Best practice in symptom assessment: a review

E McColl

Gut 2004;**53**(Suppl IV):iv49–iv54. doi: 10.1136/gut.2003.034355

Assessment of symptoms is important both in the diagnosis of gastro-oesophageal reflux disease (GORD), and in monitoring response to therapeutic interventions. Key quality aims in assessing symptoms are to gather data that are valid, reliable, unbiased, discriminating, and responsive to change. Important considerations include who should assess symptoms, method of data collection, timing of assessment, and methods of assessing multiple symptoms.

SUMMARY

Assessment of symptoms is important both in the diagnosis of gastro-oesophageal reflux disease (GORD) and in monitoring response to therapeutic interventions. Key quality aims in assessing symptoms are to gather data that are valid, reliable, unbiased, discriminating, and responsive to change. Important considerations are: who should assess symptoms; method of data collection; timing of assessment; and methods of assessing multiple symptoms. A review of the substantial literature on these topics has led to the following conclusions. (1) Due to the subjective nature of symptoms, patient self report is more appropriate than clinician assessment. (2) Symptom diaries are generally considered to be the ''gold standard'' but well designed questionnaires with an appropriate recall period may be adequate. (3) Careful attention to instrument design and data collection protocol can enhance completion rates. (4) Use of a clearly defined reference period can improve the quality of response; the choice of reference period must take into account anticipated fluctuations over time in the symptoms of interest and how memorable those symptoms are likely to be—reference periods of one week to one month are most appropriate. (5) The duration of data collection must represent an appropriate compromise between capturing anticipated day to day variation and imposing excessive respondent burden. (6) There is no single optimal strategy with respect to collecting data on multiple symptoms. Decisions should be based on the objectives and analytic strategies of individual studies, on judgements about likely covariation in symptoms, and on whether respondents themselves can make composite ratings or whether the assumptions implicit in summated ratings are justified.

Correspondence to: Ms E McColl, National Primary Care Career Scientist, Centre for Health Services Research, School of Population and Health Sciences, University of Newcastle upon Tyne, 21 Claremont Place, Newcastle upon Tyne, NE2 4AA, UK; e.mccoll@newcastle.ac.uk

INTRODUCTION

Symptoms have been defined as ''subjective experience of abnormal function, sensation, or appearance, generally indicating disorder or disease'' (personal communication from Patient Reported Outcomes Harmonization Group, 2002). Since a significant proportion of patients with GORD do not have erosive oesophagitis,[1] evaluation of GORD symptoms is a key part of diagnosis,[2] while monitoring the response of symptoms to therapy is crucial in assessing the effectiveness of the intervention.

The key objectives, in terms of data quality, in measuring symptoms are to collect data that are:

- valid (that is, are a ''pure'' measure of the target symptom and are not distorted by other interfering variables);
- reliable (that is, are reproducible in the sense that the same measurement is obtained when there has been no change in symptoms);
- unbiased (that is, symptoms are not measured in a way that systematically under or over-estimates the true value);
- discriminating (that is, the measure can distinguish adequately between individuals who actually do differ to an important degree in terms of the variable measured—such as symptom severity);
- responsive to change (analogous to discriminating, but referring to the ability to detect true change over time or in response to an intervention).

Important considerations are: who should assess symptoms (patient versus clinician); the method of data collection (questionnaires versus diaries); the timing of assessment (reference period and duration); and the number of symptoms to be assessed (in particular, the relative merits of single item measures, multiple items analysed as a profile, summated scores, or global assessments).

WHO SHOULD ASSESS SYMPTOMS

Previous findings on clinician versus patient ratings of symptoms

There is some evidence from studies in which independent measures of symptoms could be made—for example, patient reports of palpitations compared with documented arrhythmias from 24 hour ambulatory electrocardiograph monitoring[3]—of a lack of concordance between subjective reports of the incidence and severity of symptoms and objective measures of disease activity. In general, however, the subjective

Abbreviations: GORD, gastro-oesophageal reflux disease

nature of symptoms, as implied by the Patient Reported Outcomes Harmonization Group's definition above, suggests that self report is the most appropriate method of gathering data on symptom experience. Findings from studies across a range of conditions, including GORD,[4–10] show a general pattern of only weak to moderate agreement between patient and clinician ratings, both for symptoms and for quality of life impairment. Although discrepancies in both directions have been reported, the most common finding is for clinicians to underestimate the incidence and severity of symptoms, relative to patient ratings.[10] Mismatches generally appear to be greater among patients with less severe symptoms.[5] Agreement on the presence versus the absence of symptoms is usually better than agreement about symptom intensity.[11]

