Eur Respir J 2002; 19: 398–404
DOI: 10.1183/09031936.02.00063702
Printed in UK – all rights reserved

European Respiratory Journal
ISSN 0903-1936

Interpreting thresholds for a clinically significant change in health status in asthma and COPD

P.W. Jones

Interpreting thresholds for a clinically significant change in health status in asthma and COPD. P.W. Jones. ©ERS Journals Ltd 2002.

ABSTRACT: Health status (or Health-Related Quality of Life) measurement is an established method for assessing the overall efficacy of treatments for asthma and chronic obstructive pulmonary disease (COPD). Such measurements can indicate the potential clinical significance of a treatment's effect.

This paper is concerned with methods of estimating the threshold of clinical significance for three widely used health status questionnaires for asthma and COPD: the Asthma Quality of Life Questionnaire, Chronic Respiratory Questionnaire and St George's Respiratory Questionnaire. It discusses the methodology used to obtain such estimates and shows that the estimates appear to be fairly reliable; *i.e.* for a given questionnaire, similar estimates may be obtained in different studies.

These empirically derived thresholds are all mean estimates with confidence intervals around them. The presence of these confidence intervals affects the way in which the thresholds may be used to draw inferences concerning the clinical relevance of clinical trial results. A new system of judging the magnitude of clinically significant results is proposed.

Finally, an attempt is made to translate these thresholds into scenarios that illustrate what a clinically significant change with treatment may mean to an individual patient.
Eur Respir J 2002; 19: 398–404.

St George's Hospital Medical School, London, UK.

Correspondence: P.W. Jones
St George's Hospital Medical School
London SW17 0RE
UK
Fax: 44 2087255955
E-mail: pjones@sghms.ac.uk

Keywords: Airflow obstruction
asthma
chronic obstructive pulmonary disease
quality of life
respiratory rehabilitation

Received: July 13 2000
Accepted after revision August 6 2001

Health status (or Health-Related Quality of Life) measurement is now an established method for assessing therapy for patients with chronic lung disease. It permits an estimate of the overall effect of treatment and can provide an indicator of the potential clinical significance of a treatment effect. A recent comprehensive review of health status measurement in chronic obstructive pulmonary disease (COPD) has been published elsewhere [1]. This paper is concerned with methods of estimating the threshold of clinical significance for health status instrument and how these estimates may be used to interpret the results of treatment.

What is a clinically significant threshold?

The term "clinically significant threshold" appears at first sight to be self-explanatory, but there has been no agreed definition of what it is, or who should judge clinical significance [2]. Several of the methods discussed in this paper do not actually measure a clinically important difference. They are rather closer to the "just-noticeable-difference" or JND used in psycho-physics. The JND is measured in individuals. Nevertheless, this type of approach has been applied to a population of patients to estimate the clinically significant difference in a health status score between patients [3]. Most studies of the threshold for clinical significance have not been carried out between patients, but have used a within-patient "just-noticeable-change". Another variant of this approach is the "just-noticeable-effect" of a treatment. When described in this way, it becomes clear that such methods do not identify a clinical threshold, but rather, they define a just detectable level of change or effect. This may be perfectly valid, since a change in clinical state that is detectable by the patient is probably worthwhile, but strictly speaking, such a change may or may not be "clinically significant".

Some studies have used clinician judgement of change as the criterion for setting a threshold, but it is not clear how the clinicians made their judgements or what factors they took into account when judging a "clinically significant change" or "clinically significant difference". On this point, there is evidence that the factors actually used by physicians when making judgements about disease severity may be quite different from the factors that they say are important [4]. To date, all attempts to identify a clinically significant threshold have used some form of patient or clinician judgement. However one study, designed for other purposes, does permit a test of the clinically significant threshold score for one widely used health status questionnaire; the St George's Respiratory Questionnaire (SGRQ). That study used criteria for clinical significance that were not based upon patient or clinician judgement [5]. Discussed in more detail in a later section of this paper, this study is perhaps the first true test of a clinically significant threshold.

