Drug Information Journal, Vol. 28, pp. 123–140, 1994
Printed in the USA.

0092-8615/94

QUALITY OF LIFE: OVERVIEW AND PERSPECTIVES

ROBERT LAUNOIS
Université de Paris XIII, Paris, France

The debate between public bodies and the pharmaceutical industry is restricted by fundamental differences in the arguments put forward. The former highlight differences between the increasing curve in medical expenditure and progress judged by increased life expectancy, whereas the latter stress that the aims of contemporary medicine should now be to limit the results of disease and improve quality of life. The failure of health care to demonstrate beneficial effects originates from the fact that the measurement parameters used are inappropriate; new ones are required. In order to make judgments on subjective health and disease outcome, nonphysiological parameters must be used. The first part of this discussion will describe the concepts involved, the second will examine mechanisms currently available. Finally, the qualities such instruments must possess shall be assessed.

Key Words: Quality of life; Domains; Indicator

CONCEPTS

THE FIRST STAGE OF any study examining the quality of life is to define the *universe* of the area to be studied. Once defined, the universe must be categorized to define *specifics domains* to be quantified (Figure 1). In order to assess these domains, a number of *criteria* or indicators must be available in order to quantify them and appropriate *scaling procedures* must be selected. Finally, development of a definitive indicator system must consider *objectives* for which it has been designed, without which results are meaningless.

Quality of life is such a unifying concept that ultimately all facets of the being may be included: environmental factors, behavior, and lifestyle. This discussion shall be restricted to those factors which influence patients' quality of life as a result of disease or its treatment. Life may not, however, be assessed generally: at best, different aspects of life may be assessed. This has two implications:

1. It forces one to break down the overall being into its constituent parts, an approach which may not be a bad thing given the abstract nature of the concept, and
2. It forces one to define from the start domains which will be explored.

Categorization of health is a difficult stage. In a number of cases it will be performed mathematically by deriving a vector from correlation between these indicators. For convenience, health shall be defined *prospectively* by using the World Health Organization (WHO) definition which is most frequently cited: "health is not only the absence of disease or of disability, but an overall state of physical, mental, and social well-being."

Reprint address: Robert Launois, Université de Paris, Paris, France.

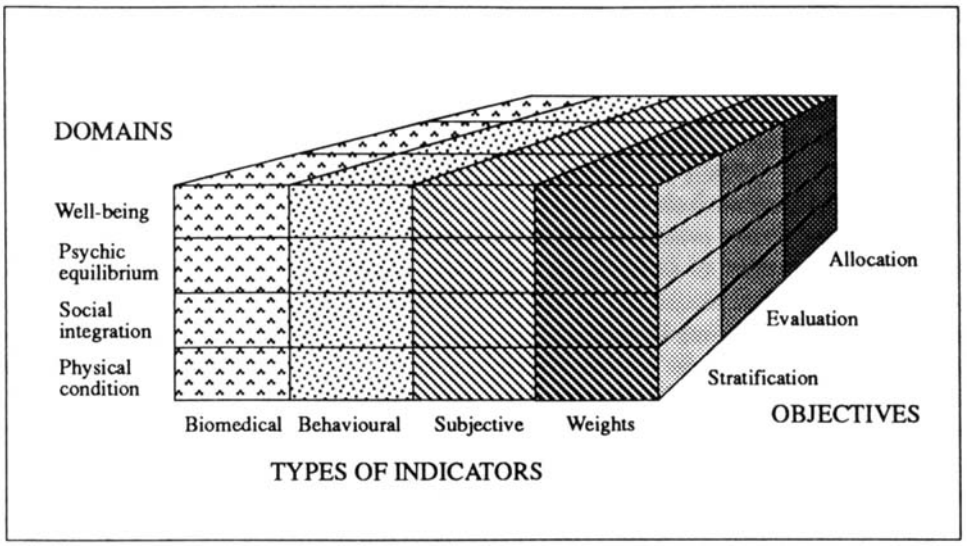

FIGURE 1. Definition of the universe and constituent domains.

The concept of well-being which combines the above will be considered in part as a separate domain. A good quality of life may, therefore, be represented by a feeling of well-being, emotional stability, appropriate social integration, and good physical state.

CHOICE OF INDICATORS

Up to this point these four domains are only concepts, that is, abstract principles. Measurement of these domains must be performed using solid recordable parameters. For each domain, a number of points has to be defined which will act as intermediaries between the abstract characteristics to be inferred and either objective or subjective measures.

Choices depend on the approach chosen to assess health problems. For some authors, the definition of health may be restricted to the absence of clinical symptoms or biological abnormalities. Other workers distinguish between those diseases which may be defined by the profession, and sickness expressed in terms of behavior. A number of definitions stress the patient's perception of illness, that is, they are based primarily on a patient's individual satisfaction or lack of satisfaction with his well-being.

Different methods for collecting information apply to these different approaches. It is important that these all be addressed simultaneously if each of the domains of quality of life are to be examined from all angles: three types of indicators may be used: bio medical, behavioral, and perceptual.

The intensity of symptoms, degree of incapacity, or level of dissatisfaction depend both on the absolute severity of the phenomenon and on the degree to which it interferes with daily life. The relative weight given to illness used to assess a domain should, in principal, be assessed by the patient himself. Frequently, however, this is defined by external observers, or graded by reference to mean population behavior. The problem may occasionally be completely masked by the use of an equal weighting system (such as, for example, the Apgar score [1]). It is important to assess the interdependence of quality of life domains. For this reason a fourth column named "weighting," which may be used to quantify the relative importance

of the different criteria, has been added. The exact nature of the weighting—equal weighting, individual scale, or external reference—will depend on the assessment instrument used.

DEFINITION OF PROCEDURES FOR SCALE CONSTRUCTION

Measurement of physiological or functional parameters is straightforward when it is derived from physical indicators. This does not apply when measurements are influenced by the emotional state of the patient. When an attempt is made to infer a characteristic by means of measurements performed using perceptual indicators, the measurement instrument used and conditions of use must be carefully defined, so that the procedure may be repeated and results verified.

The scaling procedures (2,3,4), that is, the conventions which govern allocation of values for different indicator positions, are a primary feature of standardization required for the measurement instrument. They package empirical interpretation into a unit which may be used and dictate the method of statistical assessment of findings.

When numerical symbols are not accompanied by measurement units they adopt a purely descriptive role. The number may, therefore, be used as part of a simple identification procedure. Here it corresponds more to categorization rather than to quantified measurement. Figures which are divided into such categories may be transferred and reallocated without adverse effects. In standardized discharge summaries, for example, where the figure one defines active patients and the figure two inactive patients, no information is lost where one defines inactivity and two activity.

