# Development of a symptom based outcome measure for asthma

Nick Steen, Allen Hutchinson, Elaine McColl, Martin P Eccles, Jenny Hewison, Keith A Meadows, Stephen M Blades, Penny Fowler

**Measuring symptom specific health outcome is complex, but the methodologies now exist to develop measures with the appropriate properties. As one element of a major programme to develop multidomain health outcome measures for chronic disease, a symptom based measure for asthma care has been developed for use in general practice and outpatient departments. This article outlines the development process, which used a framework recently described in the theoretical literature to show the constraints that scientific criteria place on the development of outcome measures and the means of overcoming such limiting factors. Although substantial effort is required to undertake a rigorous process of development, useful tools are the result. Two five item, symptom based outcome measures for adult asthma are described.**

**Centre for Health Services Research, University of Newcastle upon Tyne, Newcastle upon Tyne**
Nick Steen, *research associate*
Elaine McColl, *senior research associate*
Martin P Eccles, *senior lecturer*
Stephen M Blades, *visiting research associate*

**Department of Public Health Medicine, University of Hull, Hull**
Allen Hutchinson, *professor of public health medicine*
Keith A Meadows, *research fellow*

**University of Leeds, Leeds**
Jenny Hewison, *senior lecturer in psychology*

**East Gloucestershire NHS Trust**
Penny Fowler, *quality development manager*

Correspondence to:
Professor Hutchinson.

*BMJ* 1994;309:1065-8

Although health outcome measurement is now becoming a priority for the NHS, substantial scientific hurdles are yet to be overcome. In the move from the assessment of health status to the evaluation of health outcome, a recent series of papers in the *BMJ* emphasised that not all health status measures have the scientific attributes required of an outcomes measure (for instance, evidence of responsiveness to change) and not all outcome measures are health status measures in psychometric terms (for instance, peak flow reading in the management of asthma).[1 2] Other authors have suggested that health outcome indicators could be used to compare populations that are subject to a range of influences, rather than just to evaluate the effect of a specific intervention.[3]

Fitzpatrick and colleagues identified important scientific issues in the measurement of health status and health outcome.[1 2] Measures must have the properties of validity and reliability. Validity relates to the effect of systematic error; an instrument is valid to the extent that it measures what it purports to measure.[4] Reliability relates to the effect of random error; it refers to the extent to which the measure can reproduce the same results in repeated applications under unchanged circumstances. In addition to these properties, if health outcome measures are to be used as evaluative instruments they must also have the property of responsiveness to change—the ability to identify what may be small but none the less clinically important changes.

This paper considers these scientific issues in the development of a symptom based outcome measure for adults with asthma. The work was undertaken in the context of an attempt to develop practical outcome measures that clinicians could use to assess the quality of their care within primary care and hospital outpatient departments (ambulatory care). Symptom based outcome measures for both asthma and diabetes were developed to complement functional and psychosocial outcome measures (collected from patients) and biomedical indicators of outcome (collected from clinicians). The parent study, aimed at developing these suites of measures, will be reported elsewhere; this paper describes the work on symptom measures for asthma. The key developmental steps are identified sequentially; they follow closely the recommendations of Streiner and Norman on the development of health measurement scales.[5]

## Critical review of the evidence

A review of the international literature must be undertaken to determine whether an equivalent or similar piece of work has already met the project requirements. In this case work on measurement of asthma symptoms was identified but few of these health status measures had the properties of health outcome measures. The asthma symptom questionnaires developed by Kinsman *et al* and Usherwood *et al* had good internal reliability, and evidence was also given in support of their validity.[6] Patients had been used as a principal source of symptoms in the development of both questionnaires. But neither questionnaire fully met our needs: because we had already identified validated instruments that measured functional and psychosocial health status our requirement was for a questionnaire that measured clinical symptoms only. No evidence in respect to sensitivity to change was reported for either instrument. Nevertheless, the two questionnaires provided a useful source of clinical symptoms that might be included in a new questionnaire.

