Advantages and Limitations of Utility Assessment Methods in Rheumatoid Arthritis

ARIEL BERESNIAK, ANTHONY S. RUSSELL, BOULOS HARAOUI, LOUIS BESSETTE, CLAIRE BOMBARDIER, and GERARD DURU

ABSTRACT. Utility assessment and cost-utility analyses such as costs/quality-adjusted life-years (QALY) are frequently presented to demonstrate the value of new treatment options in rheumatoid arthritis (RA). However, utility indicators require various methods that introduce significant methodological challenges, which directly influence the results and ensuing reimbursement decisions. Our objective was to review and discuss these challenges and the validity of frequently used utility assessment techniques in the context of RA. Coding the intensity of preferences or variations in patient satisfaction in order to assess utility implies extreme mathematical assumptions about a patient's rationality regarding his/her preferences towards different given health states. The construction and assumptions of commonly used "direct approaches" (standard gamble, time tradeoff, visual analog scale) and indirect approaches (EQ5D, HUI, SF6D) are presented. Other approaches such as transformation in utility of data from clinical (Health Assessment Questionnaire) or quality of life instruments ("mapping technique") are analyzed as they appear to generate uncertainty and a wide variation in estimated utility values in the context of RA. Utility assessment and cost-utility analyses in RA, which form the basis of the QALY, are frequently published and often requested by health technology assessment agencies to assist reimbursement decisions. However, when interpreting the results, the medical community must take into consideration the limitations and significant uncertainty of these approaches.

In light of these findings, real cost-effectiveness analyses based on observed clinical outcomes appear to be more robust and reliable to assist decision-making, particularly in the context of RA. (First Release Oct 15 2007; J Rheumatol 2007;34:2193–200)

Key Indexing Terms:
UTILITY QUALITY-ADJUSTED LIFE-YEARS RHEUMATOID ARTHRITIS

Economists have proposed "utility" indicators for use in medicoeconomic evaluations. Some regulatory authorities responsible for reviewing and assessing the value of new health technologies, such as CADTH (Canada), NICE (United Kingdom), and PBAC (Australia), tend to favor cost per quality-adjusted life-year (QALY)-type approaches, which are based on utility assessments, for determining and comparing the estimated value of different treatment strategies for one or more diseases. Thus, an increasing number of utility assessments are being conducted to support reimbursement decisions concerning new treatment strategies in rheumatoid arthritis (RA)[1].

However, constructing utility indicators introduces a number of important issues that directly influence the results and the ensuing decisions. Moreover, the medical community may not be familiar with these methods and their limitations. Because of the surge in rheumatology-related scientific publications and reports presenting QALY-based utility assessments and cost-utility analyses, a critical review of these methods is needed for a better understanding by clinicians of their advantages and limitations.

Utility Assessment Methodology

For health economists, utility assessment addresses 2 objectives: (1) the need for a synthetic outcome indicator to relate outcomes with costs for a given medical strategy; and (2) the need for a universal indicator to allow comparison between different diseases.

There are 3 main types of medicoeconomic evaluations that compare costs and outcomes:

Cost-benefit analyses, which compare the cost of a given treatment approach against its outcomes expressed in a mon-

From LIRAES, University Paris-Descartes, Paris; CNRS, National Center of Scientific Research, Lyon, France; Data Mining International, Geneva, Switzerland; University of Alberta Hospital, Edmonton, Alberta; University of Montreal, Montreal Rheumatology Institute, Montreal, Quebec; Centre Hospitalier Universitaire de Québec-CHUL, Quebec, Quebec; and University of Toronto, Mount Sinai Hospital, Toronto, Ontario, Canada.

A. Beresniak, MD, PhD, LIRAES, University Paris-Descartes and Data Mining International; A.S. Russell, MA, MB, BChir, University of Alberta Hospital; B. Haraoui, MD, University of Montreal, Montreal Rheumatology Institute; L. Bessette, MD, Centre Hospitalier Universitaire de Québec–CHUL; C. Bombardier, MD, University of Toronto, Mount Sinai Hospital; G. Duru, PhD, CNRS, National Center of Scientific Research.