The figure may also be used to rank findings. It may be used to indicate the relative position of the indicator on a continuum scale of the feature being assessed, implying that this indicator has the same basic characteristics at different levels: "I love a little, a lot." If, however, the scale is not standardized it is not possible to compare distances between points. It is impossible, therefore, to define the distance between two gradations on the scale even if the figures concerned are equally spaced. It provides a rank order, but the distance between two levels on the scale and the relationship between these levels may not be assessed. Most scales used to assess clinical quality of life are of this type, and it is, therefore, impossible to use them to assess change. Comparative scales in specific categories must be defined: "I am better, a little better, nothing has changed."

In order to be more than descriptive, numerical symbols must be accompanied by measurement units. To be a true measurement of size, the figure must be related to a standard unit: the figure two on its own has no meaning, two meters provides information.

Once the scale has been established using a single constant measurement unit throughout its length, points along the scale may be compared between, for example, a control group and a treatment group, even if ratios are impossible to assess in the absence of a zero standard. It is often difficult or even impossible to demonstrate beyond doubt the complete disappearance of a parameter used to assess quality of life. Even if conditions of life worse than death exist, it is difficult to imagine several quality of life domains with zero scores.

Where, for the modalities comprising an indicator, it is possible to define both a natural zero, the origin, and the distance between two points on the scale, the figure then becomes arithmetic. The distance between two gradations and their ratio may then be calculated. This is a fundamental property of a metric scale (still called a ratio of proportionality scale) used to confirm, for example, that one state of health is twice more severe than another, something which could not previously have been done.

The best characteristic of a scale is its invariance, that is, the degree to which it can be manipulated without distorting its structure. In the *ordinal* scale all transformations which preserve the order in the scale do not change available information. Such a scale is said to be preserved following *monotonic* transformation. In an *interval* scale all numbers on the scale may be multiplied by a constant factor, or the origin shifted by a constant number without changing results. Such a scale is said to be preserved by *affine* transformation ($y = mx + c$). Finally, in a *ratio* scale the relationship between values is not changed if they are multiplied by the same constant factor. It is preserved by *linear* transformation ($y = mx$). The more precise the information contained in the scale the more restricted the ability to modify the scale without changing the information contained therein becomes.

SPECIFICITY OF TOOLS

If scales are to be used as measurement instruments, they must be applied appropriately if they are to produce reliable results; in other words, they must measure what they were designed to assess. The tool used to identify a problem is not necessarily that which allows progression to be followed, and the tool used to follow progress may not be useful to assess allocation of resources. The choice of a method requires initial definition of users' needs: identification of a problem, assessment of change in response to treatment, or greater and more coherent use of scarce resources.

Discriminative and Evaluative Indicators

Certain groups of indicators measure different levels of quality of life. These are *discriminative indicators*.

Changes in level which reflect the differential effects of disease and treatment measure developments; these are defined as *evaluative indicators*. As Kischner and Guyatt have shown, discriminative indicators may not necessarily be used as evaluative indicators (5).

Discriminative indicators are used to categorize a population into subgroups as a function of specific individual characteristics comprising an individual at a given point in time where no reference criteria exist to distinguish these individuals. Indicators must be chosen which are common to all: patients and healthy people; the number of grades may be limited ultimately to two categories where the feature may be present or absent. The reliability of the instrument may be confirmed by verifying that interindividual differences do not change over time. Any consistent change which parallels the score may not, therefore, be identified. In any event, the absolute score is of no importance as it is used purely to classify subjects into specific categories.

Evaluative indicators have a completely different use. They are used to measure quantitative changes in quality of life. Items are chosen as a function of their ability to demonstrate change.

Multiple response options exist and the stability of the instrument may be measured by assessing the repeatability of intraindividual change with time when treatment is not changed. Assessment of the absolute quality of life and of changes in quality of life, therefore, require different instruments. The use of a discriminative indicator in a randomized study is, therefore, doomed to failure in advance as this type of instrument should not be used to assess the effects of treatment.

Structure of Quality, Overall Quality

An economist needs overall results (6). He requires a common measurement to compare the effects of decisions high up the health care system: techniques and equipment available to a statistically average

population. The medical approach is different. A clinician's aim is to apply techniques and equipment available to him, to draw from them the maximum possible benefit.

All features of the disease must be approached, explaining why assessment has remained a multidimensional problem. These two different attitudes have produced two different approaches in the development of questionnaires. The quality of life may be assessed in two ways: by composition, by constructing the general from the specific; or globally, by first examining the whole system, automatically combining but not categorizing partial observations.

In the compositional approach a number of partial indicators may be combined either simply or by *ad hoc* weighting of selected variables. These partial indicators may or may not be combined to produce an overall score. When they remain as distinct entities in the final assessment mechanism they produce profiles. If they are combined into a single figure, the term index or combined indicator is used. In all cases, the method chosen will reflect the complexity of the situation. This is a method which has been used since the 19th century by psychometrists endeavouring to impose "the discipline of measurement and figures to aspects of the spirit." This concept was also put forward by Alvan Fenstein and the psychopathologists (7) in recommending grading of clinical judgments. A doctor in the privacy of his consulting room has no need of a questionnaire to assess his patients' quality of life.

An ear to the patients' complaints is sufficient to identify and to remedy them. This is completely different when treatments are being assessed on a group of patients. Standardized reliable measurement instruments must first be available. The medical approach is very pragmatic, it relies not on assessing all aspects of quality of life, but on specific examination of those areas which relate to the disease and its treatment. The area may be more or less exhaustively assessed dependent on circumstances and the proposed treatment.

The method used by economists and decisional analysis proponents (8,9) is completely different. It is an overall approach based on the supposition that quality of life exists as a continuum from good health to death. Most simplistically, this concept may be thought of in terms of the definition of health used by WHO, graded from well-being and love of life to death, passing successively through the presence of signs and symptoms, physical disability, reduced mental capacity, and social withdrawal.

This heuristic approach produces an unidimensional ladder but distorts reality as it only grades isolated abnormalities. Symptoms of dysfunction present most frequently in combination. It is, therefore, the overall pattern of changes in quality of life which are graded on an interval scale by use of specific scenarios or by using means of classifying states of health. An assessment of the overall quality of life involves, therefore, determining values of coefficients between zero and one attributable to each scenario. These weighting factors adjust the quantity of life as a function of quality; and are, therefore, called "quality of life coefficients" (QOLC).