Empirical evidence of mismatches between clinician and patient ratings of symptoms in GORD

Previous research into mismatch between patient and clinician ratings of symptoms was borne out in secondary analyses of data on file at AstraZeneca, which were carried out specifically for this workshop. The data related to a large sample of both male and female patients, aged 18–80 years, whose dominant symptom was heartburn. These patients were included in a multicentre, double blind, randomised, parallel group study, comparing four week courses of three drug regimens; for the purposes of these secondary analyses, data from all three arms of the trial were combined. Clinician and patient ratings, at three points in time (baseline, two, and four weeks), of the severity of three symptoms of GORD (heartburn, epigastric pain, and regurgitation) were obtained. Our analyses showed that the level of agreement was poorest at baseline; clinicians underestimated symptom severity relative to patients' ratings of heartburn in 34% of cases; underestimates of the severity of epigastric pain and regurgitation were observed in 55% and 46% of cases, respectively. At baseline, clinicians judged that 40% of patients were free from epigastric pain but only 14% of patients considered themselves to be free of this symptom; the corresponding values for regurgitation were 32% and 14%. Low clinician or investigator ratings of symptom severity at baseline represent a potential ''floor effect'' with little scope for improvement over time.

Levels of agreement between clinician and patient ratings improved over time. At four weeks, patients and clinicians agreed on heartburn severity in 78% of cases; agreement in respect of severity of epigastric pain and regurgitation was observed in 58% and 76% of cases, respectively. This convergence of patient and clinician ratings may be, in part, an artefact of the low investigator ratings at baseline coupled with patients' perceptions of improvement over time, thus bringing the two ratings more closely in line.

Implications of findings regarding clinician versus patient ratings

The implications of our findings and those of previous researchers are that when clinicians' ratings are used, estimates of absolute treatment effects at any given point in time are likely to be upwardly biased (that is, overestimate the impact of treatment) because symptom severity is underestimated. Conversely, by using clinicians' ratings, comparisons of changes in symptom severity over time or in response to a therapeutic intervention are likely to underestimate the impact of treatment because of the floor effects and trend towards convergence noted above. In a well designed randomised controlled trial, there is no reason to believe that the mismatch between clinician and patient ratings will vary across the different arms of the trial and, therefore, estimates of the effects of one treatment relative to another should not be affected. However, estimates of absolute treatment effect and, therefore, calculations of number needed to treat, are likely to be biased, as described above.

Potential explanations for mismatches between clinician and patient ratings

Several possible explanations have been proposed for the lack of concordance between patient and clinician ratings. Lack of interrater reliability between clinicians has been demonstrated.[11] Holmes and colleagues[12] have suggested that clinicians may focus primarily on frequency and intensity of symptoms in their history taking while patients also consider symptom related disability and the impact on quality of life in judging the severity of their symptoms.

Lack of a common vocabulary may also underpin discrepancies. Agreus and colleagues[13 14] reported that the terms ''pain'' and ''discomfort'' do not adequately reflect the full range of descriptors used by the general public in describing dyspepsia. Stanghellini[15] suggests that patients may attribute different meanings to ''pain'' and ''discomfort''. Similarly, ambiguity of the term ''heartburn'' has been highlighted[16 17]; Carlsson and colleagues[16] have shown how a ''word picture'' in which heartburn was defined as ''a burning feeling rising from the stomach or lower chest up towards the neck'' facilitated elicitation of experience of this symptom.

Patients may underreport (or indeed exaggerate) symptoms to clinicians. Underreporting may occur when patients wish to please the clinician or researcher[10] and thus represents a social desirability bias. Overreporting may occur if there is a perceived benefit to the patient (for example, being certificated as unable to work).

Patients and clinicians may also use different frames of reference in assessing symptoms.[10] For example, patients are likely to use an internal frame of reference—either in relation to their own past experiences of symptoms or to an ideal state of health—while health care professionals may implicitly compare a given patient to other patients whom they have treated.[18] The use of a structured record or questionnaire[16] may facilitate elicitation of symptoms by a clinician but even this will not guarantee consistency in characterising or assessing symptoms.[11]

METHOD OF DATA COLLECTION

In collecting self report data on symptoms, the two main methods of data collection are questionnaires (most usually self completed, paper and pencil questionnaires, but electronic questionnaires and interviewer administered instruments may also be used) and diaries. Approximately 25% of all clinical trials include diaries as a method of data collection.[19] Several diaries have been developed and used with GORD patients.[20 21] Structured self completion questionnaires have also been developed and validated in GORD patients.[16 22–30] Both diaries and questionnaires have advantages and disadvantages.