Address reprint requests to Dr. A.S. Russell, Rheumatic Disease Unit, University of Alberta, 562 Heritage Medical Research Centre, Edmonton, Alberta T6G 2S2, Canada. E-mail: as.russell@ualberta.ca

Accepted for publication June 21, 2007.

etary unit (e.g., if the treatment outcome is life-years saved, the assessment of the benefit requires assignment of a monetary value to a life-year).

Cost-effectiveness analyses, which compare the cost of a given treatment approach against its outcomes expressed as a clinical efficacy indicator (e.g., in RA, the cost per remission or the cost to achieve American College of Rheumatology 20% response).

Cost-utility analyses, which compare the cost of a given treatment approach against its outcomes expressed as a utility value. By definition, cost-per-QALY analyses are cost-utility analyses.

Despite these very specific and largely consensual definitions, numerous publications present, without distinction, true cost-utility analyses as "cost-effectiveness" analyses. Although these concepts are based on different methodologies that are neither equivalent nor interchangeable, a treatment associated with a high cost per QALY is often mistakenly presented as a "cost-ineffective" treatment.

This incorrect practice of presenting and interpreting the results of cost-utility analyses under the banner of "cost-effectiveness" seems to be gaining currency, for the following reasons.

1. Cost-utility analyses rely on a much larger number of assumptions than cost-effectiveness analyses. Thus, by presenting a cost-utility analysis as a "cost-effectiveness" analysis, one does not have to validate a certain number of assumptions specific to cost-utility analyses.

2. Unlike true cost-effectiveness analyses, cost-utility analyses are based on complex economic evaluation concepts that are commonly neither well known nor understood by the medical community, hence the tendency for some authors to simplify by presenting them as "cost-effectiveness" analyses.

Unfortunately, this terminological confusion, which is propagated by the scientific literature, does not make it any easier for an average reader to decipher the methods used.

Assessing utility and preferences. In economic parlance, utility is a concept associated with the preference of an individual regarding a set of objects. In medicine, the individual is the patient, and the objects are often different health states for which the patient expresses his/her preference. The concept of utility is thus used to rank different health states on the basis of a patient's preferences according to the satisfaction that he/she derives from these health states.

The different options to be ranked according to individual preferences can be monetary gains, quantities of goods, lengths of time, health states, and so on. They may also concern more complex notions involving the idea of uncertainty, such as contracts or standard gambles with different choices.

Traditionally, a mathematical relationship of the function "preferred to" type is used to represent an individual's preferences. Utility is then expressed as a numerical value referred to as a "utility function" to each option to be considered. Coding the intensity of preferences or variations in satisfaction implies extreme mathematical assumptions about the patient's rationality regarding his/her preferences concerning different given health states.

In an effort to solve this difficulty, health economists came up with the idea of using solutions derived from game theory, such as those described by von Neumann and Morgenstern[2]. Their work does, in fact, propose a set of theoretical solutions for defining direct preferences concerning different options. The principle consists in eliciting preferences, not with regard to situations per se, but with regard to contracts that involve these situations in an uncertain environment. These contracts are referred to as "lotteries" (or "standard gambles") and define the accepted probability of choosing a given situation. This approach, in order to be verified, requires assumptions about the rationality of the individual's preferences and his/her behavior toward risk. For instance, the individuals as a group must be considered to have similar risk-taking preferences (risk-neutral), which, in the real world, is seldom confirmed, especially in medicine. This patient-risk non-neutrality is one of the most significant limitations of applying Neumannian utility theory to measuring utility values in medicoeconomic evaluations because the resulting bias cannot be controlled.

Choosing the system of reference. When assessing utility, one must choose a system of reference to measure it. One may choose a range between 0 and 1. Very frequently, a utility of 0 represents the health state "death," and the utility of 1, the health state "perfect health."

Of course, 2 utility values measured in 2 different systems of reference cannot be compared. This would make no more sense than saying that it is hotter in a country where the temperature is 32° (measured in the Fahrenheit system of reference) than in a country where the temperature is 0° (measured in the Celsius system of reference). However, this is what is commonly and inappropriately done when cost-utility analyses are benchmarked ("league tables") based on different methods[3,4].

