Part 1
Developing and Validating Instruments for Assessing Quality of Life and Patient-Reported Outcomes

1
Introduction

Summary

A key methodology for the evaluation of therapies is the randomised controlled trial (RCT). These clinical trials traditionally considered relatively objective clinical outcome measures, such as cure, biological response to treatment, or survival. Later, investigators and patients alike have argued that subjective indicators should also be considered. These subjective patient-reported outcomes are often regarded as indicators of quality of life. They comprise a variety of outcome measures, such as emotional functioning (including anxiety and depression), physical functioning, social functioning, pain, fatigue, other symptoms and toxicity. A large number of questionnaires, or *instruments*, have been developed for assessing patient-reported outcomes and quality of life, and these have been used in a wide variety of circumstances. This book is concerned with the development, analysis and interpretation of data from these quality of life instruments.

1.1 Patient-reported outcomes

This book accepts a broad definition of *quality of life*, and discusses the design, application and use of single- and multi-item, subjective, measurement scales. This encompasses not just 'overall quality of life' but also the symptoms and side effects that may or may not reflect – or affect – quality of life. Some researchers prefer to emphasise that we are only interested in health aspects, as in *health-related quality of life* (HRQoL or HRQL), while others adopt the terms *patient-reported outcomes* (PROs) or *patient-reported outcome measures* (PROMs), because those terms indicate interest in a whole host of outcomes, such as pain, fatigue, depression through to physical symptoms such as nausea and vomiting. But not all subjects are 'patients' who are ill; it is also suggested that PRO could mean *person-reported outcome*. *Health outcomes assessment* has also been proposed, which emphasises that the focus is on health issues and also avoids specifying the respondent: for young children and for the cognitively impaired we may use *proxy assessment* for cognitive reasons. And for

Quality of Life: The Assessment, Analysis and Reporting of Patient-Reported Outcomes, Third Edition.
Peter M. Fayers and David Machin.

many years some questionnaires have focused on *health status* or *self-reported health* (SRH), with considerable overlap to *quality of life*.

From a measurement perspective, this book is concerned with all the above. For simplicity we will use the now well-established overall term *quality of life* (QoL) to indicate (a) the set of outcomes that contribute to a patient's well-being or overall health, or (b) a summary measure or scale that purports to describe a patient's overall well-being or health. Examples of summary measures for QoL include general questions such as 'How good is your overall quality of life?' or 'How do you rate your overall health?' that represent global assessments. When referring to outcomes that reflect individual dimensions, we use the acronym PROs. Examples of PROs are pain or fatigue; symptoms such as headaches or skin irritation; function, such as social and role functioning; issues such as body image or existential beliefs; and so on. Mostly, we shall assume the respondent is the patient or person whose experience we are interested in (*self-report*), but it could be a proxy.

The measurement issues for all these outcomes are similar. Should we use single- or multi-item scales? Content and construct validity – are we measuring what we intend? Sensitivity, reliability, responsiveness – is the assessment statistically adequate? How should such assessments be incorporated into clinical studies? And how do we analyse, report and interpret the results?

1.2 What is a patient-reported outcome?

The definition of *patient-reported outcome* is straightforward, and has been described as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" (US FDA, 2009). A PRO can be measured by self-report or by interview provided that the interviewer records only the patient's response. The outcome can be measured in absolute terms (e.g. severity of a symptom, sign or state of a disease) or as a change from a previous assessment.

1.3 What is *quality of life*?

In contrast to PRO, the term *Quality of life* is ill defined. The World Health Organization (WHO, 1948) declares health to be 'a state of complete physical, mental and social well-being, and not merely the absence of disease'. Many other definitions of both 'health' and 'quality of life' have been attempted, often linking the two and, for QoL, frequently emphasising components of happiness and satisfaction with life. In the absence of any universally accepted definition, some investigators argue that most people, in the Western world at least, are familiar with the expression 'quality of life' and have an intuitive understanding of what it comprises.

However, it is clear that 'QoL' means different things to different people, and takes on different meanings according to the area of application. To a town planner, for

example, it might represent access to green space and other facilities. In the context of clinical trials we are rarely interested in QoL in such a broad sense, and instead are concerned only with evaluating those aspects that are affected by disease or treatment for disease. This may sometimes be extended to include indirect consequences of disease, such as unemployment or financial difficulties. To distinguish between QoL in its more general sense and the requirements of clinical medicine and clinical trials the term *health-related quality of life* (HRQoL) is frequently used in order to remove ambiguity.

Health-related QoL is still a loose definition. What aspects of QoL should be included? It is generally agreed that the relevant aspects may vary from study to study but can include general health, physical functioning, physical symptoms and toxicity, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, sexual functioning and existential issues. In the absence of any agreed formal definition of QoL, most investigators circumvent the issues by describing what *they* mean by QoL, and then letting the items (questions) in their questionnaire speak for themselves. Thus some questionnaires focus upon the relatively objective signs such as patient-reported toxicity, and in effect define the relevant aspects of QoL as being, for their purposes, limited to treatment toxicity. Other investigators argue that what matters most is the impact of toxicity, and therefore their questionnaires place greater emphasis upon psychological aspects, such as anxiety and depression. Yet others try to allow for spiritual issues, ability to cope with illness and satisfaction with life.

Some QoL instruments focus upon a single concept, such as emotional functioning. Other instruments regard these individual concepts as aspects, or *dimensions*, of QoL, and therefore include items relating to several concepts. Although there is disagreement about what components should be evaluated, most investigators agree that a number of the above dimensions should be included in QoL questionnaires, and that QoL is a multidimensional construct. Because there are so many potential dimensions, it is impractical to try to assess all these concepts simultaneously in one instrument. Most instruments intended for health-status assessment include at least some items that focus upon physical, emotional and social functioning. For example, if emotional functioning is accepted as being one aspect of QoL that should be investigated, several questions could evaluate anxiety, tension, irritability, depression and so on. Thus instruments may contain many items. Although a single global question such as 'How would you rate your overall quality of life?' is a useful adjunct to multi-item instruments, global questions are often regarded as too vague and non-specific to be used on their own. Most of the general questionnaires that we describe include one or more global questions alongside a number of other items covering specific issues. Some instruments place greater emphasis upon the concept of global questions, and the EQ-5D questionnaire (Appendix E4) asks a parsimonious five questions before using a single global question that enquires about 'your health'. Even more extreme is the Perceived Adjustment to Chronic Illness Scale (PACIS) described by Hürny *et al*. (1993). This instrument consists of a single, carefully phrased question that is a global indicator of coping and adjustment: 'How much effort does it cost you to cope with your illness?' This takes responses ranging between 'No effort at all' and 'A great deal of effort'.

One unifying and non-controversial theme throughout all the approaches is that the concepts forming these dimensions can be assessed only by *subjective measures*, PROs, and that they should be evaluated by asking the patient. *Proxy* assessments, by a relative or other close observer, are usually employed only if the patient is unable to make a coherent response, for example those who are very young, very old, severely ill or have mental impairment. Furthermore, many of these individual concepts – such as emotional functioning and fatigue – lack a formal, agreed definition that is universally understood by patients. In many cases the problem is compounded by language differences, and some concepts do not readily translate to other tongues. There are also cultural differences regarding the importance of the issues. Single-item questions on these aspects of QoL, as for global questions about overall QoL, are likely to be ambiguous and unreliable. Therefore it is usual to develop questionnaires that consist of multi-item measurement scales for each concept.

1.4 Historical development

One of the earliest references that impinges upon a definition of QoL appears in the *Nichomachean Ethics*, in which Aristotle (384–322 BCE) notes: "Both the multitude and persons of refinement … conceive 'the good life' or 'doing well' to be the same thing as 'being happy'. But what constitutes happiness is a matter of dispute … some say one thing and some another, indeed very often the same man says different things at different times: when he falls sick he thinks health is happiness, when he is poor, wealth." The Greek *ευδαιμονια* is commonly translated as 'happiness' although Rackham, the translator that we cite, noted that a more accurate rendering would embrace 'well-being', with Aristotle denoting by *ευδαιμονια* both a state of feeling and a kind of activity. In modern parlance this is assuredly quality of life. Although the term 'quality of life' did not exist in the Greek language of 2000 years ago, Aristotle clearly appreciated that QoL means different things to different people. He also recognised that it varies according to a person's current situation – an example of a phenomenon now termed response shift. QoL was rarely mentioned until the twentieth century, although one early commentator on the subject noted that happiness could be sacrificed for QoL: "Life at its noblest leaves mere happiness far behind; and indeed cannot endure it … Happiness is not the object of life: life has no object: it is an end in itself; and courage consists in the readiness to sacrifice happiness for an intenser quality of life" (Shaw, [1900] 1972). It would appear that by this time 'quality of life' had become a familiar term that did not require further explanation. Specific mention of QoL in relation to patients' health came much later. The influential WHO 1948 definition of health cited above was one of the earliest statements recognising and stressing the importance of the three dimensions – physical, mental and social – in the context of disease. Other definitions have been even more general: "Quality of Life: Encompasses the entire range of human experience, states, perceptions, and spheres of thought concerning the life of an individual or a community. Both objective and subjective, quality-of-life can include cultural, physical, psychological, interpersonal, spiritual, financial, political,

temporal, and philosophical dimensions. Quality-of-life implies judgement of value placed on experience of communities, groups such as families, or individuals" (Patrick and Erickson, 1993).