Advantages and disadvantages of diaries as a mode of data collection

Diaries, in theory at least, allow for data to be recorded contemporaneously, thus reducing the risk of recall bias. They allow day to day fluctuations in symptoms to be readily captured and, therefore, facilitate calculation of symptom free days. The coincidence of symptom experience and exposure to potential exacerbating factors (such as ingestion of certain foods) can be captured more easily through contemporaneous accounts, as can details of self management behaviour.[31] However, completion of a daily diary imposes considerable respondent burden, and several studies[19 32–35] present objective evidence of non-adherence to data collection protocols, although the diary respondents

themselves reported good compliance. In particular, the problem of ''hoarding'' (retrospective completion of diary entries) has been highlighted[35]; with such retrospective data recording, recall errors may occur. Unfortunately, hoarding is almost impossible to detect unless electronic devices, which can unobtrusively record the timing of data entries, are used. The spectre of fatigue and attrition effects—a drop off in diary completion rates over time—has also been cited as a potential weakness of this method of data collection but this criticism is not supported by findings on participation and completion rates for diary studies. Burman[31] reports that typical participation rates are in excess of 80% while Roghmann and Haggerty[36] and Verbrugge[37] have shown very low rates of missing data in diary studies. These findings from previous research in other populations are borne out by secondary analysis of data on file at AstraZeneca which was carried out specifically for this workshop. In a study of just over 700 GORD patients, 81% of patients completed all required diary entries (that is, daily diary cards for 28 ± 4 days) and a further 16% had 75–99% complete entries (that is, entries on between 75% and 99% of the days for which they remained in the study). There was also no evidence of a fall off in response rates over time; on day 1, 98% of respondents completed their diaries; completion rates on days 24 and 28 were 98% and 94%, respectively.

Other potential sources of biases with diary methods are ''conditioning'', whereby respondents become more tolerant of their symptoms over time, potentially resulting in higher rates of reporting earlier in the diary period[37] and ''sensitising'', whereby symptom awareness is increased over time, resulting in higher rates of reporting as time goes by.

Advantages and disadvantages of questionnaires as a mode of data collection

Questionnaire methods require respondents to remember whether and when they experienced symptoms. Types of recall error include omission (completely forgetting that the symptom was experienced in the time frame of interest) and ''telescoping'' (misplacing an event in time, most usually by recalling it as occurring more recently than it actually did). Dahlquist and colleagues[38] have demonstrated that errors of recall in respect of minor symptoms are likely to occur when the time frame of reference exceeds one week, with minor symptoms being forgotten. Recall biases have also been noted in studies in which estimates of pain relief rely on recollections of past pain,[39 40] leading to overestimation of treatment effects.

When symptoms fluctuate over time, the need to summarise may present patients with a complex cognitive task; they may employ short cuts or heuristics[41 42] in producing their response, again threatening the validity of the answer. Steen and colleagues[18 43] experimented with one and three month reference periods in assessing symptoms of asthma and diabetes. They found that patients commonly used an ''anchor and adjust'' heuristic in respect of the three month recall period; respondents first estimated how often they had experienced the symptom of interest in the preceding month and then adjusted the three month estimate upwards or downwards from that initial anchor.

The risk of social desirability bias—for example, underreporting consumption of certain foods that may trigger symptoms—may also be greater when data are collected by means of questionnaires, especially when these are interviewer administered.[44]

Diaries or questionnaires—which mode of data collection is best

It is generally held that diaries are the ''gold standard'' method of collecting symptom data; rates of reporting minor symptoms are generally higher in diaries than in questionnaires, and the implicit assumption is that ''higher reporting is more accurate''.[37] However, adequate agreement between patterns of symptoms, as recorded in diaries and reported in questionnaires, has been demonstrated in a number of studies.[20 21] Questionnaires may, therefore, be an adequate method of data collection in situations where a contemporaneous account is not essential or where a measure of ''symptom free'' days is not required.

Good practice in diary design

Two main styles of diaries are used in health studies.[45] In a ledger style diary, the occurrence of a symptom or other health related event (such as an encounter with a health professional) is recorded and dated as the event occurs; thus entries relate only to days on which the phenomenon of interest occur. In a journal style diary, an entry is made each day (or defined time period), regardless of whether an event occurs. In this style of diary, provision must be made for recording the absence of the phenomenon on that day (for example, that the symptom of interest was not experienced). Ledger style diaries reduce respondent burden[31] but there is a greater risk of the respondent forgetting to make an entry when it would be appropriate to do so and, therefore, there is greater difficulty in distinguishing between the absence of the symptoms and missing data.[36]

