire and all had chronic airflow limitation. In the current work, patients had a variety of medical conditions, and were exposed to individualized questionnaires. With regard to the latter point, if similar 7-point scales could be interpreted in a similar fashion when used as response options in different questionnaires, interpreting the results of new instruments would be greatly facilitated.

The results are not inconsistent with the hypotheses, but provide only limited support. The magnitude of difference in score for each of small, medium, and large differences was smaller than might have been expected. Furthermore, the magnitude of the difference in questionnaire scores was as great in patients reporting no change as in those reporting small but clinically important differences. The finding raises questions about the validity of the comparison of the two periods, and of the symptom questionnaires themselves.

It is our impression that patients have much more difficulty making comparisons between two periods of time as opposed to making a single absolute rating of, for instance, the degree of change from a prior visit. This impression is supported by the trend toward greater between-patient variability in changes in HRQL score corresponding to varying effect sizes seen in the current investigation (in which small, moderate, and large effect sizes were defined according to a comparison between periods), in comparison with our previous work (in which a global rating of change defined the effect sizes). An alternative interpretation is that the greater variability in changes in HRQL score corresponding to small, moderate, and large effects is a result of the variety of questionnaires used, and the varying chronic diseases [2]. However, this alternative hypothesis would not explain the statistically significant greater number of directional discrepancies between the Global Rating and the HRQL questionnaire score in the current study.

Patients' difficulties in comparing two periods of time would suggest that there might be considerable random error in the rating of preference and magnitude of difference. Further, the number of patients was small, and the confidence intervals around the estimates of small, medium, and large effects was relatively wide. These confidence intervals include the estimates of MID, medium, and large differences from our previous study.

Although the confidence intervals include the previous estimates, the mean differences in questionnaire score corresponding to varying effect sizes was smaller than in the previous study. This may be because patients participating in *N* of 1 RCTs were expecting to see differences and thus more likely to report them, and overestimate their magnitude. This would not, however, explain the relatively large difference in mean score associated with the "no

important difference" group. This finding is to a considerable extent due to an anomalous result in a single patient.

A woman with fibrositis reported no difference in the control of her disease between periods, and yet there was a mean difference in questionnaire score of three points per question. This data point, being one of 14, contributed dramatically to the overall estimate of the "not important" difference. When this particular treatment pair was checked, it turned out that the large difference in the questionnaire score was due to a concurrent illness with symptoms that overlapped with fibrositis symptomatology (bursitis). The influence of this extreme data point was even larger when only one data point per patient was used.

In both the current and previous study, we observed large between–patient variability in the changes in symptom questionnaire score corresponding to varying estimates of drug efficacy. That is, within each category of rating of change (no difference, small, moderate, and large effect) the range of change in questionnaire score between patients was large. Some portion of this variability is certainly due to the less than perfect validity of the independent standard (in this case, the Global Rating of drug efficacy). However, it is likely that patients have different standards about the changes in symptoms that they view as important or trivial. Such variability in the clinical significance of a particular change in score or outcome is seen in physiologic as well as subjective measures. This suggests that the HRQL questionnaire results should not be used as the sole criterion of whether an individual patient has experienced important changes. On the other hand, establishing the range of changes in score that correspond to small, medium, and large effects across a group of patients remains crucial to allowing meaningful interpretation of study results. This report provides support for the plausible range within which the MID and larger effect sizes probably fall, while at the same time emphasizing the need for further work in this area.

REFERENCES

1. The German and Austrian Xamoterol Study Group: Double-blind placebo-controlled comparison of digoxin and xamoterol in chronic heart failure. Lancet 1:489–493, 1988
2. Jaeschke R, Singer J, Guyatt GH: Measurement of health status: Ascertaining the minimal clinically important difference. Controlled Clin Trials 10:407–415, 1989
3. Guyatt GH, Keller J, Jaeschke R, et al: Clinical usefulness of the N of 1 randomized control trials: Three year experience. Ann Intern Med 112:293–299, 1990
4. Guyatt G, Sackett D, Adachi J, et al: A clinician's guide for conducting randomized trials in individual patients. Can Med Assoc J 139:497–503, 1988