Another important issue concerning the concept of the clinically important threshold is the assumption that there is one score for the threshold that applies to all patients. As already discussed, judgements by patients and clinicians provide a core methodology for establishing these thresholds, but making a judgement means assigning values. Patients will differ in the value they place upon the many disturbances of daily life and wellbeing that result from chronic lung disease, so there will be inter-individual differences in the threshold score for clinical significance. As a result, any study designed to identify a threshold score will produce a mean value from many individuals. In consequence, the threshold score will apply to a population, not to an individual. However, this is not the disadvantage that it may appear. All health status questionnaires treat each patient as if they were a "typical" patient and questionnaire items form a lowest common denominator of potential similarity between patients. Thus, each patient is assessed in a standard way. Since the questionnaires are population-based, it is reasonable to use threshold scores that are also standard for that population.

Background

Clinical thresholds are used most commonly to judge whether a treatment has a clinically worthwhile effect, or whether it is superior to another treatment. A good example is a study that compared salmeterol with regular four-times daily salbutamol in patients with asthma [6]. With salmeterol, there was a statistically significant improvement in health status gain from baseline of 0.49 units, measured using the Asthma Quality of Life Questionnaire (AQLQ; fig. 1). Salbutamol produced a smaller, but still statistically significant improvement of 0.27 units. To interpret this study, it is necessary to know that the suggested threshold for clinical significance using the AQLQ is 0.5 units [7]. Thus, the improvement with salmeterol almost reached the 0.5 unit threshold of clinical significance, but not quite. It was 2% below it. Does that mean that the effect of salmeterol was not clinically worthwhile? To answer this question one should ask how confident can one be, that the threshold value of 0.5 was estimated with absolute precision and no sampling error? This paper will address these issues, although in the reverse order in which these questions are posed.

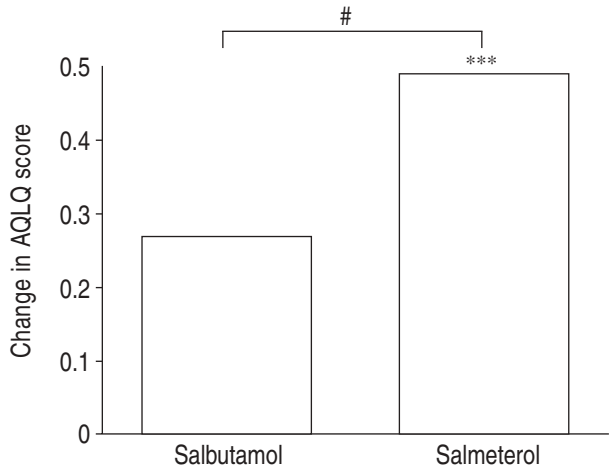

Fig. 1.–Changes in the Asthma Quality of Life Questionnaire (AQLQ) score with salmeterol compared with regular four-times daily salbutamol in patients (n) with asthma (adapted from [6]). The threshold for clincially significant change is equal to 0.5. ***: p<0.001 *versus* salbutamol; #: p=0.022.

Methods of assessing the clinically significant threshold

The development and validation of the threshold for clinical significance of three questionnaires widely used in clinical trials in asthma and COPD will be discussed. These are: AQLQ [8], Chronic Respiratory Questionnaire (CRQ) [9] and SGRQ [10]. As with all tests of the validity of a measurement instrument, determination of a threshold score for clinical significance requires the accumulation of a number of pieces of evidence, preferably obtained using different methodologies. Methods of establishing a clinical threshold will be discussed under three headings: 1) Clinician judgement; 2) Patient judgement; 3) Criterion referencing.

Clinician judgement

To establish a minimum clinically important difference (MCID) for the CRQ, JAESCHKE *et al.* [11] began with a discussion among a group of clinicians who had administered this questionnaire to patients in three clinical studies. A consensus was reached that most patients with an increase of three points on the dyspnoea domain had experienced a reduction in dyspnoea that was important to them in their daily lives, whereas most patients with an increase of one or two points had no change. Three points corresponds to a change of 0.6 per item in the "dyspnoea" domain of the CRQ. Similar discussions were held over the "emotional function" and "fatigue" domains of this questionnaire. As a result of these discussions, JAESCHKE *et al.* [11] established a hypothesis that the MCID approximated to a mean change of 0.5 per item for each of these domains of the questionnaire. The threshold for the "mastery" domain was not discussed in their report.