## Selection of symptoms for a new questionnaire

Two general practitioners, a clinical psychologist, and several asthmatic patients were asked to review the pool of items to check for completeness (to determine if any symptoms were missing) and also to offer an opinion about the appropriateness of each of the symptoms—that is, they were asked to ensure the content validity and face validity of the new questionnaire.[5] Content validity refers to the appropriateness of the selection of concepts for inclusion in the measure. Several symptoms—for example, dizziness—were excluded at this stage, since they were judged by clinicians or patients to be of little specific relevance to patients with asthma. An item about coughing at night was added, since this is a common symptom in asthma and thus is appropriate and important to include. The final pool of items is given in table I.

## Construction of items

Three particular issues were of concern in this development process—choosing a recall period, which may affect reliability, and quantifying the frequency

TABLE I—*Items included in scales for measuring symptoms of asthma*

| | Reduced scale | |
|---|---|---|
| Asthma symptoms | 1 | 2 |
| Breathlessness during exercise | ✓ | ✓ |
| Breathlessness during day when not exercising | | ✓ |
| Wheezing during day | ✓ | |
| Coughing during day | ✓ | |
| Wheezing at night | | ✓ |
| Breathlessness at night | | |
| Coughing at night | ✓ | ✓ |
| Disturbed sleep | | |
| Fear because of asthma | ✓ | |
| Feeling of tightness in chest | | ✓ |

and severity of symptoms. Recall is likely to be more accurate over a short recent period than over a protracted period. Thus a questionnaire that asks about symptoms experienced in the previous week is likely to have greater reliability than one that asks about symptoms experienced during the past three months. If prevalence of a particular symptom in a target population is low, however, few patients may endorse a question relating to its occurrence if the recall period is too restricted—for example, the previous day. Low item endorsement may reduce the utility of the measure for detecting differences in health status between individuals.[5]

It was necessary, therefore, to strike a balance between a timeframe that was too long, perhaps leading to recall bias, and too short, which might lead to low item endorsement. We therefore chose one month as the recall period for items asking about symptom frequency and three months for items asking about more memorable events such as time off work and consultations with the doctor.

The problem of assessing symptom severity was that of finding expressions that were acceptable to patients and clinicians. Some patients, for example, might find a severe wheeze during the day less of a problem than a slightly less severe attack during the night. Various forms of wording were considered and piloted with patients. The final choice was, "How much bother does the symptom cause?"

## Scaling responses

Each of the 10 questions on frequency and severity of symptoms was of the form, "On how many days in the past month have you experienced a particular symptom?" (box). It was felt to be unreasonable to expect the patient to recall the exact number of days; instead a five point scale, ranging from never to every day, was used. Bother was measured on a four point scale ranging from no bother at all to very much bother. We provided a "does not apply to me" category for patients who had not experienced that symptom during the past month.

## Testing the questionnaires

INTERPRETABILITY AND ADMINISTRATION

The new questionnaire was piloted on groups of patients to ensure that each item could be easily understood. This resulted in some minor refinements to the wording of questions.

Items where the vast majority of patients endorsed only one of the responses categories were discarded, since such items provide little information.[5]

ENDORSEMENT OF RESPONSE CATEGORIES

After a questionnaire has been tested for readability and absence of ambiguity, it should be tested for endorsement frequency. The asthma symptom questionnaire was given by clinicians to a convenience sample of patients with asthma, drawn from 32 general practice sites and seven hospital outpatient clinics. Sample size calculations were based on the requirements of the parent study. People with asthma aged 18 and over and who could read and write in English were eligible for inclusion in the study. Pregnancy was an absolute exclusion criterion, and clinicians were discouraged from recruiting patients during an acute exacerbation of the illness, since this was likely to produce a biased view of symptom frequency and severity. In practice, most patients were recruited during a routine review consultation. A total of 639 patients were recruited, 390 from general practices and 207 from outpatient clinics.