For either style of diary, a further choice lies between a highly structured format, using closed questions, and a less structured approach, with open ended items. In symptom diaries, the structured approach, using a predefined list of symptoms and associated response categories (for example, presence/absence, or rating of symptom severity) ensures equivalence of stimulus (thus increasing the reliability of the response), reduces respondent burden, and facilitates data coding and entry.[46] Most diaries used with GORD patients to date have been highly structured based, for example, on the gastrointestinal symptoms rating scale.[29]

Well designed diaries and clear instructions on how they are to be completed enhance both the quantity and quality of data collected.[47] Face to face instruction on diary protocols is desirable[31] and telephone or electronic prompts seem to be more effective than postal reminders.[38 48] Electronic diaries have been shown to be an effective and well accepted alternative to traditional paper based methods.[32 49]

Good practice in questionnaire selection and design

Considerations in selecting a questionnaire for use with GORD patients include: the purpose for which the questionnaire was designed; evidence of validity, reliability, and responsiveness to change; the population on which it was tested and validated; the time frame of reference; and practical considerations of cost and copyright.[50] Kirshner and Guyatt[51] have distinguished between instruments designed for purposes of discrimination (diagnosis and screening), prediction (of response to treatment or of prognosis), and evaluation (assessing response to an intervention). They have highlighted that the relative weight accorded to the psychometric properties of validity, reliability, discriminatory power, and responsiveness to change will depend on the purpose. Although their focus was on quality of life instruments, the same distinctions are likely to be relevant with respect to symptom questionnaires. An instrument designed primarily as a diagnostic tool[16] may not be sufficiently responsive to change for use in the context of evaluation.

Moreover, properties of validity, reliability, and so on are not universal. For this reason, a questionnaire developed and validated in, say, a secondary care population in the USA may need to be adapted and revalidated for application in a primary care setting or in another culture. Issues of cross cultural adaptation are discussed elsewhere in this issue

(Wyrwich and Staebler Tardino[52] in this supplement *(see page iv45–iv48)*).

Some symptom questionnaires are in the public domain and may be freely used by researchers and practitioners. In other cases, access may require completion of a copyright or users' agreement, possibly with explicit conditions of use and/or a charge for reproduction of the questionnaire and access to data processing instructions. Practical barriers to usage include unduly restrictive conditions of use or excessive costs.

As with diaries, careful attention to questionnaire design and administration can enhance both the quantity of response and the quality of that response.[53 54] Jenkins and Dillman[55] provide guidelines for questionnaire design and layout, based on principles of cognitive psychology. A comprehensive review of methods to enhance the rate of response to questionnaire surveys is provided by Edwards and colleagues.[56]

TIMING OF ASSESSMENTS

The duration and frequency of symptom assessment must take into account the purpose of data collection and the natural history of the disease and anticipated response to any therapeutic intervention. There is often a need to strike a balance between what is desirable and what is feasible (for example, in terms of keeping respondent burden to an acceptable level).

Use of a reference time period

Dillman[53] defines behaviour to include ''what has happened in people's lives'', and this definition includes experience of symptoms. Sudman and Bradburn[57] advise that the reliability of responses to behaviour questions can be improved by asking about a specific time period. In relation to asking about symptoms, Steen and colleagues[18] state that the choice of reference period should take into account their prevalence and incidence. If the point prevalence is low and the reference period in question is very short, few patients may respond positively. Low item endorsement may reduce the utility of the measure in detecting differences between individuals and change over time.[58] Booth and colleagues[59] have reported significant daily fluctuations in symptoms of GORD. Their findings suggest that asking only about ''today'' (or, indeed, any single day) may produce an incomplete picture of overall symptom burden, and that longitudinal data, reflecting day to day variability in symptoms, may be required. However, as already noted, recall of minor symptoms is likely to be poor when the reference period exceeds one week.[38] Steen and colleagues[18] found that one month was the maximum period over which patients could provide reliable data on frequency of symptoms of asthma or diabetes, and on the impact of those symptoms. The time frames referred to in questionnaires appropriate to GORD patients vary, although reference periods of 2–4 weeks are common. The interaction between the choice of reference period and timing of assessment must also be considered. Clearly, a questionnaire referring to symptoms ''in the last four weeks'' would be inappropriate for use two weeks after initiation of therapy.