The product of the number of years or fractions of a year spent in a given state of health with the corresponding quality of life coefficient converts the time spent in poor health into equivalent fractions of years of good health (10). If this same procedure is performed at different stages during disease progression a number of years is obtained, corrected for the quality of life years (QALY) (11). The cost of treatment may, therefore, be divided by the QALY result to produce a parameter on which the relative merits of treatment or nontreatment with two alternative therapies may be compared.

Numerator and denominator must, of course, be related to time as two identical effects on health or two identical units of expenditure will not have the same value

when they occur at different times. The reasons for this are simple: immediate action is always preferable to individuals than a delay. Resources which are not consumed immediately may always be invested elsewhere. Costs and benefits in the future will always, therefore, attract a lesser weighting than those which occur immediately.

Momentary Analysis or Follow-up

The traditional approach to the measurement of quality of life made no reference to changes with time as it did not balance quality against quantity. This is only reasonable if treatment options and outcomes are completely similar, in three areas:

1. The associated risk of death,
2. The total length of life, and finally
3. The time spent in the different stages of progression of disease throughout the observation period.

This hypothesis assumes that two therapeutic maneuvres produce their effects over the same time period (t_1), that this effect is absolutely stable hereafter (t_2), and the assumption that progression to ultimate death occurs consistently over a given period (t_3). These hypotheses appear too restrictive. The differential assessment of therapeutic options measured in terms of their utility allows in contrast an assessment of their long-term effects. If one goes beyond the realms of clinical decision making to address the question of resource allocation, it is by definition imperative to have a score which may be used for comparison over time.

Two different situations must be distinguished; treatments may be instituted simultaneously and independently in different medical fields, or alternatively treatments in one specific indication may be mutually exclusive. In the first situation the decision algorithm involves the construction of a hierarchy of possibilities as a function of their mean cost-effectiveness within the limits of budgetary constraints; classically, a list. With the publication of the list inconsistencies in choices may become obvious (11,12). The higher the unit cost of success, the less justified the corresponding investment appears. Development of simpler techniques will, for a given budget, produce better overall results in terms of public health.

In the second situation, when treatments are incompatible, comparison is reduced to the assessment of two successive procedures (13,14). The first requires selection of a group of effective strategies based on the dominance principal from all possible therapeutic strategies. In the second stage, the society chooses from the effective strategies the one which appears to be best and fixes the resources it will make available to obtain what it deems to be optimal cost-effectiveness.

INSTRUMENTS IN USE

Assessment of the quality of life must fulfill the needs of those who use it. For the doctor this is a means to rise above too biological an approach, which is unquestionably useful in severe situations but assumes only a secondary role once the life-threatening event is past. Beyond organic disease, body spirit must be examined, but subjective judgments have limitations.

The patient produces a detailed quantified description of his problems but does not prioritize them. What is important is that he should be able to explain his various complaints weighted according to their effect on his quality of life. The collective process seeks to reflect the priorities of the society.

These three approaches which originate from different concepts – clinimetry, decision theory, and health indicators – were the basis of the development of current instruments. The basis of the differences between them is the introduction and the nature of the weighting scheme adopted.

THE EYE OF THE OBSERVER

Functional Incapacity Scales

Assessment of the dependence of elderly subjects has led over recent years to the development of many scales allowing assessment of individual performance using a number of essential functions. These rely almost totally on measurements performed by those caring for the individual.

The Katz scale (15) produces a global scale based on six activities: bathing, dressing, toilet, mobility, incontinence, and feeding. Each parameter is assessed on graded scores of up to three.

The Harris score (16) examines the ability of elderly subjects to perform acts of daily living: eating, buttoning clothes, moving, going to bed, bathing, washing, dressing, tying shoelaces, and combing hair.

Two different types of activity are distinguished: primary and secondary activities. Five levels of severity are defined as a function of the incapacity or help required:

1. Alone and without difficulty,
2. Alone and with difficulty,
3. Requiring help,
4. Impossible to perform even with help, and
5. Tasks performed with difficulty but unknown as to whether help is present or not.

Each feature is graded by severity from zero to six for primary activities and from zero to three for secondary activities. Scores are added to give an overall indicator. Multiple incapacities are only scored as highly as the sum of their handicapping effects, whereas the simultaneous failure of several systems is always more handicapping than the sum of the individual effects.

The New York Heart Association (NYHA) classification (17) proposed by the association of New York cardiologists grades cardiac and vascular diseases as a function of the severity of symptoms and the performance of tasks. Four grades are distinguished:

1. Absence of symptoms during normal activity,
2. Mild symptoms during normal activity,
3. Tiredness, dyspnoea, palpitations, and angina developing on effort less than required for normal activity, and
4. Symptoms at rest.

This scale is very widely used in clinical practice and in randomized studies. It is an ordinal scale which leaves a sizeable part of its interpretation to the opinion of the doctor. A number of authors have demonstrated its limited reproducibility, and its validity has also been questioned, as correlation with functional capacity is poor.

Goldman demonstrated that it was frequently highly subjective. The NYHA score improves simply because the patient gives up activities which he finds tiring. The same author proposed a new instrument to grade these problems: the specific activity scale (SAS) (17); objective signs are stressed at the expense of symptoms. Functional capacity of patients relating to certain activities representative of their daily life are graded in metabolic equivalence or "mets."

One of the functional indices most frequently used is that of Karnofsky (18). This addresses three questions: Has the patient been able to continue work? Can he carry out his normal functions? and Can he perform basic activities of daily living? The response profile defines three performance grades spread out over 11 levels from normal activity, 10, to death, zero. The functional states described are neither exhaustive not exclusive and there are exceptions and situations which are impossible to classify. Its long history and extensive use in medical circles explain why this scale is still used despite demonstrable failings.

All of the indicators used to assess restricted activity in terms of fundamental

acts of daily living assess levels of autonomy which are too great or handicaps that are too infrequent to be of use in assessing the entire population (Figure 2). Stewart reports that 80% of a noninstitutionalized population are devoid of any specific functional failings although Kaplan and Bush reported that 50% of subjects questioned in the San Diego study reported minor problems which affected their quality of life without significantly restricting their autonomy or mobility. To assess the adverse effects of a disease or treatment a much broader concept of quality of life must be used, integrating both psychological and social domains and using opinions of patients themselves, and not those of the doctors caring for them.

Preisman and Baum (23) used such a method to assess the effects of breast cancer therapy. This was the first attempt to use visual analogue scales in oncology to produce autovaluation of the quality of life by the patient himself (LASA-P).