Quality of life and utility assessment. The essential advantage of defining utilities by means of a utility function is to be able to incorporate quality-of-life measures into a medicoeconomic evaluation and to construct a cost-utility analysis.

Actually, directly measuring quality of life — for example with the Medical Outcomes Study Short-Form-36 Health Survey (SF-36) — does not enable one to perform cost-utility analyses, since costs cannot be divided by a quality-of-life measurement. Quality-of-life scales are strictly "ordinal"-type instruments. In other words, they adequately define an order, but the graduations are not all equal and the "unit" is not defined. These scales therefore cannot be used to perform mathematical operations, such as multiplication and division, in order to arrive at results expressed as a "cost-per-unit-of-quality-of-life".

Theoretically, measuring a utility value in relation to a health state value enables one to perform "cost per unit of util-

ity" analyses, thanks to the Neumannian property of certain utility functions. From this theory, numerous preference identification methods have been derived to lead to a utility function. Traditionally, they are divided into direct and indirect methods.

Direct Methods

Utility measurement methods are said to be direct when the patients' preferences are self-reported. These methods are the standard gamble, time tradeoff, and visual analog scales (VAS).

Standard gamble. The standard gamble, a utility measurement method derived from game theory[2], involves gambling between contracts. The utility measurement procedure in the standard gamble technique is as follows:

The system of reference in which utility will be measured is chosen to be between 0 and 1. Death is often represented with a utility of 0, whereas perfect health is often represented with a utility of 1.

The subject is offered 2 alternatives. Alternative 1 (the gamble) represents 2 possible outcomes: either the patient returns to a perfect health state (probability p) or the patient dies immediately (probability 1 – p). Alternative 2 represents the health state of interest or the current health state of the patient (one of less than optimal health). The patient then is asked what probability, p, of dying now, he/she would accept in order to change from his/her current health state (to be measured) to the health state "perfect health." This is the "gamble." Thus the odds risk of death now is altered progressively until the patient is at the point of uncertainty or equivalence between clearly accepting or refusing the odds to gain perfect health. The measured utility of his/her current health state is thus 1 – p. For example, if a patient in the health state "walk with a cane" accepted a 5% risk of dying to achieve complete healing, his/her utility would be 1 – 0.05, or 0.95.

Time tradeoff. Time tradeoff is a utility measurement method that is similar but uses contracts involving a choice between situations and lengths of time. This method is less used because the underlying theory has never really been described. It has even more theoretical limitations than the standard gamble. Further, personal experience suggests patients find the questions relatively disconcerting.

The time tradeoff method presents the respondent with the task of determining how much of their life they would be willing to give up to be in a better versus a poorer health state. The time tradeoff procedure consists in proposing the following alternatives to the patient:

A: Live T years in health state X (e.g., 10 yrs in perfect health); or

B: Live t years in health state x (e.g., 20 yrs in a wheelchair), where the time T with the preferred health state must be shorter than the time t with actual health state, as in the example above where t > T (20 yrs > 10 yrs) and where X is preferred to x (e.g., perfect health is preferred to living in a wheelchair). Hence, time T is shortened until the patient feels indifferent between the 2 proposed alternatives.

The life expectancies t* for which the patient is indifferent between the 2 alternatives can then be determined, in which case it is said that the utility of health state x is equal to T/t. For example, if the patient is indifferent between situation A "10 years in perfect health" and situation B "20 years in a wheelchair," it is said that the utility of the health state is equal to 10/20 (0.5).

One of the theoretical problems posed by the time tradeoff method is that it is proposed that preferences on pairs (time duration, health state) can be calculated from single preferences of time duration on one hand, and single preferences of health states on the other hand, using a simple multiplication of the 2 utilities. This problem is the same as for the QALY method described later.

Visual analog scales. VAS are measurement techniques using normed scales or small graduated rulers on which the patient evaluates intensity of variations of his/her preferences in response to specific questions. The "0" of the scale corresponds to a health state previously selected as "origin" and the "1" corresponds to a health state previously selected as the unit.