Clinician judgement was used also by JONES *et al.* [12] to provide an estimate of the threshold of significance for the SGRQ. They used a different technique to that of JAESCHKE *et al.* [11]. A small group of physicians and nurses experienced in pulmonary care were asked to imagine two populations of patients. They were then asked to judge what would constitute a minimum clinically significant difference between these two hypothesized populations using the following: frequency of cough; frequency of wheeze; level of dyspnoea in daily life; level of depression; 6-min walking distance. These differences, estimated individually, were then applied

simultaneously in a multivariate model that used data on these variables collected from a population of patients with COPD. The calculated difference in SGRQ score between these two hypothesized populations was 3.9 units. For convenience, this was rounded to 4 units.

Patient judgement

This usually takes the form of an overall assessment of disease severity or treatment efficacy. All workers in this field have used patients' global estimates of severity or response to treatment as a "gold standard", although this type of scale has never been validated properly. Furthermore, it is not known how patients make judgements about their disease severity or the efficacy of their treatment. There is evidence that these global assessments correlate quite well with health status questionnaires, although a recent study has shown that the nature of this relationship may depend on the way in which global questions are phrased [13]. Furthermore there is evidence of a "response shift" when patients use global questions to assess their asthma severity [13].

Clinical threshold for the Chronic Respiratory Questionnaire. Patients' judgements of their change in health were used by JAESCHKE *et al.* [11] to confirm the clinician judgements concerning the MCID. They used 55 patients who had participated in two clinical trials in COPD and 20 heart failure patients who contributed to a trial of patients in sinus rhythm. The Chronic Heart Failure Questionnaire (CHQ) was used for that study, but its content is almost identical to that of the CRQ. Multiple measurements were made and it appears that each patient contributed two measurements to the analysis. The changes in CRQ/CHQ over the study period were compared with the patients' retrospective global estimate of their change, measured using a 15-point transition scale. This contained seven categories of improvement, one of no change and seven of deterioration. The authors equated six categories of change on this scale to the threshold of clinical significance. These were: "Almost the same, hardly any better (or worse)", "A little better (or worse)", and "Somewhat better (or worse)". The mean changes in CRQ/CHQ scores corresponding to this threshold were: Dyspnoea score 0.43; Fatigue score 0.64; Emotional function 0.49. No threshold for the mastery component was reported. There were a total of 63 observations in the six change categories that the authors felt corresponded to an MCID. These were made in an unknown number of patients. No distribution statistics or tests of significance were given, but there was considerable variation. For example, the MCID estimated for the dyspnoea score varied between 0.28–0.62 across the three trials that contributed results to this analysis.

A later study by these authors used a different technique. They asked patients in a pulmonary rehabilitation programme to make judgements concerning the size of health status differences between themselves and their peers [3]. This is a between-patient MCID, rather than a within-patient just-noticeable change. The results were broadly similar to those obtained using the within-patient technique, although the between-patient MCID for the dyspnoea score was 0.09 in contrast to the score of 0.43 estimated for the within-patient MCID obtained earlier [11]. When the MCID for the emotions, fatigue and mastery components were pooled (*i.e.* excluding the dyspnoea component), the pooled MCID was 0.42 (95% confidence intervals (CI) 0.32–0.53).

In their original publication concerning the MCID for the CRQ, the authors noted that "The mean change in score per question corresponding to the MCID is consistently around 0.5…" [11]. Whilst this appears to be the case, it is quite clear that there is considerable variation around this value.