The patient was asked to place a mark on a horizontal or vertical line between two extremes corresponding to the absence or maximal intensity of a given indicator. According to A. Moles: "The subject faced with such options feels obliged to find a solution between the two extremes. In order to reply to the question he must approach the 'physionomy of the phenomenon' through which he answers the question" (24).

This technique has been applied to specific aspects of morbidity: humor, energy, pain, nausea, appetite, ability to perform domestic tasks, social life, anxiety, and relief provided by treatment. Each response is graded out of 10 with an overall maximum score of 100. This technique was used in a comparative study examining hormonal and cytotoxic therapy and demonstrated that although secondary effects were more severe with cytotoxic than with hormonal treatment the quality of life was better in the former case due to a greater reduction in tumor load.

The functional living index in cancer (FLIC) (24) uses the same objectives: it assesses progression in patients suffering from malignancy using other than the traditional functional approach. Questions were developed on a semistructured approach using a panel of experts and including patients and their spouses, doctors, nurses, and a priest.

This panel established a list of 250 questions which after elimination of redundant

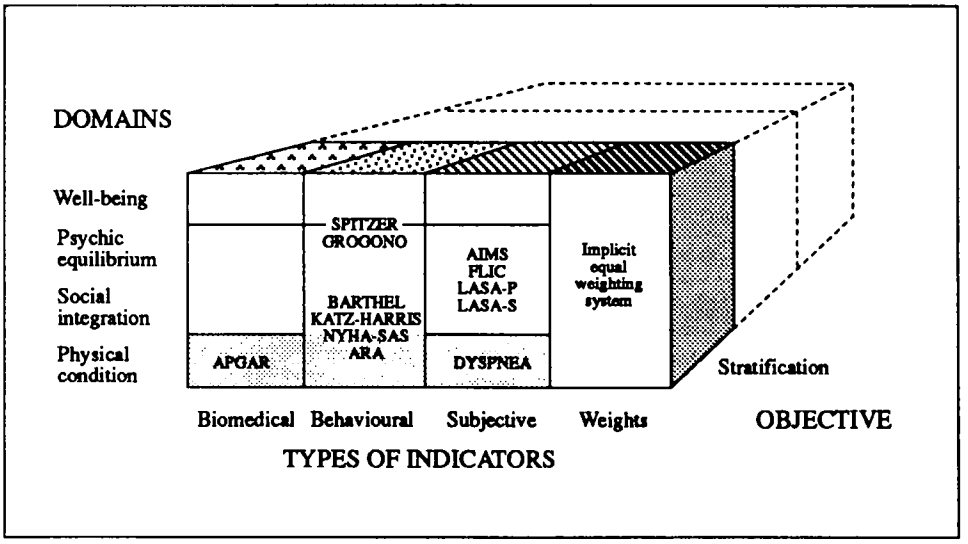

FIGURE 2. Eye of the observer (References 15–31).

or poorly designed questions led to an initial questionnaire containing 92 points which was tested on 175 patients. Multifactorial analysis was used to identify principal features, and to eliminate 52 further superfluous questions. A second questionnaire contained 40 topics and was tested on 312 patients. Following this, a second analysis was performed to confirm design stability. This led to the development of a third questionnaire containing 20 points which was assessed on 175 patients. Finally, two additional questions were added: the current version, therefore, contains 22 equally weighted questions. The scaling procedure uses both analogue scales and specific categories. A scale divided into a number of levels, from four to seven depending on the question, is used for each question. Each subject must place a vertical mark corresponding to the position which best describes his current situation. The closest value on the visual scale is scored for the trait and a global score constructed from the sum of partial scores from the different domains examined. It is an ordinal scale.

In rheumatology, English literature has recently described a whole collection of statistical instruments used to measure functional and psychological effects of rheumatoid arthritis: Health Service Questionnaire (HAQ) (27) and Arthritis Impact Measurement Scale (AIMS) (28). The AIMS scale uses 55 questions in nine areas: mobility, autonomous function, mobility, manual dexterity, domestic activity, personal hygiene, social relationships, anxiety, depression, and pain. The effects of disease were assessed in three independent areas: physical incapacity, distress, and pain. Global scores are not calculated.

Batteries of Indicators

This method was used particularly in a double-blind study (31) to assess the effects of three anti-hypertensive treatment regimes. Investigators chose five areas of quality of life: physical state, emotional state, intellectual ability, social integration, and general feeling of well-being. These domains were explored using independent multidimensional indicators. Such an approach, although exhaustive, posed a number of problems. In order to be of use, valid and practical indicators to measure each domain had to be identified. Secondly, interpretation of results may be difficult, particularly in the absence of primary endpoints when performances in different domains do not necessarily progress in the same direction. The use of batteries of multiple indicators is laden with potential problems.

THE PATIENT'S EXPERIENCE

According to Goldberg, patient preferences may be expressed either as the effects of weighting of results of partial measurement, with or without subdomain aggregation, or overall global score.

Subjective Profiles of Quality of Life

Proponents of the first school of thought defend a dissected approach to quality of life. They advocate initial identification of relevant domains based on information reported in the literature and interviews with experts and patients. Signs and symptoms gathered may be combined to assess the impact of disease on the daily life of the patient. In order to quantify responses each item has to be converted to a score. An initial questionnaire must, therefore, be designed for two purposes: to scale the indicators and to select the most relevant. Given that the items pool is designed to provide the basis for construction of the final questionnaire it is important to list many more parameters than will be used in the definitive version. For each parameter two types of questions are used, the first assessing the presence and intensity of the problems and the second the importance attributed to it by the patient as a measure

of the quality of life. This method of analysis may be used to select the relevant parameters. They consist of selecting items with the highest product between frequency and importance. The other eliminates parameters by principal component analysis; identification of redundant parameters and regrouping parameters according to their contribution (loading) to different factors.

The first approach is the most appropriate when the aim of the assessment is to "know the basis of the subject's appreciation of their quality of life" (32). Guyatt used the distinction proposed by Gerin between "central values" as a function of which patients orientate their lives and "peripheral values." Only parameters reflecting the central values were used in the final questionnaire, the others being eliminated. The assessment instrument by its nature implicitly integrates patient preferences as these were the basis of the choice of areas and items selected in each subdomain assessed.

Scale of Personal Well-being

Torrance (36) proposed that patients should be encouraged to express their feelings in terms of a range of states of health, combining different domains of quality of life. The methods used to record individuals' preferences are highly varied (37–39): "standard gamble," "time trade-off," and "category rating." The first of these methods was traditionally used to assess key preferences in situations of uncertainty. Because of this it is considered to be particularly appropriate in medical fields (Figure 3).