VAS are widely used to measure certain aspects of health such as pain, and have also been proposed as a method for determining patient preferences[5]. The respondents are asked to visually evaluate on a small ruler or straight line their level of preference on a continuum between "the most desirable" and "the least desirable." One is then supposed to be able to read the utility value directly on the scale.

In fact, there is no reason for the proposed graduations to be equal, a necessary condition for the use of utility values in cost-utility analyses. Further, numerous experiments have successfully shown that reproducibility is very low in a given patient[3]. Lastly, each patient uses his/her own system of reference, which compromises the comparability of evaluations between patients and which therefore, compromises population studies. Its great simplicity explains why this technique is still widely used.

Indirect Methods: EuroQol (EQ5D), Health Utilities Index (HUI), SF6D

Given the difficulty of operationally determining patient preferences using the above-mentioned direct methods, some authors have suggested developing indirect ones. These are usual quality-of-life questionnaires (that describe a health state profile as opposed to a preference) that have a scoring procedure developed for calculating utility values. Each health state profile is therefore associated with a specific utility value, that is, a preference regarding this health state.

These methods are said to be "indirect" because they do not measure patient preferences directly. Despite numerous methodological criticisms, they enjoy a certain degree of popularity because they are very easy to use (just fill in a ques-

tionnaire). The most widely used instruments are the EQ5D, HUI, and SF6D.

Apart from the validation of the questionnaires per se, which involved the same validation methods as those used for quality-of-life instruments, the main methodological criticisms concern the calculation of utility values.

For example, an 8-item questionnaire with 4 proposed answers for each question gives a total possible number of health state profiles of 4^8, or 65,536 possibilities, which should be ranked between 0 (death) and 1 (perfect health). This can only be done by making numerous assumptions and approximations.

The EQ5D (or EuroQol) instrument consists of 5 questions on mobility, self-care, pain, usual activities, and psychological status with 3 possible answers for each item, with 243 possible health states. Because it was impossible to generate direct valuations for all those health states, the authors used a procedure that allowed interpolation of valuations for all EuroQol states conducting direct valuations on a subset of these with VAS and time tradeoff techniques. A utility value was then attributed to each of the 243 possible health states. However, this value depends on how the questions were proposed by the authors, how these were scored, and how the question score was translated into utility values. Carr-Hill[6] mentions that the EuroQol authors[7-9] used numerous simplifications of data from VAS such as ignoring a certain number of health states, considering that a large number of very different health states could have the same utility value, and postulating that the gradation intervals on a scale are regular, although they do not provide any proof of this. Gafni and Birch[10] show that the EuroQol suffers from several major limitations and thus cannot provide a valid measure for use in economic appraisals or studies concerned with evaluating healthcare intervention, as proposed by its proponents. They state that the EuroQol does not reflect patient preferences but rather social preferences or even the preferences of its authors, depending on the technique selected to extrapolate social tariffs (i.e., utility values). This is why we recommend not to use EuroQol results to calculate QALY to assist public decisions, as such decisions would be unsubstantiated and could be highly challenged using different methods leading to divergent utility values.

The principle on which the HUI[11] was designed is very similar to that underlying the EuroQol, since it was a question of developing a health state questionnaire that was not specific to any given disease and where the scoring procedure yielded a utility value for each health state profile. The third version of the HUI (Mark III) consists of 8 items referred to as "attributes": vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain.

To this end, a small number of health state profiles were selected and assigned utility values using standard gamble techniques in normal healthy subjects — not patients. "Mapping" techniques (as described hereafter) were then used to extrapolate the utility values obtained to the thousands of possible health state profiles resulting from all the combinations of the 8 attributes[11,12].

The SF6D is a descriptive system of 6 dimensions extracted from the 8 dimensions of the SF-36 quality of life generic questionnaire to generate numbers of health profiles consisting of 6 dimensions with levels[13]. Derived preference weights have been revealed using standard gamble, time tradeoff, and VAS applied to a test population. Like EQ5D and HUI, SF6D has been tested in RA with different degrees of responsiveness[14].

These 3 instruments share the same methodological limitations: the significant uncertainty of calculated utility values.