Clinical threshold for the Asthma Quality of Life Questionnaire. A similar approach to that just described for the CRQ was used for the AQLQ [7], although for this test the MCID was calculated using those AQLQ scores that corresponded to "A little better (or worse)" and "Somewhat better (or worse)". The degree of change "Almost the same, hardly any better (or worse)" was now judged to be equivalent to "No change", a change from the original technique used by JAESCHKE *et al.* [11]. The patients were 39 adults attending a clinic. They were followed-up twice, so each patient could contribute up to two pairs of observations. The AQLQ has four component scores, the number of observations contributing to the MCID ranged from 10–23 and the MCID for each of the components ranged (mean±SD) 0.47±0.51–0.58±0.56. The MCID for the Total AQLQ score was 0.52 units. The 95% CI around this mean value were 0.29 and 0.81 units.

The studies on the AQLQ and CRQ used very similar methodologies to assess the MCID. The number of patients contributing to these estimates was not large and there is clearly a lot of variation between patients in terms of the score that corresponds to the MCID. The mean value of 0.5 for the threshold is clearly a pragmatic value since the measured values varied between the different components of the questionnaires, and none had an estimated MCID that was exactly 0.50. JAESCHKE *et al.* [11] suggested that a MCID of 0.5 might be a property of seven-point scales of this type. The study by JUNIPER *et al.* [7] appears to support this conclusion, although as noted earlier, the methodology was changed slightly between the two studies. This evidence of predictive validity lends support to the use of 0.5 as an indicator of a clinically significant change, but this value chosen by the questionnaire's authors is clearly an estimate with scatter around it.

Clinical threshold for St George's Respiratory Questionnaire in asthma. Data from a 1-yr study of nedocromil in moderate asthma [14] was used in a similar way to that used for the CRQ and AQLQ. The major differences were the much longer duration of follow-up and the use of the patients' retrospective estimate of treatment efficacy on a five-point scale: 1) "Made me worse"; 2) "No effect"; 3) "Slightly

effective"; 4) "Moderately effective"; and 5) "Very effective". The difference in score measured at baseline and after 12 months of therapy was compared with the patients' retrospective estimate of treatment efficacy made at the end of the study. The patients were not shown their baseline SGRQ questionnaires (or their scores) when they made the second SGRQ assessment and their assessment of overall treatment efficacy. There was a rank order correlation between change in health status and overall judgement of treatment efficacy [15]. This is illustrated in figure 2. There was no change in score in the 108 patients who judged the treatment to be ineffective, but a mean 4.0 unit change (95% CI 1.6; 6.4 units) in the 97 patients who judged the treatment to be slightly effective.

Clinical threshold for St George's Respiratory Questionnaire in chronic obstructive pulmonary disease. A similar approach to that formerly described was used, albeit with different wording, in the assessment of treatment efficacy in a 16 week study of salmeterol in COPD [16]. In that study, treatment efficacy was assessed by the patients using a scale worded: "no effect"; "satisfactory", "effective, "very effective". The improvement in SGRQ score in the patients who judged the treatment to be "satisfactory" was 2.0 units (95% CI 0.2; 4.1 units; n=87). The term "satisfactory" is ambiguous because it does not convey a clear meaning about a detectable therapeutic effect. "effective" is the lowest response category on this scale that is clearly compatible with efficacy. In the 109 patients who scored the treatment as being effective, the mean improvement in SGRQ score was 4.3 units (95% CI 1.8; 6.9 units). "very effective corresponded to an improvement of 8.1 units (95% CI 4.7; 11.4 units; n=55).