One of the first instruments that broadened the assessment of patients beyond physiological and clinical examination was the Karnofsky Performance Scale proposed in 1947 (Karnofsky and Burchenal, 1947) for use in clinical settings. This is a simple scale ranging from 0 for 'dead' to 100 indicating 'normal, no complaints, no evidence of disease'. Healthcare staff make the assessment. Over the years, it has led to a number of other scales for functional ability, physical functioning and *activities of daily living* (ADL), such as the Barthel Index. Although these questionnaires are still sometimes described as QoL instruments, they capture only one aspect of it and provide an inadequate representation of patients' overall well-being and QoL.

The next generation of questionnaires, in the late 1970s and early 1980s, that quantified health status were used for the general evaluation of health. These instruments focused on physical functioning, physical and psychological symptoms, impact of illness, perceived distress and life satisfaction. Examples of such instruments include the Sickness Impact Profile (SIP) and the Nottingham Health Profile (NHP). Although these instruments are frequently described as QoL questionnaires, their authors neither designed them nor claimed them as QoL instruments.

Meanwhile, Priestman and Baum (1976) were adapting linear analogue self-assessment (LASA) methods to assess QoL in breast cancer patients. The LASA approach, which is also sometimes called a visual analogue scale (VAS), provides a 10 cm line, with the ends labelled with words describing the extremes of a condition. The patient is asked to mark the point along the line that corresponds with their feelings. An example of a LASA scale is contained in the EQ-5D (Appendix E4). Priestman and Baum (1976) measured a variety of subjective effects, including well-being, mood, anxiety, activity, pain, social activities and the patient's opinion as to 'Is the treatment helping?' Others took the view that one need only ask a single question to evaluate the QoL of patients with cancer: "How would you rate your QoL today?" (Gough *et al.*, 1983), and supported their position by demonstrating a relatively strong correlation between answers to this single question and scores derived from a more extensive battery of questionnaires.

Much of the development of QoL instruments has built upon these early attempts, first with increasing emphasis on the more subjective aspects, such as emotional, role, social and cognitive functioning, but subsequently with a counter trend towards greater focus on patient-reported symptoms and other relatively objective outcomes. Frequently, one or more general or global questions concerning overall QoL are included. Implicit in all this is that psychological and social aspects, functional capacity and symptomatology all relate to QoL. Thus, if a patient is unable to achieve full physical, psychological or social functioning, it is assumed that their QoL is poorer. Although this may in general seem a reasonable assumption, it can lead to theoretical problems. In particular, many forms of functioning, especially physical functioning, may be regarded as *causal* variables that can be expected to change or affect a patient's QoL but do not necessarily reflect the true level of their QoL (see Section 2.6). For

example, a patient may have a poor QoL irrespective of whether their physical functioning is impaired; this might arise because of other factors, such as pain. Therefore scales measuring functional status assess only whether there are problems that *may* cause distress to the patient or impair their QoL; absence of problems in these specific areas does not indicate that a patient has no problems at all, nor does it necessarily indicate that a patient has good QoL. Despite these reservations, most instruments continue to focus on health status, functional status and checklists of symptoms. For clinical purposes this may be logical since, when comparing treatments, the clinician is most concerned with the differences in the symptoms and side effects due to the various therapies, and the impact of these differences upon QoL.

A number of other theoretical models for QoL have been proposed. The *expectations* model of Calman (1984) suggests that individuals have aims and goals in life and that QoL is a measure of the difference between the hopes and expectations of the individual and the individual's present experience. It is concerned with the difference between perceived goals and actual goals. The gap may be narrowed by improving the function of a patient or by modifying their expectations. Instruments such as the Schedule for Evaluation of Individual Quality of Life (SEIQoL) and the Patient Generated Index (PGI), described in Section 1.8, use Calman's expectations model as a conceptual basis and provide the facility to incorporate personal values.

The *needs* model relates QoL to the ability and capacity of patients to satisfy certain human needs. QoL is at its highest when all needs are fulfilled, and at its lowest when few needs are satisfied. Needs include such aspects as identity, status, self-esteem, affection, love, security, enjoyment, creativity, food, sleep, pain avoidance and activity. Hunt and McKenna (1992) use this model to generate several QoL measures. Somewhat related is the *reintegration to normal living* model that has also been regarded as an approach to assessing QoL. Reintegration means the ability to do what one has to do or wants to do, but it does not mean being free of disease or symptoms.

Other definitions or indicators of QoL that have been suggested are *personal well-being* and *satisfaction with life*. The *impact of illness* (or treatment) on social, emotional, occupational and family domains emphasises the illness aspect, or the *interference* of symptoms and side effects. The *existential* approach notes that preferences are not fixed and are both individual and vary over time – as was recognised so long ago by Aristotle. Having a 'positive approach to life' can give life high quality, regardless of the medical condition. Therefore it can be important to assess existential beliefs and also *coping*. A patient's perception of their QoL can be altered by influencing their existential beliefs or by helping them to cope better. The existential model of QoL leads to the inclusion of such items as pleasure in life and positive outlook on life.

Patient preference measures differ from other models of QoL in that they explicitly incorporate *weights* that reflect the importance that patients attach to specific dimensions. Different states and dimensions are compared against each other, to establish a ranking in terms of their value or in terms of patients' preferences of one state over another. These and other *utility measure* approaches to QoL assessment are derived from decision-making theory and are frequently employed in any economic evaluations of treatments.

Finally, some authorities simply circumvent the challenges of defining QoL. The US FDA (2009) notes: "Quality of life — A general concept that implies an evaluation of the effect of all aspects of life on general well-being. Because this term implies the evaluation of nonhealth-related aspects of life, and because the term generally is accepted to mean *what the patient thinks it is*, it is too general and undefined to be considered appropriate for a medical product claim." Further, they also define "Health-related quality of life (HRQL) — HRQL is a multidomain concept that represents the patient's general perception of the effect of illness and treatment on physical, psychological, and social aspects of life", but also add the hard-to-satisfy condition that "Claiming a statistical and meaningful improvement in HRQL implies: (1) that all HRQL domains that are important to interpreting change in how the clinical trial's population feels or functions as a result of the targeted disease and its treatment were measured; (2) that a general improvement was demonstrated; and (3) that no decrement was demonstrated in any domain" (US FDA, 2009).

Thus there is continuing philosophical debate about the meaning of QoL and about what should be measured. Perhaps the simplest and most pragmatic view is that all of these concepts reflect issues that are of fundamental importance to patients' well-being. They are all worth investigating and quantifying.

1.5 Why measure quality of life?

There are several reasons why QoL assessments may be included in RCTs, and it is important to distinguish between them as the nature of the measurements, and the questionnaires that are employed, will depend upon the objectives of the trial. Perhaps the most obvious reason is in order to compare the study treatments, in which case it is important to identify those aspects of QoL that may be affected by the therapy. These include both benefits, as may be sought in palliative trials that are expected to improve QoL, and negative changes, such as toxicity and any side effects of therapy.

Clinical trials of treatment with curative intent

Many clinical trial organisations have now introduced the assessment of QoL as being a standard part of new trials. An obvious reason for the emphasis towards QoL as an important endpoint is that treatment of fatal diseases can, and often does, result in limited gains in cure or prolonged survival. With some notable exceptions, little improvement has been seen in patients with major cancer diagnoses, HIV or AIDS. At the same time therapeutic interventions in these diseases frequently cause serious side effects and functional impairment.

There are numerous examples in which QoL assessments have had an unexpectedly important role in the interpretations and conclusions of RCTs, and it is perhaps surprising that it took so long for the relevance of QoL assessment to be appreciated. For example, one of the earliest randomised trials to include QoL assessment was by Coates *et al*. (1987), who reported that, contrary to their initial expectations,

continuous as opposed to intermittent chemotherapy for advanced breast cancer not only prolonged survival but most importantly resulted in a superior QoL.

Similarly, other RCTs recognised that QoL may be the principal outcome of interest. For example, an RCT comparing three anti-hypertensive therapies conducted by Croog *et al*. (1986) demonstrated major differences in QoL, and the results of a cancer chemotherapy trial of Buccheri *et al*. (1989) suggested that small treatment benefits may be more than outweighed by the poorer QoL and cost of therapy. In extreme cases, the cure might be worse than the disease.