Duration of symptom recording

Burman[31] recommends that, in diary studies, the period of recording should represent a balance between the avoidance of excessive respondent burden and the necessity to capture adequately the anticipated fluctuations in the phenomenon of interest, as discussed above. Carp and Carp[44] found that problems reported on a selected day in their diary study were not an adequate predictor of problems across the other days of that week, and concluded that short diary periods could be problematic. However, protracted periods of data recording (in excess of six months) pose significant respondent burden and are, therefore, likely to be associated with poorer compliance with diary protocol.[60] Burman[31] indicates that typical diary periods tend to be 2–4 weeks when continuous data recording is required. An alternative strategy to reduce respondent burden when longer term data are required is for diaries to be completed only on a random selection of days in the data collection period[48] although with this strategy there is the risk that respondents will forget to fill in the diary on the designated dates.

ASSESSMENT OF MULTIPLE SYMPTOMS

In patients with GORD, the most characteristic symptom is heartburn but patients also report epigastric pain, regurgitation, belching, early satiety, acid reflux, dysphagia, and other symptoms.[26 61] Moreover, for some individuals heartburn is not the dominant symptom. Furthermore, therapeutic interventions may themselves give rise to symptomatic side effects, both within and beyond the digestive tract. For these reasons, consideration must be given to the number of symptoms to be measured, whether multiple items (questions) are needed to measure each symptom, whether it is appropriate to produce an overall assessment (for example, of symptom severity), and, if so, how that score should be computed.

How many symptoms, and how many dimensions

The issue of how many symptoms to assess will depend on the purpose of assessment. For diagnostic purposes, it is the combination and coincidence of symptoms that may serve to distinguish those patients who are more likely to be experiencing GORD from those with other disorders of the gastrointestinal system. In therapeutic interventions, the reduction of heartburn may be the primary objective but patients will generally also expect reduction of other symptoms and will not welcome the development of new symptoms as a side effect of therapy. In these situations, it may be desirable to assess the response of multiple symptoms to the intervention.

The experience of symptoms is likely to be multifaceted, including timing, frequency, duration, intensity, effects on role function, and distress caused by symptom occurrence.[18] Depending on the purpose of assessment, one or more facets of symptom experience may need to be measured. Keefe[62] has reported that pain intensity measures are flawed by focusing ''only on one dimension of multidimensional experience''.

Challenges of multiple measures

Although justified by the multifaceted experience of symptoms, the use of more than one measure presents problems of multiple end points in statistical analysis and interpretation. Similar problems arise in relation to other multidimensional outcome measures, in particular health related quality of life.[63 64] One solution is to define the symptom of greatest relevance as the primary end point (and in clinical trials or surveys to base power calculations on that item) and to treat the remaining symptoms as secondary end points. Another is to combine scores on individual symptoms to produce an overall measure of, say, symptom severity. The simplest way of doing this is simply to sum the individual responses. However, there are two assumptions implicit in this approach.[18] The first is that symptoms are being measured on an interval scale and that the ''distance'' between each pair of points on the scale is the same (for example, if symptoms are categorised as ''none'', ''mild'', ''moderate'', and ''severe'', that the difference between ''none'' and ''mild'' is the same as between ''mild'' and ''moderate'' or between ''moderate'' and ''severe''). The second is that all symptoms are given the same weight (that is, are of equal importance). Neither of these assumptions may be warranted.

Global assessments of symptoms

An alternative to asking individually about each symptom, and combining responses as described above, is to ask respondents to make a global assessment of their symptoms, using either a visual analogue scale or a set of verbal descriptors. This approach is also used in quality of life research.[65] Global assessments may, however, be cognitively demanding for respondents, requiring them to: judge which symptoms to include in the global rating; retrieve from memory specific and relevant information on individual symptoms; synthesise the information retrieved into a single overall judgement; and to ''map'' the adjudged response onto one of the response categories offered.[66] Respondents are likely to use a range of cognitive strategies and heuristics[41 42] in making their assessment, and may not use comparable frames of reference, thus threatening the validity and reliability of responses. Making a global rating may also be difficult when some symptoms are improving while others are unchanging or deteriorating—Ziebland and colleagues[67] observed that ''the global transition item … obscures a feature of (rheumatoid arthritis) activity whereby some functions may improve while others are in decline''. None the less, several studies in which global ratings have been compared with summated symptom scores have demonstrated reasonable agreement between the two approaches.[67–71] Global assessments of change over time or in response to an intervention have also been successfully used in quality of life research.[67 72 73]

ACKNOWLEDGEMENT

E McColl is a Primary Care Career Scientist, funded by the United Kingdom Department of Health's Primary Care Development Programme.