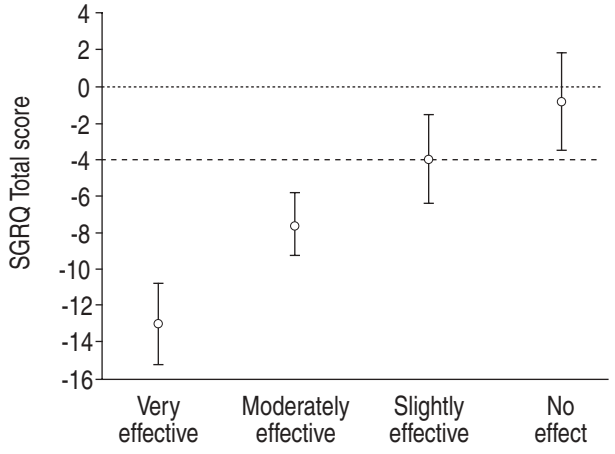

Fig. 2. – Changes in the St George's Respiratory Questionnaire (SGRQ) score following 1 yr of therapy with nedocromil sodium or placebo in patients with asthma, compared with the patients' retrospective estimate of the treatment's efficacy [14, 15]. The patients were blind to their previous and current SGRQ scores when they judged the efficacy of the treatment. Circles represent the mean; whiskers represent 95% confidence intervals. ·····: no change from baseline; - - -: the threshold for clinical significance. Very effective: n=139; Moderately effective: n=164; Slightly effective: n=97; No effect: n=108.

Criterion referencing

This is a method that compares a health status score to a criterion of health. Such criteria may be the occurrence of hospital admission, death, need for a major change in treatment or a clear change in clinical state of the patient. The principle underlying this approach is that health status questionnaire scores should be worse in patients who have major health events compared with those who do not.

In a prospective study, SGRQ scores were obtained from 238 patients as they were being discharged from hospital, following an acute exacerbation of COPD [5]. They were then followed for 1-yr. During that period, 109 patients were admitted to hospital or died, whilst 129 patients did not have one or other of these major health-related events. The difference in SGRQ score between these two groups at baseline was 4.8 units (95% CI 1.6; 8.0 units).

In another study in COPD, SGRQ scores were related to the Medical Research Council (MRC) dyspnoea grade [17]. In 32 patients who were housebound (MRC dyspnoea grade 5), the SGRQ scores were 3.9 units (95% CI 1.8; 9.4 units), worse than in 32 patients who had major impairment of daily activity due to dyspnoea, but were not housebound (MRC dyspnoea grade 4).

Summary

The threshold for the CRQ and AQLQ appears to be around 0.5. Similarly, the estimate for the SGRQ threshold is consistently around 4 units, regardless of the method of estimation and the number of the subjects contributing to the estimate. It is noteworthy that both teams of questionnaire developers have rounded their recommended threshold values to the nearest convenient number, rather than advise the use of a precise threshold value obtained from one particular study. These threshold scores provide an indicator of the change that is compatible with a clinically significant change for the patient.

Interpreting changes in health status measurements

Changes in questionnaire score following treatment have been related to the estimated threshold for clinical significance in a number of different ways.

Comparison with the average estimate for the clinical threshold

This is the simplest approach, although perhaps not the best. The mean treatment effect is compared with the threshold for clinical significance as illustrated in figure 1. If the score lies below the threshold, this benefit is deemed, by inference, to be subclinical. The limitations to this approach are clear following the preceding discussion of the methods of estimating these thresholds. Whilst they appear to be reliable,

i.e. similar thresholds are found in different settings, they are estimates and do have confidence intervals around them. The mean treatment effect of 0.49 in the study illustrated in figure 1 lies fractionally below the estimated clinically significant threshold of 0.5 recommended by the author of the AQLQ (although as noted earlier, the actual measured estimate of the threshold was 0.52). A response of 0.49 using the AQLQ units still lies well within the CI for the threshold. This result should be considered clinically significant.

As already described, there is variance in any estimate of a threshold for clinical significance, so it may be more appropriate to use the lower CI of the estimated threshold when making a judgement as to whether a treatment was clinically worthwhile. One problem with this approach is that the size of the CI depends upon the number of observations contributing to an estimate. The threshold for the overall score in AQLQ was estimated from ten observations [7] and calculation of the lower CI from the published data in that study produces a lower CI for the MCID of 0.29. If, however, there were 100 observations and the estimated threshold and its scatter were exactly the same (*i.e.* threshold estimate=0.52 with an SD of 0.41), the lower 95% CI would now be 0.48. This shows the limitations of using the lower CI for the estimated threshold as the cut-off point for judging whether a result is clinically significant. The cut-off point will be heavily dependent upon the number of observations contributing to its estimation. Quite clearly, the larger the number of observations, the more precise the estimate, but how many observations is enough?