Example from the literature

Testa *et al*. (1993) describe an RCT evaluating hypertensive therapy in men. Two angiotensin-converting enzyme inhibitors, captopril and enalapril, were compared. In total, 379 active men with mild to moderate hypertension, aged 55 to 79, were randomised between the treatment arms. QoL was one of the main outcome measures. Several QoL scales were used, including an Overall QoL scale based on a mean score from 11 subscales.

In order to interpret the magnitude of the differences in QoL that were observed, stressful life events that produced an equivalent change in QoL scores were considered, and the responses to the Overall QoL scale were re-calibrated accordingly. Overall QoL scores shifted positively for captopril by 0.11 units, and negatively for enalapril by 0.11. Negative shifts of 0.11 corresponded to those encountered when there was 'major change in work responsibility', 'in-law troubles' or 'mortgage foreclosure'. On the basis of these investigations, a clinically important change was deemed to be one between 0.1 and 0.2.

It was concluded that, although the therapies were indistinguishable in terms of clinical assessments of efficacy and safety, they produced substantial and different changes in QoL.

Clinical trials of treatment with palliative intent

One consequence of ageing societies is the corresponding increased prevalence of chronic diseases. The treatment outcome in such diseases cannot be cure but must relate to the improvement of the well-being of patients thus treated. The aim is to palliate symptoms, or to prolong the time without symptoms. Traditionally, clinical and not QoL outcomes have been the principal endpoints. For example, in an RCT of therapy for advanced oesophageal cancer, absence of dysphagia might have been the main outcome measure indicating success of therapy. Nowadays, in trials of palliative care, QoL is more frequently chosen as the outcome measure of choice. Symptom relief is now recognised as but one aspect of palliative intervention, and a comprehensive assessment of QoL is often as important as an evaluation of symptoms.

Example from the literature

Temel *et al*. (2010) randomly assigned patients with newly diagnosed metastatic non–small-cell lung cancer to receive either early palliative care integrated with standard oncologic care or standard oncologic care alone. Quality of life and mood were assessed at baseline and at 12 weeks with the use of the Functional Assessment of Cancer Therapy–Lung (FACT-L) scale and the Hospital Anxiety and Depression Scale. The primary outcome was the change in the quality of life at 12 weeks.

Of the 151 randomised patients, 27 died by 12 weeks and 107 (86% of the remaining patients) completed assessments. Patients assigned to early palliative care had a better quality of life than did patients assigned to standard care (mean score on the FACT-L scale, 98.0 vs. 91.5; $p = 0.03$). In addition, fewer patients in the palliative care group than in the standard care group had depressive symptoms (16% vs. 38%, $p = 0.01$). Despite the fact that fewer patients in the early palliative care group than in the standard care group received aggressive end-of-life care (33% vs. 54%, $p = 0.05$), median survival was longer among patients receiving early palliative care (11.6 months vs. 8.9 months, $p = 0.02$).

Temel *et al*. conclude that, among patients with metastatic non-small-cell lung cancer, early palliative care led to significant improvements in both quality of life and mood. As compared with patients receiving standard care, patients receiving early palliative care had less aggressive care at the end of life but longer survival.

Example from the literature

Fatigue, lethargy, anorexia, nausea and weakness are common in patients with advanced cancer. It had been widely believed at the time that progestagens, including megestrol acetate (MA), might have a useful function for the palliative treatment of advanced endocrine-insensitive tumours. Beller *et al*. (1997) report a double-blind RCT of 240 patients randomised to 12 weeks of high- or low-dose MA, or to matching placebo. Nutritional status was recorded, and QoL was measured using six LASA scales, at randomisation and after four, eight and 12 weeks.

Patients receiving MA reported substantially better appetite, mood and overall QoL than patients receiving placebo, with a larger benefit being seen for the higher dose. Table 1.1 shows the average change from the baseline at time of randomisation. No statistically significant differences were observed in the nutritional status measurements. Side effects of therapy were minor and did not differ across treatments.

The authors conclude that high-dose MA provides useful palliation for patients with endocrine-insensitive advanced cancer. It improves appetite, mood and overall QoL in these patients, although not through a direct effect on nutritional status.

Table 1.1 Average difference in QoL between baseline and subsequent weeks

LASA scores	Placebo	Low dose MA	High dose MA	*p*-value (trend)
Physical well-being	5.8	6.5	13.9	0.13
Mood	−4.1	0.4	10.2	0.001
Pain	−5.3	−6.9	1.9	0.13
Nausea/vomiting	−1.4	8.7	7.2	0.08
Appetite	9.7	17.0	31.3	0.0001
Overall QoL	−2.7	2.8	13.1	0.001
Combined QoL measure	−2.1	2.4	12.3	0.001

Source: Beller *et al.*, 1997, Table 3. Reproduced with permission of Oxford University Press.

Improving symptom relief, care or rehabilitation

Traditionally, medicine has tended to concentrate upon symptom relief as an outcome measure. Studies using QoL instruments may reveal other issues that are equally or more important to patients. For example, in advanced oesophageal cancer, it was found that many patients say that fatigue has a far greater impact upon their QoL than dyspnoea. Such a finding is contrary to traditional teaching, but has been replicated in many other cancer sites.

Example from the literature

Smets *et al.* (1998) used the Multidimensional Fatigue Inventory (MFI-20, Appendix E14) to assess fatigue in 250 patients who were receiving radiotherapy with curative intent for various cancers. Patients rated their fatigue at two-weekly intervals during treatment and within two weeks after completion of radiotherapy.

There was a gradual increase in fatigue during radiotherapy and a decrease after completion of treatment. After treatment, 46% of the patients reported fatigue as being among the three symptoms causing most distress, and 40% reported having been tired throughout the treatment period.

Smets *et al.* conclude that there is a need to give preparatory information to new patients who are at risk of fatigue, and interventions including exercise and psychotherapy may be beneficial. They also suggested that their results might be underestimations because the oldest and most tired patients were more inclined to refuse participation.

Rehabilitation programmes, too, have traditionally concentrated upon physical aspects of health, functioning and ability to perform ADLs; these physical aspects were most frequently evaluated by healthcare workers or other observers. Increasingly, patient-completed QoL assessment is now perceived as essential to the evaluation of successful rehabilitation. Problems revealed by questioning patients can lead to

modifications and improvement in the programme, or alternatively may show that some methods offer little benefit.

Example from the literature

Results from several small trials had suggested that group therapy, counselling, relaxation therapy and psychoeducation might have a role in the rehabilitation of patients following acute myocardial infarction. Jones and West (1996) report an RCT that examined the impact of psychological rehabilitation after myocardial infarction. In this trial, 2328 patients were randomised between policies of no-intervention and intervention consisting of comprehensive rehabilitation with psychological therapy and opportunities for group and individual counselling. Patients were assessed both by interview and by questionnaires for anxiety and depression, state anxiety, expectations of future life, psychological well-being, sexual activity and functional disability.

At six months, 34% of patients receiving intervention had clinically significant levels of anxiety, compared with 32% of no-intervention patients. In both groups, 19% had clinically significant levels of depression. Differences for other domains were also minimal.

The authors conclude that rehabilitation programmes based upon psychological therapy, counselling, relaxation training and stress management seem to offer little objective benefit to myocardial infarction patients.

Facilitating communication with patients

Another reason for assessing QoL is to establish information about the range of problems that affect patients. In this case, the investigator may be less interested in whether there are treatment differences, and might even anticipate that both study arms will experience similar levels of some aspects of QoL. The aim is to collect information in a form that can be communicated to future patients, enabling them to anticipate and understand the consequences of their illness and its treatment. Patients themselves often express the wish for more emphasis upon research into QoL issues, and seek insight into the concomitants of their disease and its treatment.

Example from the literature

The Dartmouth COOP, a primary care research network, developed nine pictorial Charts to measure patient function and QoL (Nelson *et al.*, 1990). Each Chart has a five-point scale, is illustrated, and can be self- or office-staff administered. The Charts are used to measure patients' overall physical functioning,

emotional problems, daily activities, social activities, pain, overall health and QoL. The QoL item is presented as a ladder, and was later used in the QOLIE-89 instrument (Appendix E10, question 49).

Nelson *et al*. report results for over 2000 patients in four diverse clinical settings. Most clinicians and patients reported that the Charts were easy to use and provided a valuable tool. For nearly half of the patients in whom the Charts uncovered new information, changes in clinical management were initiated as a consequence.

It was concluded that the COOP Charts are practicable, reliable, valid, sensitive to the effects of disease and useful for measuring patient function quickly.