Limitations of composite outcome indicators (QALY). The QALY indicator was proposed in the early 1980s in order to take into account both a patient's quality of life and the length of time during which this quality of life is experienced. It is therefore a composite indicator, whose formula is as follows:

QALY = number of life-years gained (survival) * utility

The assumptions underlying the calculation of QALY are well known and widely published[15]. The methodological criticisms of QALY are fueling an international debate between their supporters and detractors. Aware of the limitations of this indicator, its supporters believe that it is better to have an imperfect synthetic indicator than to have none at all, while the detractors of QALY think that it is better not to have any synthetic indicator than to generate invalid results and comparisons that affect medical decision-making.

One set of criticisms concerns the multiplicative assumption, which is the same problem for the time tradeoff method: utility of one pair (time duration, health states) should be equal to the product of the utility of each component of the pair (time-duration utility and health-state utility). The validity of this approach using utility "multiplication" was tested[4] in a patient population, and it has not been possible to verify it in real-world situations, which calls into question the validity of multiplying utilities to calculate QALY.

The following is a simple example, which demonstrates the difficulties that may occur when multiplying time utilities and health-states utilities to calculate QALY to compare interventions. Suppose that there is a choice to be made regarding a travel destination. The first option would be a 2-month stay in a city known to have temperatures of 5°C. The alternative would be a 1-month stay in another city with temperatures of 25°C.

The preferences regarding duration and temperatures are defined by the utility functions from which an individual will derive preferences for a longer or shorter stay and higher or lower temperatures. The number of QALY for this decision is defined as duration * temperature. The first option would then lead to a QALY value of 10 (2 mo × 5°) and the second, to a QALY value of 25 (1 mo × 25°). Therefore, the second option would appear to be preferred over the first.

Now consider the same example with exactly the same temperatures, but this time the temperatures are expressed

using the Fahrenheit scale rather than Celsius. With corresponding temperatures of 41°F and 77°F, this results in 82 and 77 QALY, respectively. In this case, the first option would be preferred. However, this is the opposite conclusion as compared to the original calculation! In this case it is therefore impossible to distinguish which of the 2 options is preferable, without including the preference of an individual. The same concept applies for assessing patients' health preferences.

Another important criticism of QALY is the fact that their value will depend directly on the method chosen to assess utility (direct or indirect methods). As reported by several authors, one can obtain different QALY figures simply by changing one's utility assessment method. Despite the methodological debate over the validity of the QALY indicator, some health technology assessment agencies still prefer to use cost-per-QALY analyses to assist decision-making and have suggested cost-per-QALY "commonly accepted thresholds" above which a product will not be reimbursed and below which it may.

Thus, it may simply become a matter of finding the right utility assessment method to maximize or minimize results expressed as QALY, depending on what one wishes to demonstrate. Marra, *et al*[16] showed that utility scores yielded by indirect methods — the HUI, EQ5D, and SF6D questionnaires — in patients with RA were statistically significantly different. Similarly, Conner-Spady and Suarez-Almazor[17] found significant differences in the utility scores from different instruments (EQ5D, SF6D, and HUI) and warned about the validity of their use in cost-utility analyses.

This is also why the QALY indicator is not recommended in certain countries, such as France, where good pharmacoeconomic practice recommendations state the following (pharmacoeconomic practice recommendation No. 25, 2002):
The (QALY) aggregation rule poses many problems in terms of both methodology and philosophy;
The limited robustness of this approach allows the manipulation of the conclusions of a study;
In the current state of research, it is not recommended that public health decisions be based on study results expressed in terms of QALY...given the possibility of arriving at divergent results from the same observed data.

Issues Concerning the Conversion of Health States Questionnaires into Utility Values

In order to skip a specific data collection, some authors propose "inferring" a utility value from another type of questionnaire, such as a quality-of-life or health state questionnaire. This approach, referred to as "mapping," consists in hypothesizing about utility values without having obtained them from patients, whether by direct or indirect methods. Mapping is thus a technique that consists in establishing a link, although it may not exist, between 2 measures so that by knowing the value of one measure (e.g., health state score) and the mathematical relationship that describes the link, one can calculate the value of the other measure (e.g., a utility value).