Number of patients exceeding the threshold

Some patients within a population may have a worthwhile benefit from treatment, even though the mean effect on the population may not achieve the threshold for clinical significance. In an attempt to address this problem, an approach has been developed based upon a technique that is applied to the results of trials that use dichotomous outcomes, such as death *versus* survival or stroke *versus* no stroke. Such data allow the calculation of relative risk and risk difference between groups. These estimates can be used to permit calculation of the number of patients who would need to be treated (NNT) to prevent one event. This approach requires patients to be categorized into those in whom health gain exceeded the threshold for significance and those in whom it did not. The proportion of patients who achieved a clinically significant improvement in one limb of a study can then be compared with the proportion in the other. This method has been used to reanalyse data from a three-limbed crossover trial of salmeterol, regular salbutamol and placebo [18]. The number of patients who would need to be treated with salmeterol to produce a single patient with a clinically significant benefit compared to control treatment (*i.e.* the NNT) was 4.5.

As with any measurement made upon a sample of a whole population, NNTs are estimates that should have confidence limits around them. In the example quoted above, the 95% CIs for NNT, calculated from the reported data, are 4.1 and 8.3. In other words, the most favourable estimate of the NNT is 4, whereas the least favourable estimate is around 8. Note that the CIs for the NNT are not symmetrical around the mean, the upper CI for the NNT is rather further from the mean than the lower CI.

To test the influence of the threshold of clinical significance on the NNT when calculated using this method, a sensitivity analysis using data from a trial of salmeterol in COPD was carried out [16]. Using the established SGRQ threshold of 4 units, the NNT for salmeterol was 4.5. When the threshold was changed to 5 units, the NNT was still 4.5. When the threshold was set to 3 units, the NNT was 5.0. In fact, the NNT appeared to be very stable, regardless of the clinical threshold. When the SGRQ threshold was set as low as 1 unit, the NNT for salmeterol was still only 5.2. This analysis may be interpreted to suggest that the NNT is independent of the exact value of the threshold for clinical significance. Another view would be that this method of examining clinical trial results just reflects the proportion of patients who responded to the treatment compared to placebo, but tells us nothing about whether that response was clinically significant.

Confidence interval method of judging therapeutic response

The existing methods of assessing the clinical relevance of a therapeutic response have shortcomings. An alternative method that overcomes some of these limitations is proposed here. In this approach, the size of treatment response in relation to the threshold for clinical significance is categorized using both the mean value for the treatment effect and its CIs. The different categories of response are described later and illustrated in figure 3.

No effect. This occurs when the lower CI of the treatment effect crosses zero. The treatment has not produced a statistically significant effect, because either it is ineffective or the study was under-powered.

No clinically significant effect. The lower CI of the treatment effect lies above zero, but the upper CI does not include the clinically significant threshold. The treatment effect may be statistically significant, but not clinically significant.

Not significantly less than the threshold. This type of result will occur when the mean treatment effect lies between zero and the clinical threshold, the lower CI for the treatment does not include zero and the upper CI includes the clinically significant threshold. In this situation, the mean effect is not significantly lower than the clinical threshold, so the treatment effect should be considered clinically significant. For brevity, this treatment effect could be called "small but clinically significant".

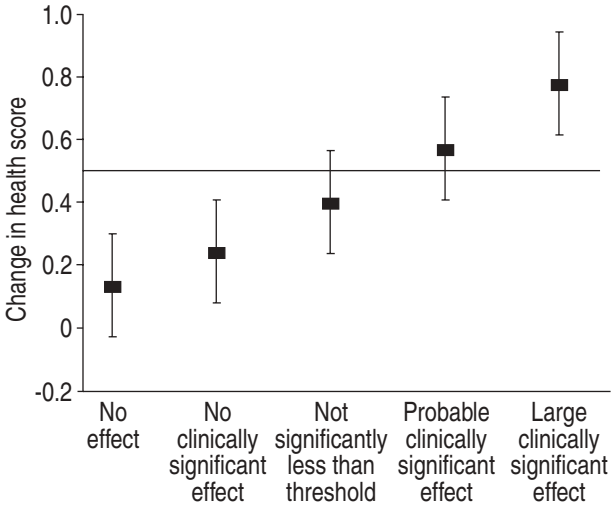

Fig. 3. – Method of categorizing clinical trials results using the threshold for clinical significance and the confidence intervals around the mean treatment effect. See text for a further explanation. Solid line represents the threshold for minimum clinically significant effect. Boxes represent 95% confidence intervals.