Patient preferences

Not only does a patient's self-assessment often differ substantially from the judgement of their doctor or other healthcare staff, but patients' preferences also seem to differ from those of other people. Many patients accept toxic chemotherapy for the prospect of minimal benefit in terms of probability of cure or prolongation of life, contrary to the expectations of medical staff. Therefore QoL should be measured from the patient's perspective, using a patient-completed questionnaire.

Example from the literature

Slevin *et al*. (1990) asked 106 consecutive patients with solid tumours to complete questionnaires about their willingness to receive chemotherapy. They were told that the more-intensive regimen was likely to have considerable side effects and drawbacks, such as severe nausea and vomiting, hair loss, frequent tiredness and weakness, frequent use of drips and needles, admission to hospital, decreased sexual interest and possible infertility. They were given different scenarios, such as (i) small (1%) chance of cure, (ii) no cure, but chance of prolonging life by three months and (iii) 1% chance of symptom relief only. All patients knew they were about to commence chemotherapy, and thus considered the questions seriously. Cancer nurses, general practitioners, radiotherapists, oncologists and sociodemographically matched controls were asked the same questions.

Table 1.2 shows the percentage of respondents that would accept chemotherapy under each scenario. There are major and consistent differences between the opinions of patients and the others, and also between the different healthcare staff.

Slevin *et al*. comment that patients appear to regard a minute chance of possible benefit as worthwhile, whatever the cost. They conclude: "It may be that the only people who can evaluate such life and death decisions are those faced with them."

Table 1.2 Percentage of respondents willing to accept intensive or mild chemotherapy with a minimum chance of effectiveness

	Controls	Cancer nurses	General practitioners	Radiotherapists	Oncologists	Cancer patients
Number	100	303	790	88	60	100
Cure (1%)						
Intensive regimen	19	13	12	4	20	53
Mild regimen	35	39	44	27	52	67
Prolonging life by 3 months						
Intensive regimen	10	6	3	0	10	42
Mild regimen	25	25	27	13	45	53
Relief of symptoms (1%)						
Intensive regimen	10	6	2	0	7	43
Mild regimen	19	26	21	2	11	59

Source: Adapted from Slevin *et al.*, 1990, Table II. Reproduced with permission of BMJ Publishing Group Limited.

These results have been closely replicated by others. For example, Lindley *et al.* (1998) examined QoL in 86 breast cancer patients, using the SF-36 and the Functional Living Index – Cancer (FLIC). They note that 'the majority of patients indicated a willingness to accept six months of chemotherapy for small to modest potential benefit'.

Late problems of psychosocial adaptation

Cured patients and long-term survivors may have continuing problems long after their treatment is completed. These late problems may be overlooked, and QoL reported in such patients often gives results that are contrary to expectations.

Example from the literature

Bjordal *et al.* (1994), in a study of long-term survivors from a trial of radiotherapy for head and neck cancer, unexpectedly found that the hypofractionated patients reported slightly better QoL than those who received conventional therapy. Hypofractionated patients had slightly better EORTC QLQ-C30 mean scores for role, social and emotional function and better overall QoL (Table 1.3), and reported less fatigue. However, both groups reported high levels of symptoms 7–11 years after their radiotherapy, such as dryness in the mouth and mucus production, and high levels of psychological distress (30% being clinical 'cases').

The authors conclude that clinicians need to be aware of these problems and that some patients would benefit from social support or medication. It was proposed that the GHQ-20 (General Health Questionnaire) could facilitate the screening for patients whose psychological distress might be treated.

Table 1.3 Quality of life in head and neck cancer patients 7–11 years after curative treatment

	Conventional radiotherapy (n = 103)	Hypofractionated radiotherapy (n = 101)	p-value
EORTC QLQ-C30 Function scales (mean scores)			
Physical function	74	79	NS
Role function	72	83	0.03
Social function	73	83	0.02
Emotional function	77	84	0.02
Cognitive function	80	83	NS
Overall QoL	61	69	0.04
EORTC QLQ-C30 Symptom scales (mean scores)			
Pain	19	15	NS
Fatigue	32	25	0.04
Emesis	6	5	NS
GHQ scores			
Mean score	20.8	19.7	NS
% cases	*31%*	*32%*	NS

Source: Bjordal *et al.,* 1994. Reproduced with permission of Elsevier.

Medical decision-making

QoL can be a predictor of treatment success, and several studies have found that factors such as overall QoL, physical well-being, mood and pain are of prognostic importance. For example, in cancer patients, pre-treatment assessment of QoL has been shown to be strongly predictive of survival, and a better predictor than performance status (Gotay *et al*., 2008; Quinten *et al.*, 2009). On this basis, it is possible to argue for the routine assessment of QoL in therapy trials.

It is not clear in these circumstances whether QoL scores reflect an early perception by the patient of disease progression or whether QoL status in some way influences the course of disease. If the former, the level of QoL is merely predictive of outcome. If it affects outcome, there could be potential to use improvement in QoL as an active form of therapy. Whatever the nature of the association, these findings underline the importance of evaluating QoL and using it when making medical decisions. Similar results have been observed in various disease areas. For example, Jenkins (1992) observed that preoperative QoL partially predicts the recovery process in

heart surgery patients. Changes in QoL scores during treatment have also been shown to have prognostic value.

Example from the literature

Coates *et al*. (1997) showed that patients' self-assessment of QoL is an important prognostic factor of survival in advanced cancer patients. Adult patients with advanced malignancy from 12 institutions in 10 countries completed the EORTC QLQ-C30 questionnaire. Baseline patient and disease characteristics were recorded.

Follow-up information was obtained on 656 patients, of whom 411 had died. In addition to age and performance status, the QLQ-C30 global QoL scale and the scales of physical, role, emotional, cognitive and social function were each predictive of subsequent survival duration. Table 1.4 shows the association of survival with the scores for overall physical condition (Q29) and overall quality of life (Q30). In this table, items Q29 and Q30 were each divided about their respective medians, and the hazard ratios show that patients with high scores were less likely to die than those below the median. For example, the hazard ratio of 0.89 for Q29 indicates that the rate of death in patients with high scores was only 89% of the death rate in those with low scores.

Coates *et al*. conclude that QoL scores carry prognostic information independent of other recorded factors.

Table 1.4 Prognostic significance for survival of two single-item QoL scores in patients with cancer, after allowing for performance status and age

QoL Variable	Hazard ratio	95% Confidence Interval	*p*-value
Physical condition (Q29)	0.89	0.82 to 0.96	0.003
Overall QoL (Q30)	0.87	0.80 to 0.94	0.001

Source: Coates *et al*., 1997. Reproduced with permission of Elsevier.

1.6 Which clinical trials should assess QoL?

It would be inappropriate to suggest that *all* RCTs, even in cancer, HIV or chronic diseases, should make a formal assessment of QoL. Clearly, there are situations where such information is not relevant. For example, when evaluating a potentially curative treatment that is not very likely to have adverse side effects, or if the treatments and side effects are similar in the various study arms, it might be unnecessary to make such assessment. However, many trial groups now insist that the investigators should at least consider the QoL implications and should positively justify *not* including the assessment of QoL.

When is QoL assessment a relevant endpoint? Gotay and Moore (1992) propose the following classification of trials for QoL purposes:

1. QoL may be the main endpoint. This is frequently true in palliative care, or when patients are seriously ill with incurable disease.
2. Treatments may be expected to be equivalent in efficacy, and a new treatment would be deemed preferable if it confers QoL benefits.
3. A new treatment may show a small benefit in cure rates or survival advantage, but this might be offset by QoL deterioration.
4. Treatments may differ considerably in their short-term efficacy, but if the overall failure rate is high then QoL issues should be considered.

Furthermore, despite the optimism of those who launch trials that seek a survival breakthrough, all too often completed trials show a limited survival advantage. Thus in these cases the relevance of QoL assessment has to be considered, since any gain in therapeutic efficacy would have to be weighed against possible negative effects pertaining to QoL.

1.7 How to measure quality of life

Ask the patient

Observers are poor judges of patients' opinions. Many studies have shown that independent assessments by either healthcare professionals or patients' relatives differ from the responses obtained when patients complete self-reported questionnaires. In some conditions observers appear to consistently overestimate QoL scores, in others, underestimate. There is general agreement that patients' opinions vary considerably from the expectations of both staff and relatives. It has been suggested that observers tend to underestimate the impact of psychological aspects and tend to emphasise the importance of the more obvious symptoms. The impacts of pain, nausea and vomiting have all been reported as being underestimated. Expected symptoms and toxicity tend to be accepted and hence ignored by clinical staff. Studies of nausea and vomiting in cancer patients receiving chemotherapy have found that doctors assume these symptoms are likely to occur and, as a consequence, often report only the more severe events. However, patients who are reported as having no problems may assert that they suffered quite a lot of vomiting (Fayers *et al*., 1991).