REFERENCES

1 **Venables TL**, Newland RD, Patel AC, *et al*. Omeprazole 10 milligrams once daily, omeprazole 20 milligrams once daily, or ranitidine 150 milligrams twice daily, evaluated as initial therapy for the relief of symptoms of gastro-oesophageal reflux disease in general practice. *Scand J Gastroenterol* 1997;**32**:965–73.
2 **Dent J**. Definitions of reflux disease and its separation from dyspepsia. *Gut* 2002;**50**(suppl IV):iv17–20.
3 **Barsky AJ**, Cleary PD, Barnett MC, *et al*. The accuracy of symptom reporting by patients complaining of palpitations. *Am J Med* 1994;**97**:214–21.
4 **Corley DA**, Cello JP, Koch J. Accuracy of endoscopic databases for assessing patient symptoms: comparison with self-reported questionnaires in patients infected with the human immunodeficiency virus. *Gastrointest Endosc* 2000;**51**:129–33.
5 **Fontaine A**, Larue F, Lassauniere JM. Physicians' recognition of the symptoms experienced by HIV patients: how reliable? *J Pain Symptom Manage* 1999;**18**:263–70.
6 **Justice AC**, Rabeneck L, Hays RD, *et al*. Sensitivity, specificity, reliability, and clinical validity of provider-reported symptoms: a comparison with self-reported symptoms. Outcomes Committee of the AIDS Clinical Trials Group. *J Acquir Immune Defic Syndr Hum Retrovirol* 1999;**21**:126–33.
7 **Justice AC**, Chang CH, Rabeneck L, *et al*. Clinical importance of provider-reported HIV symptoms compared with patient-report. *Med Care* 2001;**39**:397–408.
8 **Kwoh CK**, Ibrahim SA. Rheumatology patient and physician concordance with respect to important health and symptom status outcomes. *Arthritis Rheum* 2001;**45**:372–7.
9 **Sandmark S**, Carlsson R, Fausa O, *et al*. Omeprazole or ranitidine in the treatment of reflux esophagitis. Results of a double-blind, randomized, Scandinavian multicenter study. *Scand J Gastroenterol* 1988;**23**:625–32.
10 **Stephens RJ**, Hopwood P, Girling DJ, *et al*. Randomized trials with quality of life endpoints: are doctors' ratings of patients' physical symptoms interchangeable with patients' self-ratings? *Qual Life Res* 1997;**6**:225–36.
11 **Heading RC**, Wager E, Tooley PJH. Reliability of symptom assessment in dyspepsia. *Eur J Gastroenterol Hepatol* 1997;**9**:779–81.
12 **Holmes WF**, MacGregor EA, Sawyer JP, *et al*. Information about migraine disability influences physicians' perceptions of illness severity and treatment needs. *Headache* 2001;**41**:343–50.
13 **Agreus L**, Talley NJ, Svardsudd K, *et al*. Identifying dyspepsia and irritable bowel syndrome: the value of pain or discomfort, and bowel habit descriptors. *Scand J Gastroenterol* 2000;**35**:142–51.
14 **Agreus L**, Svardsudd K, Nyren O, *et al*. The epidemiology of abdominal symptoms: prevalence and demographic characteristics in a Swedish adult population. A report from the Abdominal Symptoms Survey. *Scand J Gastroenterol* 1994;**29**:102–9.
15 **Stanghellini V**. Review article: pain versus discomfort—is differentiation clinically useful? *Aliment Pharmacol Ther* 2001;**15**:145–9.
16 **Carlsson R**, Dent J, Bolling-Sternevald E, *et al*. The usefulness of a structured questionnaire in the assessment of symptomatic gastroesophageal reflux disease. *Scand J Gastroenterol* 1998;**33**:1023–9.
17 **Locke GR**, Talley NJ, Weaver AL, *et al*. A new questionnaire for gastroesophageal reflux disease. *Mayo Clin Proc* 1994;**69**:539–47.
18 **Steen IN**, McColl E, Hutchinson A. Developing symptom based outcome measures. In: Hutchinson A, McColl E, Christie M, *et al*, eds. *Outcome measures in primary and outpatient care*. Reading: Harwood Academic Publishers, 1996:23–44.
19 **Shiffman S**, Hufford MR, Paty JA. Subject experience diaries in clinical research. Part 1: The patient experience movement. *Appl Clin Trials* 2001;February:46–56.
20 **Junghard O**, Lauritsen K, Talley NJ, *et al*. Validation of seven graded diary cards for severity of dyspeptic symptoms in patients with non ulcer dyspepsia. *Eur J Surg Suppl* 1998;**583**:106–11.
21 **Sandha GS**, Hunt RH, Veldhuyzen Van Zanten SJ. A systematic overview of the use of diary cards, quality-of-life questionnaires, and psychometric tests in treatment trials of Helicobacter pylori-positive and -negative non-ulcer dyspepsia. *Scand J Gastroenterol* 1999;**34**:244–9.
22 **Leibbrand R**, Cuntz U, Hiller W. Assessment of functional gastrointestinal disorders using the Gastro-Questionnaire. *Int J Behav Med* 2002;**9**:155–72.