It is important to note that this method of classification has a built-in safeguard against the possibility that a small under-powered study using an ineffective treatment will produce a result compatible with a clinically significant effect. In a trial such as this, the mean value could lie between zero and the threshold, but the CIs will be wide because of the small study numbers. Whilst the upper CI may include the threshold for clinical significance, the lower CI will include zero, so the result would not be statistically significant.

Probable clinically significant effect. This will occur when the mean score lies above the threshold, but the lower CI includes it. This is the result that is described currently as being "clinically significant". Strictly speaking this should be termed "moderate probability of clinically significant effect", since it is possible that the "true" treatment effect will lie below the clinically significant threshold, but more probable that it will lie on or above it.

Large clinically significant effect. When the lower CI of the treatment effect lies above the threshold, *i.e.* the mean effect is significantly greater (statistically) than the threshold, this may be termed a "large clinically significant effect", although strictly speaking this should be called "high probability of clinically significant effect". Health status changes of this size have been described in pulmonary rehabilitation [19, 20].

This approach has the value of recognizing uncertainty in measurement and categorizing the size of the response using an explicitly stated set of rules. It does not remove uncertainty around the estimates for the clinical threshold, but does reduce some of the effects of this uncertainty.

To what do the thresholds correspond in "real life"?

Health status measurement is the science of turning patient's symptoms, sense of wellbeing and physical, social and emotional impairment into numbers that permit scientific analysis. These numbers still have little meaning for clinicians so it may be useful to convert the numbers back into some example of "real life" settings.

Back-converting the St George's Respiratory Questionnaire thresholds

The SGRQ has discrete items that have mainly yes/no answers. Each item also has its own weight, so it is possible to put together clinical scenarios that correspond to a 4 unit change in SGRQ score. Table 1 contains five scenarios that correspond to a 4 unit improvement in SGRQ. They illustrate different patterns of change in a patient's symptoms, daily life and wellbeing that will produce a clinically significant change in SGRQ score. The content of each of these scenarios matches very closely items contained in the SGRQ, and to score 4 units a patient would have had to have a change in each of the effects of their disease contained in any given scenario.

Conclusion

Health status measurements form an established part of the process of assessing treatment efficacy.

Table 1. – Five scenarios that illustrate different patterns of change to a 4 unit change in the St George's Respiratory Questionnaire Score

Patient no.	Scenario
1	Attacks of wheeze change from most days a week to a few days a month, no more morning chest tightness, no longer breathless on playing sports and games and now only restricted in one or two activities that the patient wants to do compared to most things (which was the case previously).
2	No more disturbed sleep due to coughing, now able to play tennis and no longer embarrassed by cough and breathing in public.
3	No longer takes a long time to wash or dress, can now walk up stairs without stopping and go out for entertainment.
4	Things no longer seem to require too much effort, no longer has to stop for rests while doing housework and can now carry things upstairs.
5	No longer has to walk more slowly than other people, no longer breathless on getting washed and dressed or on bending over.

Patients one and two would be typically young adult asthmatic patients, whereas the scenarios illustrated by patients three to five would occur in patients with chronic obstructive pulmonary disease.

Estimates of threshold for clinical significance appear to be reliable between methods of assessment, but they should be used with care, taking into account the fact that they cannot be obtained without measurement error. The same applies to any such threshold, for example forced expiratory volume in one second [21] and walking distance tests [22]. Clinicians, formulary committees and regulatory authorities are now requesting evidence that treatments have clinically significant benefits, and these need to be able to be demonstrated. When used appropriately, these thresholds can be used to aid judgements about the clinical relevance of a treatment′s effect. This analysis was written in response to numerous requests to the author to clarify these issues in the context of health status measurement, but the issues it has addressed apply equally to any area of measurement in pulmonary medicine.