Observers frequently misjudge the absolute levels of both symptoms and general QoL. In addition, the patients' willingness to trade QoL for possible cure may be misjudged. Many patients are willing to accept unpleasant or toxic therapy for seemingly modest benefits in terms of cure, although a few patients will refuse treatment even when there is a high chance of substantial gain. Physicians and nurses are more likely to say that they would be unwilling to accept the therapy for such small potential benefit. When patients with cancer choose between two treatments, if they believe their disease is likely to be cured they may be willing to accept a treatment that adversely affects their QoL.

Observers, including health professionals, may tend to base their opinions of overall QoL upon physical signs such as symptoms and toxicity. However, in many disease areas, conventional clinical outcomes have been shown to be poorly correlated with patients' assessment of QoL. Thus, for example, in patients with asthma, Juniper *et al*. (1993) observed that correlations between clinical assessments and how patients felt and functioned in day-to-day activities were only modest.

Example from the literature

An early investigation conducted by Jachuk *et al*. (1982) into QoL concerned the effect of hypotensive drugs. Seventy-five patients with controlled hypertension each completed a questionnaire, as did a relative and doctor. A global, summary question was included, about whether there was overall improvement, no change or deterioration.

As Table 1.5 shows, while the physicians assessed all patients as having improved, approximately half the patients thought there was no change or deterioration, and all but one patient was assessed by their relatives as having deteriorated. Patients attributed their deterioration as due to decline in energy, general activity, sexual inactivity and irritability. Physicians, focusing upon control of hypertension, largely ignored these factors. Relatives commonly thought there was moderate or severe impairment of memory, energy and activity, and an increase in hypochondria, irritability and worry.

Nowadays, clinical practice places greater emphasis on patient–physician communication, and it is most unlikely that such extreme results would be observed if this study were to be repeated. The modern physician would be expected to have a far greater awareness of patients' feelings, leading to smaller differences.

Table 1.5 The results of overall assessments of QoL by 75 patients with controlled hypertension, their attending physicians, and the patients' relatives

	Improved	No change	Worse	Total
Physician	75	0	0	75
Patient	36	32	7	75
Relative	1	0	74	75

Source: Jachuk *et al.*, 1982. Reproduced with permission of the Royal College of General Practitioners.

1.8 Instruments

A large number of instruments have been developed for QoL assessment, and we provide a range of examples to illustrate some of the approaches used. Those reproduced (in the Appendix) have been chosen on the grounds of variety, to show particular features, and because these particular instruments are among the most widely used in clinical trials.

The aims and content of each instrument are described, together with an outline of the scoring procedures and any constructed multi-item scales. Most of the instruments use fairly simple forms of scoring, and the following basic procedure is usually used.

First, the successive levels of each categorical item are numbered increasingly. For example, a common scheme with four-category items is to grade responses such as 'not at all', 'a little', 'quite a bit' and 'very much' as being 0 to 3 respectively or, if preferred, 1 to 4, as it makes no difference after standardising the final scores. Second, when a scale contains multiple items, these are usually summed. Thus a four-item scale, with items scored 0 to 3, would yield a working score ranging from 0 to 12. Finally, the working score is usually standardised to a range of 0 to 100, and called the *scale score*. This enables different scales, possibly with different numbers of items and/or where the items have different numbers of categories, to be compared. In our example this would be achieved by multiplying by 100/12. We term this procedure the *standard scoring method*. A number of instruments are now advocating the use of *T*-scores, also called *norm-based scoring*, as described in Chapter 9.

Generic instruments

Some instruments are intended for general use, irrespective of the illness or condition of the patient. These *generic* questionnaires may often be applicable to healthy people, too. Some of the earliest ones were developed initially with population surveys in mind, although they were later applied in clinical trial settings.

There are many instruments that measure physical impairment, disability or handicap. Although commonly described as QoL scales, these instruments are better called *measures of health status* because they focus on physical symptoms. They emphasise the measurement of general health and make the implicit assumption that poorer health indicates poorer QoL. One weakness about this form of assessment is that different patients may react differently to similar levels of impairment. Many of the earlier questionnaires to some degree adopt this approach. We illustrate two of the more influential instruments, the Sickness Impact Profile (SIP) and the Nottingham Health Profile (NIP). Some scales specifically address activities of daily living, and we describe the Barthel questionnaire.

Few of the early instruments had scales that examine the subjective non-physical aspects of QoL, such as emotional, social and existential issues. Newer instruments, however, emphasise these subjective aspects strongly, and also commonly include one or more questions that explicitly enquire about overall QoL. We illustrate this approach by the SF-36. Later, brief instruments that place even less emphasis upon physical functioning have been developed. Two such instruments are the EQ-5D, that is intended to be suitable for use with cost–utility analysis, and the SEIQoL, which allows patients to choose those aspects of QoL that they consider most important to themselves.

Sickness Impact Profile (SIP)

The SIP of Bergner *et al*. (1981) is a measure of perceived health status, as measured by its impact upon behaviour. Appendix E1 shows an extract of SIP – the full questionnaire takes 16 pages. It was designed for assessing new treatments and for evaluating health levels in the population, and is applicable across a wide range of types and severities of illness. The SIP consists of 136 items, and takes about 20–30 minutes to complete. The items describe everyday activities, and the respondents have to mark those activities they can accomplish and those statements they agree with. It may be either interviewer- or self-administered. Twelve main areas of dysfunction are covered, but there is no global question about overall health or QoL. It has been shown that the SIP is sensitive even to minor changes in morbidity. However, in line with its original design objectives, it emphasises the impact of health upon activities and behaviour, including social functioning, rather than on feelings and perceptions – although there are some items relating to emotional well-being.

The items are negatively worded, representing dysfunction. Data from a number of field studies were compared against assessments made by healthcare professionals and students, leading to 'scale values'. These scale values are used as weights when summing the individual items to obtain the scale score. The standard scoring method is used for each of the 12 dysfunction scales. Two higher-order dimensions, summarising physical and psychosocial domains respectively, are recognised and these are scored in a similar manner.

Nottingham Health Profile

The NHP of Hunt *et al*. (1981) measures emotional, social and physical distress (Appendix E2). The NHP was influenced by the SIP, but asks about feelings and emotions directly rather than by changes in behaviour. Thus, although the authors did not develop or claim it to be a QoL instrument, it does emphasise subjective aspects of health assessment. It was based upon the perceptions and the issues that were mentioned when patients were interviewed. When it was developed, the idea of asking patients about their feelings was a novel concept.

The version 2 contains 38 items in six sections, covering sleep, pain, emotional reactions, social isolation, physical mobility and energy level. Each question takes a yes/no answer. As with the SIP, each item reflects departures from normal, and items are weighted to reflect their importance. Earlier versions included seven statements about areas of life that may be affected by health, with the respondent indicating whether there has been any impact in those areas. These statements were less applicable to the elderly, unemployed, disabled or those on low income than are the other items, and are usually omitted. The NHP forms a *profile* of six scores corresponding to the different sections of the questionnaire, and there is no single summary index.

The NHP is short compared to the SIP, and is easy to complete. The wording is simple and easily understood. It is often used in population studies of general health evaluation, and has been used in medical and non-medical settings. It is also frequently used in

clinical trials, although it was not designed for that purpose. However, it tends to emphasise severe disease states and is perhaps less sensitive to minor – yet important – changes and differences in health state. The NHP assesses whether there are any health problems, but is not sufficiently specific to identify particular problems. Some items do not apply to hospitalised patients, and the developers do not recommend it for these patients. Its simplicity, while being for many purposes an advantage, means that it does not provide suitable coverage for the conditions that apply to patients in many clinical trials.

Medical Outcomes Study 36-Item Short Form (SF-36)

The SF-36 developed by Ware *et al*. (1993) evaluates general health status, and was intended to fill a gap between the much more lengthy questionnaires and other relatively coarse single-item measures (Appendix E3). It is designed to provide assessments involving generic health concepts that are not specific to any age, disease or treatment group. Emphasis is placed upon physical, social and emotional functioning. The SF-36 has become the most widely used of the general health-status measures. It can be either self-assessed or administered by a trained interviewer.

As the name implies, there are 36 questions addressing eight health concepts (simpler 12- and eight-question forms are also available). There are two summary measures: physical health and mental health. Physical health is divided into scales for physical functioning (10 items), role-physical (four), bodily pain (two) and general health (five). Mental health comprises scales for vitality (four items), social functioning (two), role-emotional (three) and mental health (five). In addition, there is a general health transition question, which asks: 'Compared to one year ago, how would you rate your general health now?' There is also a global question about the respondent's perception of their health: 'In general, would you say your health is: (excellent, very good, good, fair, poor)?' Most questions refer to the past four weeks, although some relate to the present. A few questions, such as those for 'role-physical', take yes/no responses, while some, such as the physical functioning items, have three categories (limited a lot, limited a little, not limited at all), and other items have five or six categories for responses.