The protocol on which it is based is simple: three states of health (S1, S2, and S3) are carefully detailed and shown to a subject who must choose between the following options: Either treatment *A* which guarantees situation S2, treatment *B* which may have two possible outcomes: state S1 of probability p or state S3 of probability $1 - p$. States S1, S2, and S3 are arranged in a hierarchy with S2 occupying a position between S1 and S3. When the value of p is varied from zero to one this produces a threshold value where the patient is unable to decide between the two options. This value may be used to assess the utility of the first of these therapeutic possibilities.

The dilemma faced by patients suffering from coronary artery disease highlights the use of such a system. Mr. X suffering from angina may be offered two possibilities: either long-term therapy or the risks of a bypass operation. The outcome of the first choice in the short term is without doubt: he will live. The second choice is more risky as the changes of surgical success have been estimated by his general practitioner to be 90% in this case. The patient is caught between two possible courses of action. He may either choose the high risk situation which includes a not insignificant risk of failure or adopt the secure option but, by definition give up any possibility of improving his functional state. A problem then arises in that if the patient opts for the secure course, he will be better off than if the worst outcome of the high risk approach were to occur, that is, death, but worse off than if the operation succeeds. In order to decide he must assess the relative desirability of remaining in his present state with angina compared to the best and worst possible outcomes following the higher risk option.

The dilemma may be solved using a standard gamble based on population statistics. The structure of the gamble is identical to that of the initial problem. Choice is limited to a certain outcome and a risk outcome; survival without sequelae or death. Two differences exist, however, by comparison with the initial dilemma:

1. The decision rests on a hypothetical situation removing emotional overlay which played a part in the initial problem, and

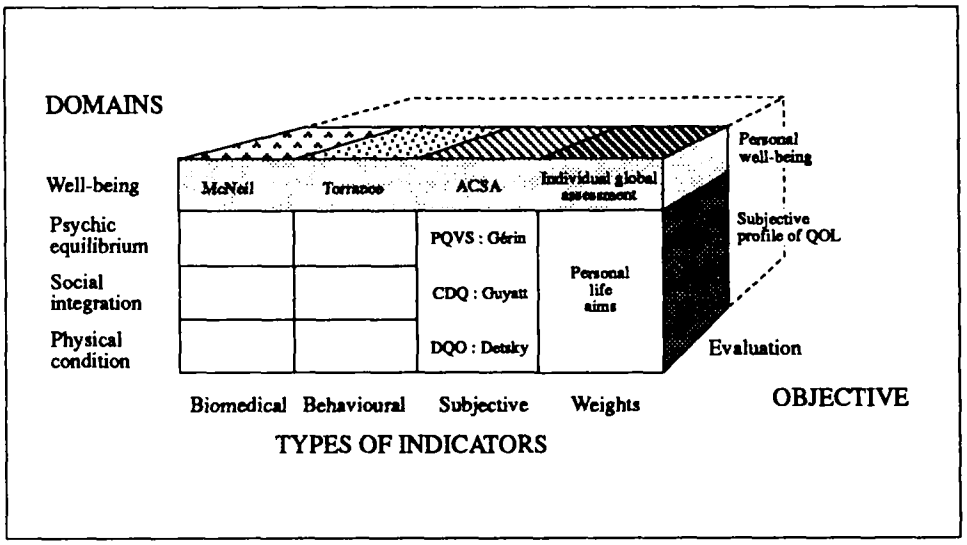

FIGURE 3. The patient's experience (References 32–41).

2. Risk calculation is not based on personal assessment but on objective measurement.

By varying probabilities attributable to the higher risk situation it is possible to assess the psychological value which the subject attributes to the certainty situation. Where the chances of success of the higher risk approach are reduced to 1% the patient must choose between the certainty of living with angina or the risk of undergoing an operation which is unlikely to succeed. The risk is not worth the gamble and the patient chooses the safe option. If, in contrast, however, the operation death rate is low (1%) the probability of surviving the operation is raised and the patient in this situation will opt for the gamble.

Where the chances of success are low, the patient will favor the status quo. In the contrary situation, he will tend to lean toward the higher risk approach. The only difference between these two situations is in the value p, the probability of success. As this increases the subject is less likely to choose the safe option and more likely to take the higher risk option. Finally, there is a threshold coefficient value where the patient is unable to choose between the two options. This value may be used to assess the current quality of life of the patient. If pain is severe or frequent, the value of the threshold coefficient is low.

If the patient will undergo anything to escape his current condition, the operation proposal is accepted even where the chances of success are limited, confirming the patient's poor state of health. If the pain is mild, the critical value for the coefficient is higher, the patient's present condition approaches that of good health; the patient does not accept the operation proposal unless it is almost certain to succeed.

The utility/preference approach has a number of advantages. First, this method produces a detailed measurement which combines mortality, morbidity, resultant physical sensory, and socio-emotional and cognitive effects, symptoms of the disease, and secondary effects of treatment into one single score. It allows calculation of a weighted life expectancy as a function of quality of life, which may not be done with specific profiles used to study the multiple effects of disease over time. Results and

costs may be brought together when they may be related to a fundamental domain. Secondly, the score directly reflects patient preference and is not influenced by weighting factors defined by the healthy population or by the practitioners caring for the patient. The instrument may be specific for the disease if appropriate parameters are chosen to define the areas to be addressed. The method has an undisputed scientific basis: decision in the face of uncertainty, described by Von Neumann and Morgenstern. Despite the indisputable applications of this mechanism, it cannot be denied that there are restrictions to its use. First, replies vary as a function of the context in which questions are set and second, it is not always possible to identify clinical variables which form the basis of the overall score. Finally, the sensitivity of a given indicator must be demonstrated in different disease states.

COLLECTIVE PREFERENCES

Measurement of collective preferences uses a group of individuals designed to represent the public interest to weight differences in states of health. The intensity of a problem may be fully reported by the patient but the importance it is given depends on the judgment of the healthy population. Whatever approach is used, quality profile or utility measurement, assessment of the significance or relative desirability of a given state of health is defined by external observation.

Profiles of Normalized Quality of Life

These use a single self-completed questionnaire which assesses different aspects of the quality of life. In contrast to multiple indicators which may be grouped together in batteries, the results of which may be combined into subscores for each domain, this is a large group of general indicators said to apply to all diseases. The best known are the sickness impact profile (SIP) (42) and the Nottingham Health Profile (NHP) (43). The SIP consists of 136 questions grouped into two domains: physical and psychological state, and five specific independent categories. The ensemble may be used to provide a global score. Each question assesses change in behavior and measures intensity of the upset. An interval scale using apparently equal gradations is used to assess the relative severity of each functional problem. This system was presented in 1975 to 108 Seattle HMO members and 25 health professionals. Each point was scored between zero and 15. Subdomain and overall global scores were calculated by dividing the sum of individual scores into the maximum possible score.