References

1. Jones PW. Health status measurement in chronic obstructive pulmonary disease. *Thorax* 2001; 56: 880–887.
2. Wright JG. The minimal important difference: who′s to say what is important? *J Clin Epidemiol* 1996; 49: 1221–1222.
3. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol* 1996; 49: 1215–1219.
4. Kirwan JR, Chaput de Saintonge DM, Joyce CRB, Currey HLF. Clinical judgment in rheumatoid arthritis. II. Judging "current disease activity" in clinical practice. *Ann Rheum Dis* 1983; 42: 648–651.
5. Osman LM, Godden DJ, Friend JAR, Legge JS, Douglas JG. Quality of life and hospital re-admission in patients with chronic obstructive pulmonary disease. *Thorax* 1997; 52: 67–71.
6. Rutten van-Mölken MPMH, Custers F, Van Doosrlaer EKA, *et al.* Comparing the performance of four different instruments in evaluating the effects of salmeterol on asthma quality of life. *Eur Respir J* 1995; 8: 888–898.
7. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol* 1994; 47: 81–87.
8. Juniper EJ, Guyatt GH, Epstein RS, Ferrie PJ, Jaeschke R, Hiller TK. Evaluation of impairment of health related quality of life in asthma: development of a questionnaire for use in clinical trials. *Thorax* 1992; 47: 76–83.
9. Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. *Thorax* 1987; 42: 773–778.
10. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure for chronic airflow limitation - the St George′s Respiratory Questionnaire. *Am Rev Respir Dis* 1992; 145: 1321–1327.
11. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clin Trials* 1989; 10: 407–415.
12. Jones PW, Quirk FH, Baveystock CM. The St George′s Respiratory Questionnaire. *Respir Med* 1991; 85: 25–31.
13. Barley EA, Jones PW. A comparison of global questions *versus* health status questionnaires as measures of the severity and impact of asthma. *Eur Respir J* 1999; 14: 591–596.
14. Jones PW. the Nedocromil Sodium Quality of Life Study Group. Quality of Life, symptoms and pulmonary function in asthma: long-term treatment with nedocromil sodium examined in a controlled multicentre trial. *Eur Respir J* 1994; 7: 55–62.
15. Jones PW, Lasserson D. Relationship between change in St George′s Respiratory Questionnaire score and patients′ perception of treatment efficacy after one year of therapy with nedocromil sodium. *Am Rev Respir Crit Care Med* 1994; 149: A211.
16. Jones PW, Bosh TK. Changes in quality of life in COPD patients treated with salmeterol. *Am J Resp Crit Care Med* 1997; 155: 1283–1289.
17. Bestall JC, Paul EA, Garrod R, Garnham R, Jones PW, Wedzicha JA. Usefulness of the Medical Research Council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. *Thorax* 1999; 54: 581–586.
18. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998; 316: 690–693.
19. Lacasse Y, Wong E, Guyatt GH, King D, Cook DJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronic obstructive pulmonary disease. *Lancet* 1996; 348: 1115–1119.
20. Griffiths TL, Burr ML, Campbell IA, *et al.* Results at 1 year of outpatient multidisciplinary pulmonary rehabilitation: a randomised controlled trial. *Lancet* 2000; 355: 362–368 (published erratum appears in *Lancet* 2000; 355: 1280).
21. Redelmeier DA, Goldstein RS, Min ST, Hyland RH. Spirometry and dyspnea in patients with COPD. *Chest* 1996; 109: 1163–1168.
22. Redelmeier DA, Bayoumi AM, Goldstein RS, Guyatt GH. Interpreting small differences in functional status: the six minute walk test in chronic lung disease patients. *Am J Respir Crit Care Med* 1997; 155: 1278–1282.