### Format of questions

Please encircle the number that best describes how you have been in the last month.
In the last month, on how many days have you wheezed during the day?

| Never | On one or a few days | On several days | On most days | Every day |
|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 |

How much bother did the wheezing cause?

| Does not apply to me | No bother at all | Not much bother | Much bother | Very much bother |
|---|---|---|---|---|
| 8 | 0 | 1 | 2 | 3 |

Response rates provide some indication of the acceptability of the questionnaire to patients. For the asthma symptoms questionnaire the overall response rate, after two reminders, was 93%. (Response rate before one reminder was 78% and after one reminder, 88%.) Non-response to individual questions was also low; of those questionnaires returned, 98% contained complete data for all 10 symptoms. It is likely that both these figures will be higher than those arising from a similar administration to a random sample; in this development phase clinicians were asked to exclude patients who they felt would have difficulty completing the questionnaire. None the less, these response rates provide some indication that the questionnaire is acceptable to and understandable by patients.

INTERNAL RELIABILITY

Homogeneity refers to the extent to which an instrument measures a single concept or trait. All items in a scale should measure the same concept, perhaps summing the responses to the individual items to obtain a simple index. The index would not be easy to interpret if the items of which it consists actually measured different concepts. If a scale is homogeneous, the scores on the individual items would be expected to be correlated with each other. A widely used indicator of internal reliability or homogeneity is Cronbach's $\alpha$.[8] Low values of $\alpha$ would indicate poor internal reliability (the items do not all measure the same concept), but high values suggest that the items correlate so well that the information could be obtained from a subset of items.

There is no hard and fast rule as to what constitutes an optimum value of $\alpha$ but somewhere in the range 0·70-0·85 would seem reasonable. The internal reliability of the 10 item questionnaire was 0·93, suggesting that some items are redundant and could be eliminated without losing too much information. This was confirmed when we examined the partial correlation coefficients[9] of the individual symptom scores with the total score. Regression analysis suggested that a five item scale would explain over 95% of the variation in the 10 item scale.

Two possible sets of five items were identified through the regression analysis (table I). Shortness of breath on exercise was common to both sets. Each set contained two additional questions relating to daytime symptoms and two questions relating to nocturnal symptoms. The internal reliability of these two reduced scales was 0·86 and 0·87. In this developmental phase of the work 10 items have been retained pending an assessment of their performance in planned inteventions in ambulatory care, but these results suggest that either of the five item groups would be adequate in future investigations.

TEST-RETEST RELIABILITY

Reliability of the questionnaire should be checked by administering the questionnaire twice, 2-14 days apart, to the same group of patients.[5] Test-retest reliability of the asthma symptom questionnaire will be assessed in a future study.

### From items to scales

In deriving a scale from a set of items, consideration must be given to the relative weighting of those items. Clinicians considered that symptoms which caused patients a lot of bother should be given a higher weighting than symptoms which caused little bother. The mean level of bother corresponding to each response category for each symptom was calculated. The statistical package ELIM was used to model the data,[10] generalised linear models[11] were fitted using weighted least squares. This indicated that there was a symptom effect (the mean level of bother varied from symptom to symptom); that there was a response category effect (the mean level of bother changed as patients reported higher symptom frequency); but that there was no interaction effect between symptoms and response categories. That is, the way in which bother depends on symptom frequency is the same for all symptoms.

This suggested a weighting system that used relative weights for each of the symptoms (these range from 0·81 for coughing during the day to 1·30 for fear due to asthma) and relative weights for each of the response categories (never=0; one or a few days=1·4; several days=2·2; most days=2·7; and every day=3·7). The performance of the weighted and unweighted scales is compared in the following sections.