The designers of the SF-36 selected, standardised and tested the items so that they can be scored using the standard scoring method. More recently, norm-based scoring has been advocated (see Chapter 9).

Most of the items appear broadly sensible. However, the physical functioning scale, in common with many similar scales, poses questions about interpretation. Questions ask whether your health limits you in 'vigorous activities, such as running, lifting heavy objects, participating in strenuous sports' or in 'walking more than a mile'. It is not clear how those who never participate in such activities should respond – for example, suppose someone who never participates in sports has severely impaired health: if they respond 'No, not limited at all' they will receive a score indicating better functioning than might be expected. Some questionnaires therefore restrict physical functioning questions to activities that are expected to be applicable to everyone, while others stress that the questions are hypothetical ('we wish to know whether you *could* participate in sports if you wanted to').

For health economic evaluations, the SF-6D has been derived from the 12-item SF-12 subset of the SF-36 questionnaire (Brazier and Roberts, 2004). The SF-6D provides a preference-based single index measure for health, estimated from values of the SF-12.

EuroQol (EQ-5D)

The EQ-5D of Brooks *et al*. (1996) is another general-purpose instrument, this time emphasising both simplicity and the multi-country aspects (Appendix E4). It takes about two minutes to complete and aims to capture physical, mental and social functioning. It is intended to be applicable over a wide range of health interventions. The EuroQol group, acknowledging its simplicity, recommend using it alongside other instruments.

Most of the questionnaires we describe are *profile* instruments because they provide a descriptive profile of the patient's functional health and symptom experience. For medical decision-making, therapeutic benefits have to be contrasted against changes in QoL: if more efficacious therapy is associated with poorer QoL outcomes, is it worthwhile? Answering this involves weighing QoL against survival and combining them into a single summary score, most commonly by determining patient 'preference ratings' or 'utilities'. Overall benefits of treatment or management policies can then be contrasted. The EQ-5D, like the SF-6D which is based on the SF-36 and described above, is described as a *utility measure* or *preference measure*, in which the scores are weighted on the basis of preferences (or utilities) for discrete health states or combinations of health states derived from a reference sample.

Five dimensions of QoL are recognised: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. In the first version of the EQ-5D each of these was addressed by a simple three-level response scale; a revised version, the EQ-5D-5L, extended this to five levels per item as shown in Appendix E4 (Herdman *et al*., 2011). The principal EQ-5D question is represented by a 20 cm vertical VAS, scored from 0 to 100, on which the respondent should mark 'your own health state today', ranging from best imaginable health state to the worst imaginable health state. A single index is generated for all health states. Perhaps because of its extreme simplicity, the EQ-5D has been less frequently used as the outcome measure for clinical trials. It has been used most widely for general healthcare evaluation, including cost–utility evaluation. It is especially used in the UK, where the government-funded National Institute for Health and Care Excellence (NICE) declares: "Health effects should be expressed in QALYs. The EQ-5D is the preferred measure of health-related quality of life in adults" (NICE, 2013).

Schedule for Evaluation of Individual Quality of Life (SEIQoL) and the Patient Generated Index (PGI)

The SEIQoL (Hickey *et al*., 1996) and the PGI (Ruta *et al*., 1994) are examples of instruments that were developed to assess QoL from the individual's perspective. For both instruments, the respondents identify areas of life that are particularly important to them, and the current level of functioning in each of these domains is evaluated. The

practical procedure is as follows. First, the patient is invited to nominate the five most important aspects of their quality of life. Most patients readily list five domains, but if they find it difficult a standard list of prompts is used. Second, the patient is asked to score each nominated item or aspect, according to its severity. The third and final stage is to provide relative weights for the importance of each domain. Although the first stage is similar for both the PGI and SEIQoL, the two instruments differ in the way they implement the second and third stages. The PGI is somewhat simpler than the SEIQoL, and is described first.

For the second stage, the PGI (Appendix E5) invites the patient to score their chosen items using scales from 0, 'the worst you could imagine', to 10, 'exactly as you would like to be'. Then, for the third stage, the PGI asks patients to 'spend' a total of 10 imaginary points to improve areas of their life. At the second stage of the SEIQoL, the patient is offered a vertical 10 cm VAS for each of their chosen areas and asked to rate themselves on the scale between 'worst possible' and 'best possible'. Each SEIQoL scale generates a score between 0 and 100. For the third stage, obtaining importance weights, there are two approaches. The original SEIQoL then made use of a judgement analysis in which the patients grade a series of presented cases (Joyce *et al*., 2003). A simpler direct-weighting approach is adopted for the SEIQoL-DW (Browne *et al*., 1997). In this, patients are provided with a plastic disc that consists of five overlapping segments corresponding to the five domains that the patient has nominated; each segment can be rotated around the central pivot, allowing its exposed size to be adjusted relative to the other segments. The patient is asked: 'How do the five domains compare in importance to each other?' This procedure generates five weights that sum to 100%. For both instruments, the investigator calculates a score by multiplying the individual's self-rating in each of their chosen areas by the relative weight that they assigned to it, and summing the products over the five areas.

Both the SEIQoL and the PGI recognise that sometimes seemingly trivial problems may be of major significance to certain patients, while other issues that are thought by observers to be important may in fact be considered unimportant. Martin *et al*. (2007) review studies using PGI, and Wettergren *et al*. (2009) review use of SEIQoL-DW. Overall, patient-generated outcome measures are cumbersome to implement, make greater cognitive demands than traditional instruments. They appear to be useful primarily in complementing other measures and in guiding management decisions for individual patients, but may be less practical for clinical trial settings or when comparing groups of patients.

Disease-specific instruments

Generic instruments, intended to cover a wide range of conditions, have the advantage that scores from patients with various diseases may be compared against each other and against the general population. On the other hand, these instruments fail to focus on the issues of particular concern to patients with disease, and may often lack the sensitivity to detect differences that arise as a consequence of treatment policies which are compared in clinical trials. This has led to the development of disease-specific

questionnaires. We describe three contrasting questionnaires that are used in a single disease area – cancer – and very different questionnaires that are widely used in epilepsy and asthma.

It may be observed that, even in the three cancer-specific instruments, there is substantial variation in content and wording. Although all of these questionnaires assess similar content areas, the relative emphasis placed on any given QoL domain and the specific ways in which questions are posed vary considerably. For example, in comparison to the FACT-G, the RSCL and the EORTC QLQ-C30 include a relatively larger number of questions addressing physical symptoms. There are also semantic and stylistic differences: when assessing depression, the generic SF-36, the FACT-G and the QLQ-C30 use the following item phrasing, respectively: (i) "Have you felt so down in the dumps that nothing could cheer you up?" and "Have you felt downhearted and blue?"; (ii) "I feel sad"; and (iii) "Did you feel depressed?" These items differ in both the degree to which they rely on idiomatic expressions and in the directness with which the questions are posed.

European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30

The EORTC QLQ-C30 is a cancer-specific 30-item questionnaire (Aaronson *et al*., 1993); see Appendix E6. The QLQ-C30 questionnaire was designed to be multidimensional in structure, appropriate for self-administration and hence brief and easy to complete, applicable across a range of cultural settings and suitable for use in clinical trials of cancer therapy. It incorporates five functional scales (physical, role, cognitive, emotional and social), three symptom scales (fatigue, pain, and nausea and vomiting), a global health-status/QoL scale, and a number of single items assessing additional symptoms commonly reported by cancer patients (dyspnoea, loss of appetite, insomnia, constipation and diarrhoea) and the perceived financial impact of the disease.

In the QLQ-C30 version 3.0 all items have response categories with four levels, from 'not at all' to 'very much', except the two items for overall physical condition and overall QoL, which use seven-point items ranging from 'very poor' to 'excellent'. The standard scoring method is used. High scale scores represent high response levels, with high functional scale scores representing high/healthy levels of functioning, and high scores for symptom scales/items representing high levels of symptomatology/problems (Fayers *et al*., 2001).

The QLQ-C30 is available in a range of languages and has been widely used in multinational cancer clinical trials. It has been found to be sensitive to differences between patients, treatment effects and changes over time.

EORTC disease- or treatment-specific modules

The EORTC QLQ-C30 is an example of an instrument that is designed to be modular, with the core questionnaire evaluating those aspects of QoL which are likely to be

relevant to a wide range of cancer patients. For each cancer site particular issues are often important, such as specific disease-related symptoms or aspects of morbidity that are consequences of specific forms of therapy. The QLQ-ELD14, described by Wheelwright *et al.* (2013), is one of several modules that address additional issues (Appendix E7). This supplements the core QLQ-C30 with an additional 14 items for elderly patients with cancer.