The NHP uses a two-part questionnaire. The first part consists of 38 questions with "Yes" or "No" responses, covering six domains: sleep, physical mobility, pain, effective reactions, social isolation, and emotional reaction. The second part assesses seven independent variables: work, salary, domestic work, interpersonal relationships, social life, family life, and sexual life, holidays and pastimes. Results are scored zero or one. Domains are not grouped together but points assessing each domain are weighted as a function of their relative severity. The reference technique used is pair comparison: each item in a domain is compared successfully to all other points within that domain. The subjectively more severe point is noted in each case. This system was used on a pilot group of 1,200 laymen without medical training to assess the frequency of points deemed more severe than others. Symptoms and problems were graded in a hierarchy, comparing mean standard deviation to frequency.

Profiles are not without merit: their reproducibility and validity have been well established (Figure 4). They also allow assessment of different domains of quality of life in one combined scale without using multiple measurement scales. This is easier both for investigators and patients. They do have problems, however, notably, they

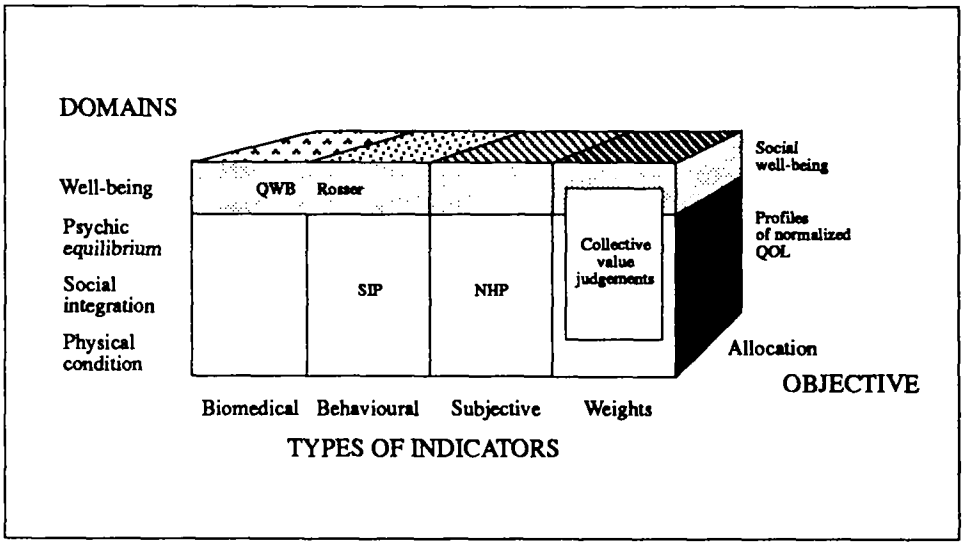

FIGURE 4. Collective preferences (References 42–44).

do not allow the specific consequences of a given disease on quality of life to be assessed. For example, physical autonomy may be assessed by means of a number of features assessing motor defects, particularly walking difficulties.

In venous diseases, walking, far from being a handicap is considered to be therapy, whereas standing upright and immobile, which is not listed in the NHP or the SIP, is a real problem for patients suffering from venous disease. The failure of the parameters used to relate to the specific problem leads inevitably to sensitivity failings or even validity problems as the functional defects explored may not be relevant. Deciding on the relative importance of different domains is also a problem. Where these do not always progress in the same direction, they must be weighted. In the absence of an overall score, overall assessment may be left to the evaluators' subjective judgment.

Measurement of Social Well-being

Quality of life may be assessed for each subject and related to a reference value established in a control group from the general population. The goal of measurement is not to assess "the importance which each of us attaches to our lives," but to produce an overall morbidity indicator through which the effects of actions influencing health may be judged. The works of Bush (8) and Rosser (44) addressed these concepts. This supplied economists with the tools necessary to divide fixed resources between competing projects.

Bush assessed the effects of disease by means of two criteria: functional incapacity and subjective problems. Functional incapacity was assessed in three domains: physical autonomy (PAC), mobility within the living area (MOB), and social activity (SAC). The corresponding scales were ordinal and contained four, five and five grades, respectively. The first assessed the autonomy of patient movement: mobility with or without difficulty, restriction to a wheelchair, or bed-bound. The second domain stressed practical surroundings and the distance which could be travelled. The third assessed social functions the individual could perform. These functions, of course, varied for each category within the population. In active subjects this was work, in those less than 15

years old, scholastic activities, and for retirees, pastimes were assessed. Finally, these three scales were combined but not added together.

Following elimination of impossible situations 29 functional levels were obtained from the 100 initial possible situations (4 × 5 × 5). The picture was completed with a list of 21 signs or symptoms. This allowed integration of inconsistent complaints (shivering, fever) to a precise diagnosis, vague symptoms (headache, dizziness), incapacitating disease even if individuals involved had not declared them (back pain) and simple deficiency (amputation). The combination of the 29 functional levels and 21 signs in five age groups produced, after removal of impossible situations, 343 case types. Each patient could be attributed to one such scenario which could be placed in a hierarchy to obtain a coefficient corresponding to the quality of life for each individual.

Scoring of preferences was performed using an equal appearing interval scale and 867 individuals from the general population were questioned. They were asked to score each scenario between zero and 15. The mean score from the group of assessors was calculated and a weighting coefficient between zero and one obtained reflecting the relative desirability of each scenario relative to death or good health.

The Rosser indicator worked on the same principal. Two scales were used. The first assessed functional adaptation of the individual to the environment in which he lived and classified physical incapacities; it used objective "signs": the absence of handicap or incapacity, minor alterations of social life, major alterations of social life, alterations of physical capacity, inability to perform normal functions, restricted to armchair, bed-ridden, or unconscious. The second scale assessed subjective impressions of disease by measuring the "suffering" associated with the disease process; an ambiguous term which was chosen for its encompassing definition: "pain and/or mental disorder and/or psychological effects of incapacity." The eight incapacity levels were combined with four "suffering" grades to produce 32 states of health. After removal of impossible combinations (unconsciousness and the presence of suffering, whatever the intensity) a system of 29 levels of classification of patients was obtained.