### Validity

The key problem with the estimation of validity is that there is no "gold standard" to act as a criterion against which performance of the scale of interest can be judged.[1] Instead, the criterion of expert clinical opinion is sometimes used. For a sample of patients, clinicians were asked to complete the Duke severity of illness index.[12] This included a question in which the doctors were asked to rate, on a five point scale, the level of symptoms experienced by the patient during the previous week.

Clinicians were asked to consider what other criteria might be appropriate to assess concurrent and predictive validity. They expected that a high level of symptoms would be associated with adverse occurrences such as the number of asthma attacks and chest infections experienced by the patient, a higher frequency of consultations, and time off work (or impairment of other activities) due to asthma. Concurrent validity refers to the extent to which responses on the scale under investigation correlate with scores on a criterion measure, where both measurements are made at the same time. In assessing concurrent validity, therefore, we related patients' symptom scores and doctors' assessment of symptom levels to adverse occurrences in the previous three months, as reported in the same questionnaire. Predictive validity is established by relating responses on the scale of interest at a given time to a criterion some time in the future[5]; we related patients' symptom scores at baseline to adverse events as reported in an identical follow up questionnaire three months later.

TABLE II—*Concurrent and predictive validity of the 10 item asthma symptom questionnaire. Values are Spearman rank-order correlation coefficients*

| | 10 Item asthma symptom score | | Doctors' evaluation of symptom level | |
|---|---|---|---|---|
| Criteria | Concurrent validity* | Predictive validity† | Concurrent validity* | Predictive validity† |
| No of asthma attacks | 0·45 | 0·44 | 0·32 | 0·27 |
| Chest infections | 0·47 | 0·37 | 0·36 | 0·30 |
| Routine consultations | 0·53 | 0·57 | 0·37 | 0·42 |
| Unplanned consultations | 0·36 | 0·53 | 0·35 | 0·26 |
| Impaired activity | 0·56 | 0·53 | 0·34 | 0·34 |

*Scale scores were correlated with the adverse occurrences (the criteria) which occurred during the three months before the questionnaire.

†In assessing predictive validity, scale scores were correlated with the adverse occurrences which occurred in the three months after the questionnaire.

Some of these criteria are influenced by factors other than level of symptoms. Thus, although correlation of the criteria with the symptom score should be high, it is not expected to be perfect. Table II shows how well the total asthma symptom score correlates with the criteria identified by the panel of experts contrasted with the clinicians' rating of symptom level. Correlations of the criteria with the asthma symptom score were higher than the corresponding correlations with the clinicians' ratings. This may be due to the coarseness of the clinicians' rating scale, the different time scale, or possibly because the information about symptoms provided by clinicians is less reliable than that provided by patients themselves. The doctors' ratings are probably based on information provided by the patients; they are a less direct assessment of the level of symptoms experienced by the patient.

TABLE III—*Coefficients (99% confidence intervals) for correlation of alternative forms of asthma symptom score with criterion index*

| | Concurrent validity | Predictive validity |
|---|---|---|
| Scale | Correlation coefficient (99% confidence interval) | Correlation coefficient (99% confidence interval) |
| Total symptom score | 0·68 (0·62 to 0·74) | 0·59 (0·48 to 0·69) |
| Weighted symptom score | 0·68 (0·62 to 0·74) | 0·58 (0·47 to 0·68) |
| Reduced scale 1 | 0·65 (0·59 to 0·71) | 0·58 (0·46 to 0·68) |
| Reduced scale 2 | 0·66 (0·59 to 0·72) | 0·57 (0·45 to 0·67) |
| SF-36 general health perception scale | 0·58 (0·46 to 0·68) | 0·53 (0·33 to 0·68) |

Validity of different forms of the asthma symptoms questionnaire was further investigated by considering their correlation with a single index formed from the criteria identified above (table III). To provide context, correlation of the criterion index with a validated general health status scale—the general health perception scale from the SF-36[13]—is also reported.