Functional Assessment of Cancer Therapy – General (FACT-G)

The Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System is a collection of QoL questionnaires targeting chronic illnesses. The core questionnaire, or FACT-G, was developed by Cella *et al.* (1993) and is a widely used cancer-specific instrument (Appendix E8). Similar to the EORTC QLQ-C30, the FACIT questionnaires adopt a modular approach and so a number of supplementary modules specific to a tumour type, treatment or condition are available. Non-cancer-specific FACIT questionnaires are also available for other diseases, such as HIV infection and multiple sclerosis.

The FACT-G version 4 contains 27 items arranged in subscales covering four dimensions of QoL: physical well-being, social/family well-being, emotional well-being and functional well-being. Items are rated from 0 to 4. The items are labelled from 'not at all' to 'very much', which is the same as for the QLQ-C30 but with the addition of a central 'somewhat'. Some items are phrased negatively, and should be reverse-scored. Subscale scores are derived by summing item responses, and a total score is derived by summing the subscale scores. Version 3 included an additional item after each subscale, enabling patients to weight each domain on an 11-point scale from 'not at all' to 'very much so'. These questions were of the form: 'Looking at the above 7 questions, how much would you say your PHYSICAL WELL-BEING affects your quality of life?' A similar set of items is optional for version 4.

Individual questions are phrased in the first person ('I have a lack of energy'), as compared with the QLQ-C30 which asks questions in the second person ('Have you felt weak?'). Both questionnaires relate to the past week, both make similar claims regarding validity and sensitivity and both target similar patients. Yet the FACT-G and the QLQ-C30 are conceptually very different from each other, with the QLQ-C30 emphasising clinical symptoms and ability to function, in contrast to the FACT-G which addresses feelings and concerns (Luckett *et al.*, 2011).

Rotterdam Symptom Checklist (RSCL)

The RSCL (de Haes *et al.*, 1996) is another instrument that is intended for measuring the QoL of cancer patients (Appendix E9). In the past the RSCL was used extensively in European cancer clinical trials, although less so nowadays. It covers broadly similar ground to the EORTC QLQ-C30 and has a similar number of questions. As its name implies, greater emphasis is placed upon the symptoms and side effects that are commonly experienced by cancer patients.

There are two features that are worthy of special note. First, the RSCL has an introductory text explaining 'for all symptoms mentioned, indicate to what extent you have been bothered by it …' This is in contrast to the QLQ-C30 and most other QoL instruments, which merely inquire about the presence of symptoms. Thus one patient might have 'a little' stomach ache but, when asked if it bothers them, might respond 'not at all'; another might respond that the same ache bothers them 'quite a bit'. What is less clear is whether most patients read the questionnaire with sufficient care to appreciate the subtle significance of the instructions. The second feature relates to the ADL scale. Here, too, there are explicit instructions, stating: 'We do not want to know whether you actually do these, but only whether you are able to perform them presently.' Thus a patient might not 'go shopping' but is requested to indicate whether they could if they wanted to. This is in marked contrast with the equivalent scale on the SF-36 that not only asks about actual functioning but also includes some strenuous tasks which are perhaps less likely to be applicable to the chronically ill.

The RSCL consists of 30 questions on four-point scales ('not at all', 'a little', 'quite a bit', 'very much'), a question about activity level, and a global question about 'your quality of life during the past week' with seven categories. There are two main scales – physical symptom distress and psychological distress – in addition to the scales for activity level and overall valuation. The standard scoring method is used.

Quality of Life in Epilepsy (QOLIE-89)

In contrast with the previous examples, the QOLIE-89 is a 13-page, 89-item questionnaire aimed at patients with epilepsy (Devinsky *et al*., 1995); Appendix E10 shows extracts. It is based upon a number of other instruments, in particular the SF-36, with additional items from other sources. It contains five questions concerning worry about seizures, and questions about specific 'bothersome' epilepsy-related limitations such as driving restrictions. Shorter versions with 31 and 10 items are available. The QOLIE-89 contains 17 multi-item scales that tap into a number of health concepts, including overall QoL, emotional well-being, role limitations owing to emotional support, social support, social isolation, energy/fatigue, seizure worry, health discouragement, attention/concentration, language, memory, physical function and health perceptions. An overall score is derived by weighting and summing the scale scores. There are also four composite scores representing issues related to epilepsy, cognition, mental health and physical health.

The QOLIE-89 has been developed and tested upon adults. Epilepsy is a serious problem for younger patients too, but children and adolescents experience very different problems from adults. Adolescents may be particularly concerned about problems of forming relationships with friends of the opposite sex, and anxious about possibilities of marriage and their dependence upon parents. Children may feel excluded from school or other activities, and may be teased by other children. Very young children may be unable to complete the questionnaire alone, and parents or others will have to assist. Thus QoL questionnaires intended for adults are unlikely to be satisfactory for

younger age groups. One example of a generic QoL questionnaire that has been used for children with epilepsy is the 16-dimensional 16D, which Apajasalo *et al*. (1996) used in young adolescents aged 12–15, comparing normal children against patients with epilepsy. One interesting feature of the QOLIE-89 is that there are five questions about general QoL issues. Questions 1, 2, 3, 49 and 89 use various formats to enquire about health perceptions, overall QoL, overall health and change in health.

Paediatric Asthma Quality of Life Questionnaire (PAQLQ)

The PAQLQ developed by Juniper *et al*. (1996) has been designed to measure the problems that children between the ages of seven and 17 experience as a result of asthma; extracts from the self-administered version are shown in Appendix E11. The PAQLQ has 23 items relating to three dimensions, namely symptoms, activity limitations and emotional function. Items are scored from 1 to 7. Three of the activity questions are 'individualised', with the children identifying important activities at the beginning of the study. There is a global question, in which children are asked to think about all the activities they did in the past week, and to indicate how much they were bothered by their asthma during these activities. The items reflecting each dimension are averaged, forming three summary scales that take values between 1 and 7.

Parents often have a poor perception of their child's health-related QoL, and so it is important to ask the children themselves about their experiences. Since children may have difficulty in completing the self-administered questionnaire, Juniper *et al*. (1996) suggest using the interviewer-administered version, administered by a trained interviewer who has experience of working with children. Children may be strongly influenced by adults and by their surroundings, and so detailed guidelines and interviewing tips are provided. The PAQLQ has been tested in children aged between seven and 17 years, and has demonstrated good measurement properties in this age group.

Instruments for specific aspects of QoL

The instruments described above purport to measure general QoL, and include at least one general question about overall QoL or health. In many clinical trials this may be adequate for treatment comparison, but sometimes the investigators will wish to explore particular issues in greater depth. We describe four instruments that are widely used in clinical trials to explore anxiety and depression, physical functioning, pain and fatigue. These domains of QoL are particularly important to patients with chronic or advanced diseases. Many other instruments are available, both for these areas and others. Additional examples are *coping* (Hürny *et al*., 1993), *satisfaction* (Baker and Intagliata, 1982), *existential beliefs* (Salmon *et al*., 1996) and *self-esteem* (Rosenberg, 1965). Since these questionnaires evaluate specific aspects of QoL, in order for a patient assessment to be called 'quality of life' these instruments would normally be used in conjunction with more general questionnaires.

Hospital Anxiety and Depression Scale (HADS)

The HADS was developed by Zigmond and Snaith (1983) and was initially intended as a clinical screening tool to detect anxiety and depression (Appendix E12). It has become widely used in clinical trials for a wide range of conditions, including arthritis, bowel dysfunction, cancer, dental phobia, osteoporosis and stroke. The HADS consists of 14 questions that are completed by the patients. Each question uses a four-point scale. Seven of these questions were designed to address anxiety, and the other seven depression. The HADS deliberately excludes items that may be associated with emotional or physical impairment, such as dizziness and headaches; it emphasises the psychological signs or consequences of anxiety and depression.

Two particular features of the HADS are interesting from the point of view of scale design. The questions addressing anxiety and depression alternate (odd and even items, respectively), and half of the questions are worded positively and half negatively (e.g. 'I feel cheerful' and 'I get sudden feelings of panic').

Each item is scored 0 to 3, where 3 represents the state associated with the most anxiety or depression. The items are summed after suitable ordering, yielding two subscales ranging from 0 to 21. Based upon psychiatric diagnosis, HADS ratings of 11 or more are regarded as definite cases that would normally require therapy; ratings of 7 or less are non-cases; those scoring 8–10 are doubtful or borderline cases that are usually referred for further psychiatric assessment.

Another widely used instrument is the Beck Depression Inventory (BDI) (Beck *et al.*, 1961), which measures existence and severity of depression. It can be either self-rated or administered orally and emphasises cognitive rather than affective symptoms. The PHQ-9 is another popular depression questionnaire, and is briefer than the BDI (Spitzer *et al.*, 1999).