23 **Pope CE**. The quality of life following antireflux surgery. *World J Surg* 1992;**16**:355–8.
24 **Rattner DW**, Brooks DC. Patient satisfaction following laparoscopic and open antireflux surgery. *Arch Surg* 1995;**130**:289–93.
25 **Revicki DA**, Wood M, Wiklund I, *et al*. Reliability and validity of the Gastrointestinal Symptom Rating Scale in patients with gastroesophageal reflux disease. *Qual Life Res* 1998;**7**:75–83.
26 **Rothman M**, Farup C, Stewart W, *et al*. Symptoms associated with gastroesophageal reflux disease: development of a questionnaire for use in clinical trials. *Dig Dis Sci* 2001;**46**:1540–9.
27 **Rush DR**, Stelmach WJ, Young TL, *et al*. Clinical effectiveness and quality of life with ranitidine vs placebo in gastroesophageal reflux disease patients: a clinical experience network (CEN) study. *J Fam Pract* 1995;**41**:126–36.
28 **Shaw M**, Talley NJ, Adlis S, *et al*. Development of a digestive health status instrument: tests of scaling assumptions, structure and reliability in a primary care population. *Aliment Pharmacol Ther* 1998;**12**:1067–78.
29 **Svedlund J**, Sjodin I, Dotevall G. GSRS—a clinical rating scale for gastrointestinal symptoms in patients with irritable bowel syndrome and peptic ulcer disease. *Dig Dis Sci* 1988;**33**:129–34.
30 **Velanovich V**, Vallance SR, Gusz JR, *et al*. Quality of life scale for gastroesophageal reflux disease. *J Am Coll Surg* 1996;**183**:217–24.
31 **Burman ME**. Health diaries in nursing research and practice. *Image J Nurs Sch* 1995;**27**:147–52.
32 **Hyland ME**, Kenyon CA, Allen R, *et al*. Diary keeping in asthma: comparison of written and electronic methods. *BMJ* 1993;**306**:487–9.
33 **Hyland ME**, Crocker GR. Validation of an asthma quality of life diary in a clinical trial. *Thorax* 1995;**50**:724–30.
34 **Hufford MR**, Shiffman S. Methodological issues affecting the value of patient-reported outcomes data. *Expert Rev Pharmacoeconomics Outcomes Res* 2002;**2**:119–28.
35 **Stone AA**, Shiffman S, Schwartz JE, *et al*. Patient non-compliance with paper diaries. *BMJ* 2002;**324**:1193–4.
36 **Roghmann KJ**, Haggerty RJ. The diary as a research instrument in the study of health and illness behavior: experiences with a random sample of young families. *Med Care* 1972;**10**:143–63.
37 **Verbrugge LM**. Health diaries. *Med Care* 1980;**18**:73–95.
38 **Dahlquist G**, Wall S, Ivarsson JI, *et al*. Health problems and care in young families—an evaluation of survey procedures. *Int J Epidemiol* 1984;**13**:221–8.
39 **Feine JS**, Lavigne GJ, Dao TT, *et al*. Memories of chronic pain and perceptions of relief. *Pain* 1998;**77**:137–41.
40 **Price DD**, Milling LS, Kirsch I, *et al*. An analysis of factors that contribute to the magnitude of placebo analgesia in an experimental paradigm. *Pain* 1999;**83**:147–56.
41 **Bradburn NM**, Rips LJ, Shevell SK. Answering autobiographical questions: The impact of memory and inference on surveys. *Science* 1987;**236**:157–61.
42 **Bradburn NM**. Temporal representation and event dating. In: Stone AA, Turkkan JS, Bachrach C, *et al*, eds. *The science of self-report: implications for research and practice*. Mahwah, NJ: Lawrence Erlbaum Associates Inc, 2000:49–61.
43 **Steen IN**, Hutchinson A, McColl E, *et al*. Development of a symptom-based outcome measures for asthma. *BMJ* 1994;**309**:1065–9.
44 **Carp FM**, Carp A. The validity, reliability and generalizability of diary data. *Exp Aging Res* 1981;**7**:281–96.
45 **Sudman S**, Lannom LB. *Health care survey using diaries*. Washington DC: National Center for Health Services Research, USDHHS, 1980 (PHS Research Report Series PHS 80–3279).
46 **Rakowski W**, Julius M, Hickey T, *et al*. Daily symptoms and behavioral responses. Results of a health diary with older adults. *Med Care* 1988;**26**:278–97.
47 **Kimble LP**, Dunbar SB, McGuire DB, *et al*. Cardiac instrument development in a low-literacy population: the revised Chest Discomfort Diary. *Heart Lung* 2001;**30**:312–20.
48 **Norman GR**, McFarlane AH, Streiner DL, *et al*. Health diaries: strategies for compliance and relation to other measures. *Med Care* 1982;**20**:623–9.
49 **Hufford MR**, Stone AA, Shiffman S, *et al*. Paper vs. Electronic diaries: compliance and subject evaluations, *Appl Clin Trials* 2002;August:38–43.