TABLE IV—*Change in asthma symptom score by patients' perception of change in their asthma*

| Perception of change | Mean change (95% confidence interval) |
|---|---|
| Much better (n=26) | −15·0 (−23·1 to −6·9) |
| A little better (n=36) | −6·0 (−11·1 to 0·9) |
| About the same (n=142) | −0·4 (−2·9 to 2·2) |
| A little worse (n=40) | 5·2 (−0·4 to 10·8) |
| Much worse (n=5) | −1·0 (−14·5 to 12·6) |

In terms of correlation with the criterion index, there is very little difference between the weighted and unweighted symptom scores. The performance of the two reduced scales is almost identical and their correlation with the criterion index is of the same order as the correlation of the total symptom score with the criterion index. In general the symptom scores seem to correlate more highly than the general health perception scale, but there is some overlap in the reported confidence intervals.

A further aspect of validity is construct validity; this refers to the extent to which the results obtained with a measure accord with the pattern of responses that would be expected on theoretical grounds. One indication of construct validity is the discriminatory power of the measure—its ability to discriminate or distinguish between groups that are known to differ.[5] We assessed the discriminatory power of our measure of asthma symptoms outcome by means of a case-control study. A total of 229 patients with asthma and 448 age and sex matched controls were drawn at random from the disease registers and practice lists of two general practices in Newcastle upon Tyne; sample sizes were determined to detect a difference of 10% between cases and controls with 80% power. All those sampled were asked to complete a questionnaire containing nine of the 10 questions on symptom frequency (the question relating to fear due to an asthma attack was not given to controls and was omitted). It was hypothesised that the asthmatic patients would have poorer health (more frequent symptoms) than the controls. A statistical test was needed that would take into account the ordinal nature of the response categories (that is, that "every day" was worse than "on most days"); the statistic of choice was the Mantel-Haenszel $\chi^2$ value.[14] Differences between cases and controls were large for all symptoms, indicating that our measure had satisfactory discriminatory power.

### Responsiveness and sensitivity to change

An outcome measure should be able to detect change in health status over time. The degree of sensitivity required will depend on the purpose of the measure. A measure intended to assess outcome for a single patient will need to be more sensitive than one intended to assess outcome for groups of patients. The 10 item asthma symptom questionnaire was administered twice to 250 patients, three months apart. The patients were also asked directly about how they thought their asthma had changed during that three month period.

There was reasonable agreement between the patients' assessment of change and the change in their asthma symptom score calculated by taking the difference of the scores obtained at baseline and follow up (table III). Analysis of variance[8] indicated that the mean score varied between groups of patients with the same perception of change in health. There was a significant trend across groups ($F_{1244}$=27·2); the deviation from linearity was not significant.

A preliminary assessment of the relative performance of different forms of the asthma symptom questionnaire has been based on the 26 patients who had perceived much improvement in their asthma (table IV). For all forms of the symptom scale there was a significant reduction in score for these 26 patients between baseline and follow up. The magnitudes of the *t* statistics derived from the paired *t* tests suggest that there is little difference in the performance of the different forms of the scale. The unweighted scale performed as well as the weighted scale, both five item scales performed as well as the full length version. In contrast, there was no change in self reported health status measured on the SF-36 general health perception scale. Thus there is some preliminary evidence to suggest that, for asthma at least, the symptom based outcome measure is much more sensitive to change in a patient's asthma than is a generic measure of general health perception.

Effect size has been proposed as a measure of the responsiveness of a questionnaire.[15,16] This is simply the difference between baseline and follow up scores divided by the standard error of baseline scores. Comparison of effect size across the different scales (table IV) shows little difference in the relative performance of the different forms of the symptom measure and that the generic measure does not seem to be responsive to change.