McGill Pain Questionnaire (MPQ)

Pain is a frequent symptom in many disease areas, and can be distressing. Not surprisingly, many instruments have been developed to assess pain. One such instrument, used extensively in clinical trials, is the Brief Pain Inventory (BPI) short form (Cleeland and Ryan, 1994). Further, many QoL instruments contain one or more items assessing pain. Examples include simple numerical rating scales, in which the respondents rate themselves in the range from 0 for no pain up to 10 for worst imaginable pain, and the more descriptive items as seen in the EORTC QLQ-C30 and the FACT-G.

The MPQ is one of the most widely used tests for the measurement of pain (Melzack, 1975). The MPQ full version has 20 main groups of items, each with between two and six adjectives as response categories, such as flickering, quivering, pulsing, throbbing, beating, pounding. It takes five to 15 minutes to complete. Based upon a literature search of terms used to describe pain, the MPQ uses a list of descriptive words that the subject ticks. The words are chosen from three classes of descriptors – sensory (such as temporal, spatial, pressure, thermal), affective (such as tension, fear) and evaluative (such as intensity, experience of pain). There is a six-point intensity scale for present

pain, from no pain through to excruciating pain. Three major measures are derived: a pain rating index using numerical scoring, the number of descriptive words chosen, and the value from the pain intensity scale. Pain-rating index scores can be calculated either across all items or for three major psychological dimensions, called sensory–discriminative, motivational–affective and cognitive–evaluative.

The short version, termed the SF-MPQ (Melzack, 1987), is shown in Appendix E13. It has 15 items that are graded by the respondent from none (0) through to severe (3). There is also a 10 cm VAS, ranging from 'no pain' through to 'worst possible pain', and the same six-point intensity scale as in the full version. It takes two to five minutes to complete. Each description carries a weight that corresponds to severity of pain. This leads to a summary score that ranges from 0 ('no pain') to 5 ('excruciating pain'). The SF-MPQ has subscales for affective and sensory components of pain, as well as a total score. In 2009 the SF-MPQ was revised to include an additional 7 items for neuropathic pain, and all 22 items are now rated from 0 to 10.

Pain is a complicated and controversial area for assessment, although some of the problems serve to illustrate general issues in QoL assessment. For example, the Zung (1983) self-rating Pain and Distress Scale measures physical and emotional distress caused by pain, rather than severity of pain itself. This recognises that severity of pain, either as indicated by pain stimuli or by the subject's verbal description, may result in different levels of distress in different patients. One level of pain stimulus may produce varying levels of suffering, as determined by reactions and emotions, in different patients. Also, pain thresholds can vary. Another issue to be considered when assessing pain is that analgesics can often control pain very effectively. Should one be making an allowance for increasing dosages of, say, opiates when evaluating levels of pain? In some studies it may be appropriate to measure 'uncontrolled pain', in which case one might argue that it is irrelevant to enquire about analgesics. On the other hand, high doses of analgesics can be accompanied by disadvantages.

Multidimensional Fatigue Inventory (MFI)

The MFI of Smets *et al*. (1995) is a 20-item self-report instrument designed to measure fatigue (Appendix E14). It covers five dimensions, each of four items: general fatigue, physical fatigue, mental fatigue, reduced motivation and reduced activity. There are equal numbers of positively and negatively worded statements, to counter possible response bias, and the respondent must indicate to what extent the particular statement applies to him or her. The five-point items take responses between 'yes, that is true' and 'no, that is not true', and are scored 1 to 5, where 5 corresponds to highest fatigue. The five scale scores are calculated by simple summation.

Four of the scales appear to be highly correlated, with mental fatigue behaving differently from the others. This suggests that there may be one or two underlying dimensions for fatigue. However, for descriptive purposes, and for a better understanding of what fatigue entails in different populations, the authors suggest that the separate dimensions be retained and that the five scales should not be combined. If a global score is required, the general fatigue scale should be used.

Many other fatigue questionnaires exist. Some are designed to be disease specific, targeting for example patients with arthritis or with cancer; some assume fatigue is unidimensional, while others use for example a three-dimensional model. Fatigue modules have been developed to complement the EORTC-Q30 and the FACT-G.

Barthel Index of Disability (BI)

Disability scales were among the earliest attempts to evaluate issues that may be regarded as related to QoL. They are still commonly employed, but mainly as a simple indication of one aspect of the patient's overall QoL. The BI (Mahoney and Barthel, 1965) was developed to measure disability, and is one of the most commonly used of the class of scales known as ADL scales. ADL scales focus upon a range of mobility, domestic and self-care tasks, and ignore issues such as pain, emotions and social functioning. The assumption is that a lower ADL score implies a lower QoL.

The BI is used to assess functional dependency before and after treatment, and to indicate the amount of nursing care that is required. It has been used widely for assessing rehabilitation outcome and stroke disability, and has been included in clinical trials. Unlike any of the other scales that we have described, it need not be completed by the patient personally but is more intended for administration by a nurse, physiotherapist or doctor concerned with the patient's care. It therefore provides an interesting contrast against the subjective self-assessment that has been adopted by many of the more recent measures. The BI examines the ability to perform normal or expected activities. Ten activities are assessed, each with two or three response categories, scored 5, 10 or 15; items are left blank and scored 0 when patients fail to meet the defined criteria. Overall scores range for 0 (highest dependency) to 100 (least dependency). It takes about one minute to assess a patient.

The original BI uses a crude scoring system, since changes in points do not appear to correspond to equivalent changes in all the scales. Also, patients can be at the highest (0) point on the scale and still become more dependent, and can similarly exceed the lowest (100) value. Modified versions of the BI largely overcome these deficiencies. For example, Shah *et al*. (1989) expanded the number of categories and propose changes to the scoring procedures (Appendix El5). The BI and its modified versions continue to be used widely as a simple method of assessing the effectiveness of rehabilitation outcome.

Many ADL scales exist, and the Katz *et al*. (1963) index is another widely used example. The concept here is that loss of skills occurs in a particular sequence, with complex functions suffering before others. Therefore six items were chosen so as to represent a hierarchical ordering of difficulty. A simplified scoring system is provided, in which the number of activities that require assistance are summed to provide a single score. Thus while '0' indicates that no help is required, '6' means that there is dependency for all the listed activities. The Katz index has been used with both children and adults, and for a wide range of conditions. In contrast to the BI, which measures ability, the Katz index measures independence.

Instrumental activities of daily living (IADL) scales include items that reflect ability to live and adapt to the environment. This includes activities such as shopping and travelling, and thus these scales evaluate one's ability to live independently within the community, as opposed to needing help with basic functions such as dressing and washing oneself. One example is the Functional Activity Questionnaire (FAQ), which was designed for use in community studies of normal ageing and mild senile dementia (Pfeffer *et al*., 1982).

1.9 Computer-adaptive instruments

Instruments developed in the twentieth century were mainly fixed format and paper based. Most are intended for self-completion by patients or other respondents, although some are designed with other modes of administration in mind such as by interviewer or over telephone. More recently, advances in computer technology have led to a new generation of instruments designed specifically for computer administration. These instruments are typically 'adaptive', in the sense that there is a large pool of potential items and each respondent is presented with a different set of items that are dynamically selected according to the respondent's previous answers. To some extent this mirrors the usual dialogue between a physician and patient: if a patient has said that they have no trouble walking long distances, why ask if they are housebound? It is more informative and efficient to tailor the interview as it progresses. However, computer-adaptive instruments use statistical algorithms to identify dynamically the optimal choice of items. They also calculate scores that are calibrated using a consistent metric across patients, enabling comparisons of PROs both between individual patients and in groups of patients.

Many of the instruments described above are available in their full original paper-based form, as short-form versions and, more recently, teams such as the EORTC group are generating in computer-adaptive versions (Petersen *et al*., 2010). Other instruments are designed mainly for computer adaptive use; a prominent example is PROMIS (Cella *et al*., 2010; Reeve *et al*., 2007), which is developing a wide range of instruments that measure concepts such as pain, fatigue, physical function, depression, anxiety and social function (www.nihpromis.org); PROMIS instruments are available as computer adaptive tests that require three to seven items for precise scores, or four- to 10-item short form versions. Computer-adaptive instruments are described in Chapter 8.

1.10 Conclusions

Definitions of QoL are controversial. Different instruments use different definitions, and frequently no specific model for QoL is stated formally. There is a wide range of QoL instruments available, although this range is likely to be reduced once the purpose

of evaluating QoL is considered. In a clinical trial setting, the disease area and therapies being evaluated will usually limit the choice. Common features of the instruments are that the patients themselves are asked, there are frequently several subscales, the scales are often based upon multiple items, and the scales represent constructs that cannot be measured directly. In Chapters 3–7 we shall explain methods for constructing such scales. Most importantly, we describe the desirable measurement properties of scales. We show how to 'validate' scales, and how to confirm whether an instrument appears to be consistent with the hypothetical model that the designers intended.