Once a stratification was available to grade the population of the function of the severity of the nature of their complaints, the question of relative severity inevitably emerged. Rosser and Kind in 1978 proposed that the concept of severity should be standardized using a relativity scale. In-depth analyses which were semistructured and lasted three to four and a half hours were performed using 30 health professionals, 20 patients, and 20 healthy subjects. Six marker conditions were selected to reflect the diversity of the 29 levels proposed. The description used to assess the domain of suffering was physical pain. The authors made no reference to the high or low morale of patients. First, the assessors were asked to grade scenarios in order of severity beginning with the least severe. Rosser and Kind then asked assessors to attribute a positive number to each marker, without defining in advance the upper limit. The only constraint used was that the numbers used should be in a ratio to the respective severity of conditions.

The precise question used was "how many times more serious do you estimate two to be by comparison with state one in the final analysis?" In order to enable assessors to be fully aware of the consequences of their choice, Rosser and Kind stressed that: "this ratio should indicate either the minimum number of mild cases which you feel are equivalent to one severe case or the relative proportion of a given resource which you feel would be justifiable in the treatment of a severe case, by comparison with a mild case."

The same procedure was used to assess other marker states and the 23 remaining intermediate situations. For each state, the value attributed to the nth state was calculated by its relationship with the n-1 state, without comparing all the n states with

each other. The relationship given by relative positions of a given state related to its predecessor was obtained by simple multiplication of the ratio by the figure corresponding to good health, defined as the origin on the scale. The overall group assessment was obtained finally by taking the median of all scores awarded: these median values, therefore, assessed a loss of utility resulting from a change in quality of life.

The third version of this indicator published by Kind and Rosser in 1982 calibrated medians by dividing them by a pivotal value; the figure attributed to death in the 1978 version, which transformed the relativity scale into an individual scale the extreme limits of which were one for good health and zero for death. The subtraction of scores corresponding to the changing quality of life from the ideal health score produced the corresponding quality of life coefficient.

REQUIRED QUALITY OF THE INSTRUMENTS

To be credible, measurement of quality of life must be pertinent, receivable, reliable, sensitive, and valid (45–50).

Content Value

The content value requires two conditions to be fulfilled: exhaustivity (the universe of complaints must be represented) and representation. The contents of a proposed instrument must cover the entire field in the area one is proposing to study, and it must contain a representative sample of terms or complaints from all those possible. A poorly defined universe is one of the worst possible types of error: it results in inadequate matching of the instrument to the universe it is designed to explore. A second source of bias originates from failure to adapt the relevant questions. The method by which questions have been chosen to construct the scale should always be specified. Construction of an instrument may depend on consultation with experts, or on statements made by the patients themselves. The choice of final questions may be based on methods which may or may not be scientific. The simplest method is to multiply the frequency of complaints by their severity, however, more sophisticated analyses such as principal component analysis may also be used.

Face Value

Face value depends on the quality of preparation: are the questions specifically precise for the domains and subdomains explored? Do they relate to a clearly defined period of time? Do they examine a fixed state of health, or a change in state of health? Are they worded in terms of capacity or performance? Is the procedure for combining different elements adequate?

Reliability

A scale is reliable if in measuring the same phenomenon on a number of occasions it produces similar results. To determine reliability the size of random measurement error must be assessed. If this is low the instrument provides a consistent measurement of the universe assessed. A number of authors describe this criterion as reproducibility, others refer to the precision of the instrument. Three methods exist to assess reliability: internal coherence, test-retest reliability, and interassessor reliability:

1. Internal coherence: the indicator is coherent when different elements are not contradictory. Such coherence exists when each facet of a domain and each domain within the instrument assess dimensions which are complementary and are not redundant. The Cronbach alpha coefficient is the most frequently used statistical measurement for this assessment.
2. Test-retest reliability: this is defined by

the similarity of successive measurements at different points in time and relating to the same feature measured by the same technique.

3. Inter-assessor reliability: this measures the agreement between different observers, assessing the same situation. The Kappa coefficient is the statistical parameter used for ordinal data and the intra-class correlation coefficient for continuum data.

Sensitivity

The sensitivity of an instrument is its capacity to detect clinically significant changes even if they are of low amplitude. An indicator is maximally sensitive when it detects all changes in a given variable over and above the imprecision due to measurement error. Guyatt (50) formulated a broadened definition of sensitivity by the term "responsiveness," which combined both reproducibility and sensitivity *per se*. Two further requirements must be fulfilled:

1. The questionnaire used must produce almost identical scores in stable subjects over time, that is, it must be reproducible, and
2. It must be able to demonstrate changes which occur when the subject's state of health improves or deteriorates.

Construct Validity

An instrument is said to be constructually valid if it measures what it truly purports to measure. This assumes both the absence of random error and systematic bias. Reliability is, therefore, a prerequisite, but is not sufficient for validity. For perfect validity, there must be no consistent error. In the absence of an undisputed reference standard, the validity of a measurement scale is obtained by comparing its results either to other indicators of quality of life assessing the same domain or to clinical indicators, and measuring any divergence or convergence. Only too often, instrument validation is performed through intuition.

CONCLUSION

The choice of an indicator depends on the answers to the four following questions: Does the user require an indicator producing discriminative or evaluative results? Does he wish to assess the overall quality of life or specific facets of the quality of life? Is the instrument to be used to follow patients over time, or at one point in time? Which opinion is to be used: that of the doctor, that of the population, or that of the patient? Only too often, the available instruments are used blind without clearly addressing these questions.

REFERENCES

1. Apgar V. A proposal for a new method of evaluation of the new-born infants. *Curr Research Anesth Analg.* July-Aug 1953:260–267.
2. Stevens S. On the theory of scales and measurement. *Sciences.* 1946;103:667–680.
3. Haski M, Moskowitz H. L'échelle sensorielle de Stevens-Moskowitz. *Rev Franç Marketing.* 1980; 2(8)1:5–18.
4. Boss JF. Quelques aspects de la measure des attitudes: les échelles multi-dimensionnelles. *Rev Franç Marketing.* 1970;34:23–44.
5. Kischner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis.* 1985;38(1):27–36.
6. Dupuy J-P. La science de décision en matière de santé: quelques éléments de réflexion. *Economie et Santé.* 1971;3:2–29.
7. Guelfi JD, Boben D. Echelle d'évaluation en psychiatrie. *EncyL Med Chir Psychiatrie.* 37200 A^{10} 10;1989:1–10.
8. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med.* 1977;96:716–721.
9. Sonnenberg FA. Théorie de la décision et de la stratégie médicale. In: Launois R, Régnier F, eds. *Décision thérapeutique et qualité de vie.* Paris: John Libbey Eurotext; 1992.
10. Patrick DL, Bush JW, Chen M. Method for measuring levels of well-being for a health-status index. *Health Services Research.* 1973:229–245.
11. Torrance G, Zipurky A. Cost effectiveness of antemartum prevention of RH Immunisation. *Clinics Perinatal.* 1984;11:267–281.
12. Williams A. Is it a wild use of resources? In Oli-