50 **Bentzen N**, Christiansen T, McColl E, *et al*. Selection and cross-cultural adaptation of health outcome measures. *Eur J Gen Pract* 1998;**4**:27–33.
51 **Kirshner B**, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* 1985;**38**:27–36.
52 **Wyrwich KW**, Staebler Tardino VM. A blueprint for symptom scales and responses: measurement and reporting. *Gut* 2004;**53**(suppl IV):iv45–8.
53 **Dillman DA**. *Mail and telephone surveys: The total design method*. New York: John Wiley and Sons Inc, 1978.
54 **Dillman DA**. *Mail and internet surveys: the tailored design methods*. New York: John Wiley and Sons Inc, 2000.
55 **Jenkins CR**, Dillman DA. Towards a theory of self-administered questionnaire design. In: Lyberg L, Biemer P, Collins M, *et al*, eds. *Survey measurement and process quality*. New York: John Wiley and Sons Inc, 1997:165–96.
56 **Edwards P**, Roberts I, Clarke M, *et al*. Increasing response rates to postal questionnaires: systematic review. *BMJ* 2002;**324**:1183–91.
57 **Sudman S**, Bradburn N. *Asking questions: a practical guide to questionnaire design*. San Francisco: Jossey-Bass, 1982.
58 **Streiner DL**, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press, 1989.
59 **Booth MI**, Stratford J, Dehn TCB. Patient self-assessment of test-day symptoms in 24-h pH-metry for suspected gastroesophageal reflux disease. *Scand J Gastroenterol* 2001;**36**:795–9.
60 **Gold DR**, Weiss ST, Tager IB, *et al*. Comparison of questionnaire and diary methods in acute childhood respiratory illness surveillance. *Am Rev Respir Dis* 1989;**139**:847–9.
61 **Kuykendall DH**, Rabeneck L, Campbell CJ, *et al*. Dyspepsia: how should we measure it? *J Clin Epidemiol* 1998;**51**:99–106.
62 **Keefe F**. Self-report of pain: issues and opportunities. In: Stone AA, Turkkan JS, Bachrach C, *et al*, eds. *The science of self-report: implications for research and practice*. Mahwah, NJ: Lawrence Erlbaum Associates Inc, 2000:317–37.
63 **Fayers P**, Machin D. *Quality of life: assessment, analysis and interpretation*. Chichester: John Wiley and Sons Inc, 2000.
64 **Fairclough DL**. *Design and analysis of quality of life studies in clinical trials: interdisciplinary statistics*. Boca Raton, FL: Chapman and Hall/CRC, 2002.
65 **Rowan K**. Global questions and scores. In: Jenkinson C, ed. *Measuring health and medical outcomes*. London: UCL Press, 1994:54–76.
66 **Tourangeau R**, Rips LJ, Rasinski K. *The psychology of survey response*. Cambridge: Cambridge University Press, 2000.
67 **Ziebland S**, Fitzpatrick R, Jenkinson C, *et al*. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. *Ann Rheum Dis* 1992;**51**:1202–5.
68 **Crabtree HL**, Hildreth AJ, O'Connell JE, *et al*. Measuring visual symptoms in British cataract patients: the cataract symptom scale. *Br J Ophthalmol* 1999;**83**:519–23.
69 **Hagg O**, Fritzell P, Oden A, *et al*. Simplifying outcome measurement: evaluation of instruments for measuring outcome after fusion surgery for chronic low back pain. *Spine* 2002;**27**:1213–22.
70 **Russell ML**, Preshaw RM, Brant RF, *et al*. Disease-specific quality of life: the Gallstone Impact Checklist. *Clin Invest Med* 1996;**19**:453–60.
71 **Talley NJ**, Verlinden M, Jones M. Validity of a new quality of life scale for functional dyspepsia: a United States multicenter trial of the Nepean Dyspepsia Index. *Am J Gastroenterol* 1999;**94**:2390–7.
72 **Juniper EF**, Guyatt GH, Willan A, *et al*. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol* 1994;**47**:81–7.
73 **Jaeschke R**, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–15.