TABLE V—*Change in total symptom score for 26 patients reporting much improvement in their asthma*

| | Mean score | | | | |
|---|---|---|---|---|---|
| Scale | Baseline | Follow up | *t* Value | P value | Effect size |
| Total symptom score | 14·5 | 8·5 | 3·65 | <0·001 | 0·66 |
| Weighted symptom score | 14·9 | 9·7 | 3·78 | <0·001 | 0·70 |
| Reduced scale 1 | 8·1 | 4·9 | 3·85 | <0·001 | 0·65 |
| Reduced scale 2 | 8·1 | 4·7 | 3·57 | <0·001 | 0·64 |
| SF-36 general health perception scale* | 58·9 | 55·8 | −0·79† | 0·72† | −0·14 |

*The SF-36 is scaled such that an increase in the score represents an improvement in health status.
†The SF-36 was administered to a subset of patients only; calculations are based on 13 cases.

## Discussion

A symptom based outcome measure has been developed for use in two situations—as a measure of a specific intervention such as starting steroid treatment, and as a comparative indicator between general practices or outpatient clinics. In undertaking such a development, certain "rules" must be followed. Much time and effort can be saved by drawing on the work of those who have previously carried out research in the field of interest. If new symptom based measures need to be developed, they must have the same attributes of reliability and validity as all other forms of outcome measure. Above all, if the measures are to be used to evaluate the effects of care they should be responsive and sensitive to change. The work reported here shows that symptom based outcome measures with these desirable properties can be developed. Such condition specific measures may be more responsive to change than generic measures, which reinforces the support for condition specific outcome measures.

For the measure reported here there seems to be no advantage in assigning weights to different symptoms; there was little difference in the relative performance of the weighted and unweighted index. There is also some evidence to suggest that the scale may be shortened to five items without significantly reducing its performance. Use of the measure alongside specific interventions in primary care settings is being planned and should facilitate its refinement. This next phase of the development will focus on the clinical utility of the symptom questionnaire. In particular we shall investigate whether it can be used as part of a patient profile that will help in management of individual patients.

This research was funded within the ambulatory care research programme under a Department of Health research grant. Many practitioners in general practice and hospitals throughout the Northern region gave us their support and time in this study. We are very grateful to them. Our administrative team—Christine Hutchinson, Linda Duckworth, and Sylvia Hudson—played a central role in ensuring success, and we were assisted by Angela Watson during the pilot work.

1 Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I. Applications and issues in assessment. *BMJ* 1992;305:1074-7.

2 Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II. Design, analysis, and interpretation. *BMJ* 1992;305:1145-8.

3 McColl AJ, Gulliford MC. *Population health outcome indicators for the NHS: a feasibility study*. London: Faculty of Public Health Medicine of the Royal College of Physicians, 1993.

4 Wilkin D, Hallam L, Dogget M. *Measures of need and outcome for primary health care*. Oxford: Oxford University Press, 1992.

5 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press, 1989.

6 Kinsman RA, Luparello T, O'Banion K, Spector S. Multidimensional analysis of the subjective symptomatology of asthma. *Psychosom Med* 1973;35:250-67.

7 Usherwood TP, Scrimgeour A, Barber JH. Questionnaire to measure perceived symptoms and disability in asthma. *Arch Dis Child* 1990;65:779-81.

8 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-334.

9 Kirkwood BR. *Essentials of medical statistics*. London: Blackwell Scientific, 1988.

10 Aitkin M, Anderson D, Francis B, Hinde J. *Statistical modelling in GLIM*. Oxford: Clarendon Press, 1989.

11 McCullagh P, Nelder JA. *Generalised linear models*. London: Chapman and Hall, 1989.

12 Parkerson GR Jr, Broadhead WE, Tse C-KJ. The Duke severity of illness checklist (DUSOI) for measurement of severity and comorbidity. *J Clin Epidemiol* 1993;46:379-93.

13 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, *et al.* Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992;305:160-4.

14 Mantel N, Haenszel W. Statistical aspects of analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719-48.

15 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27:S178-89.

16 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Control Clin Trials* 1991;12:142-58S.

*(Accepted 25 July 1994)*