ver M, Ashley-Miller M, Woods D (eds.). *Screening for risk of coronary heart disease*. London: John Wiley; 1986.
13. Doubilet P, Weinstein M, McNeil B. Use and misuse of the term 'cost effective' in medicine. *N Engl J Med*. 1986;314(4):253–255.
14. Eisenberg J. Clinical Economics: a guide to economic analysis of clinical practice. *JAMA*. 1989; 262(20):2879–2886.
15. Katz S, Ford AB, et al. Studies of illness in the aged: the index of ADL: a standardised measure of biological and psychological functions. *JAMA*. 1963;185:314–319.
16. Harris A. Handicapped and impaired in Great Britain. London HMSO, 1971. Quoted by Culyer A. Measuring health; lessons of Ontario. Ontario Economic Council, 1978.
17. The Criteria Committee of the New York Heart Association. In: *Diseases of the Heart and Blood Vessels*. Boston Mass: Little Brown; 1964.
18. Goldman L. Comparative reproducibility and validity of system for assessing cardio-vascular functional class; advantages of a new specific activity scale. *Circulation*. 1981;39:207–210.
19. Karnovsky DA, Abelman WH, et al. The use of nitrogen mustard in the palliative treatment of carcinoma. *Cancer*. 1948;634–656.
20. Spitzer WO, Dobson AJ, et al. Measuring the quality of life of cancer patients. *J Chronic Dis*. 1981;34:585–597.
21. Mahoney FL, Barthel DW. Functional evaluation: the Barthel Index. *Rehabilitation*. 1979;22–23:61–65.
22. Grogono AW, Woodgate DJ. Index for measuring health. *Lancet*. 1971;1024–1026.
23. Priestman T. Evaluation of quality of life in patients receiving treatment for advanced breast cancer. *Lancet*. 1976;899–901.
24. Moles A. *Les sciences de l'imprécis*. Paris: Editions du Seuil; 1990.
25. Schipper H, et al. Measuring the quality of life of cancer patients: the functional living index. *J Clin Oncology*. 1984;2(5):472–485.
26. Selby P, Chapman JAW, et al. The development of a method of assessing the quality of life for cancer patients. *Br J Cancer*. 1984;50(1):13–22.
27. Guillemin F, Briançon S, Pourel J. Mesure de la capacité fonctionnelle dans la polyarrhrite rhumatorde: adaption française du health assessment questionnaire (HAQ). Forthcoming, *Revue du Rhumatisme*.
28. Meenan RF. The AIMS approach to health status measurement: conceptual background and measurement properties. *J Rheumatology*. 1982; 9(5):785–788.
29. Steinbrocker O, Traeger C. Therapeutic criteria in rheumatoid arthritis. *JAMA*. 1949;140(8): 659–662.
30. Mahler A, Weinberg D, et al. The measurement of dyspnea. *Chest*. 1984;85(6):751–758.
31. Croog SW, Levine S, et al. The effects of antihypertensive therapy on the quality of life. *N Eng J Med*. 1986;314(26):1657–1664.
32. Gérin P, Dazord A, et al. L'évaluation de la qualité de vie dans les essais thérapeutiques. Aspects conceptuels et présentation d'un questionnaire. *Thérapie*. 1989;44:355–364.
33. Gérin P. Dazord A, et al. L'évaluation de la qualité de vie dans les essais thérapeutics. In Pharmacologie clinique: actualité et perspectives. Strauch G, Husson JM (eds.). *Colloque Inserm*. 1989;185:159–181.
34. Guyatt G. The questionnaire in the assessment of cardio-respiratory disease: the McMaster approach. Workshop on the assessment of the effect of drug therapy on the quality of life in cardio-respiratory disease, Broadway, Worcestershire, April 25–26, 1985.
35. Detsky A, McLaughlin J, et al. Quality of life of patients on long-term total parental nutrition at home. *J Intern Med*. 1986;1:26–23.
36. Torrance G, Feeny D. Utilities and quality adjusted life years. *Inter J Techno Assessment Health Care*. 1989;5:559–575.
37. Gadreau M. Une mesure de la santé. Collection de l'Institut de Mathématiques Economiques No. 17, Sirey Paris, 1978.
38. Launois R. L'évaluation économique des stratégies theérapeutiques. *Réalités Industrielles, Annales des Mines*. Juillet–août 1991:81–86.
39. Launois R, Orvain J, Ounis I. Apport d'une mesure des utilités: Infections respiratoires récidivantes. *Rev Epidémio Santé Publ*. 1992;40:1–10.
40. McNeil B, Weichselbaum R, et al. Speech and survival. Trade-offs between quality and quantity of life in laryngeal cancer. *N Engl J Med*. 1981;305(17):982–987.
41. Bernheim J. L'auto-évaluation anannestique comparative (ACSA). I. Description d'une méthode de mesure de la qualité subjective de la vie des malades cancéreux. *Psychol Med*. 1983;15: 1615–1617.
42. Bergner M, Bobbit RA, et al. The Sickness Impack Profile: conceptual formalisation and methodology for the development of a health status measure. *Intern J Health Serv*. 1976;2: 393–415.
43. Buquet D. Indicateur de santé perceptuelle de Nottingham. Manual d'utilisation. Inserm. Unité 164, Mai 1988.
44. Rosser RM, Kind P, A scale of valuation of states of illness, is there a social consensus? *Intern J Epidemiol*. 1978;74:347–357.
45. Kaplan R, Bush JW, Berry C. Health status: type of validity and the index of well-being. *Health Services Research*. 1976;11:478–506.
46. Churchill GA. A paradigm for developing better measures of marketing constructs. *J Marketing Research*. 1979;16:64–73.
47. Ware J, Brook R et al. Choosing measures of

health status for individuals in general populations. *Am J Public Health*. 1981;71(6):620–625.
48. Ware J. Standards for validating health measures: definition and content. *J Chronic Dis*. 1978;40(6):743–480.
49. Israel L, Waintraub L. Méthodes de'évaluation psychométriques en gériatrie. Le choix d'un instrument et ses critères de fiabilité. *Presse Méd*. 1983;12(48):3124–3128.
50. Guyatt G, Walter S. et al. Measuring change over time. Assessing the usefulness of evaluative instruments. *J Chronic Dis*. 1987;40(2):171–178.