More specifically, in the context of RA, by knowing, for example, the value of a patient's HAQ score, we could look for the mathematical relationship that would enable us to calculate the utility of his/her health state without having to gather this information from the individual. Mapping techniques are commonly used in econometric science to infer the value of an economic variable at a given point from data provided by a time series, or to predict the value of this variable from the values of other economic variables (e.g., the value of production, knowing the amount of work and the amount of capital spent by a company). Most often, econometric studies are very rigorous and use batteries of statistical tests to validate the forecasting model and its underlying assumptions.

It is not, for example, possible in econometric science to present a forecast without clearly stating the assumptions underlying the model that was used, without having performed all the necessary tests to assess its robustness and quality, and without providing all of the confidence intervals (CI) for the predicted values. These precautions are very seldom, if ever, taken in medicoeconomic publications that present forecasting cost-utility models. The authors generally provide a simple adjustment, which may lead to mathematical incongruities, as illustrated by the following example.

Michaud and Wolfe[18] proposed a means of converting HAQ scores into EuroQol utility values. The observed values of the HAQ scores collected from 42,751 questionnaires are presented by category. The corresponding values for the American version of the EuroQol (EQ5D-US) are mean utility values from questionnaires where the value of the HAQ score falls in a given category.

To illustrate this concept, we will try to perform a linear regression on this 2-entry (HAQ and EuroQol) dataset (Figure 1). Since the raw data on 42,751 pairs (EQ5D-US, HAQ) have not been published, we will assume, in our example, that all the values for each HAQ score category are at the center of the category and that all the utility values are close to their mean for each category (without taking their differences into account). These assumptions result in considerably reducing observational differences and intentionally improving the quality of the predictive model in our example. Let us try to perform a linear regression on this 2-entry (HAQ and EuroQol) dataset. Assuming that there is a close relationship between the EQ5D-US and the HAQ, there are 2 measures, a and b, such that:

$$\text{EQ5D-US} = a\,\text{HAQ} + b$$

A linear regression from 42,751 observations enables us to do the following: estimate, on a selective basis and by CI, the values of the measures a and b; test their significance (are they significantly different from 0?); calculate an estimate of the coefficient of linear correlation (the closer the coefficient is to 1, the better the model will be).

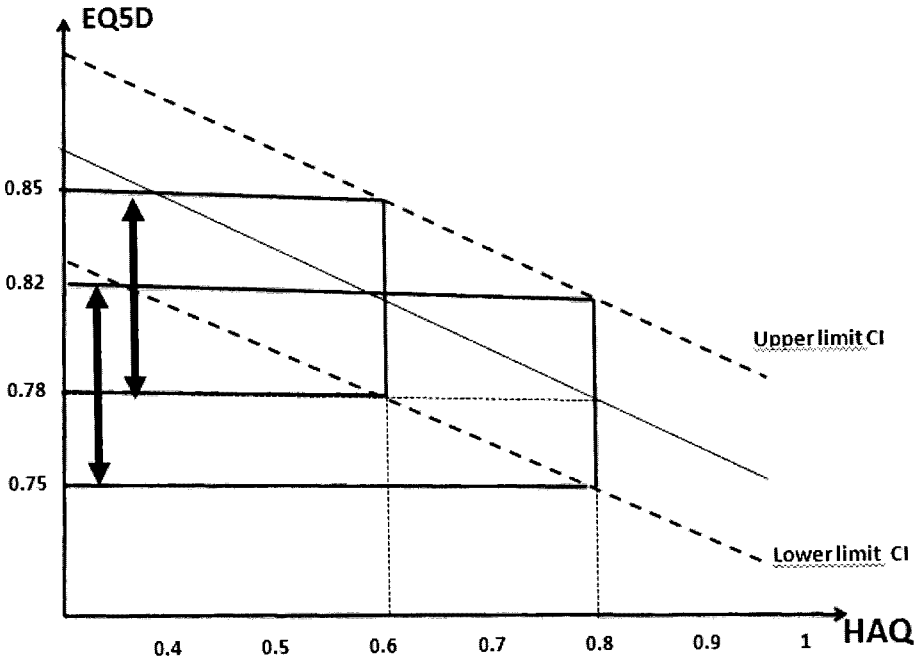

Figure 1. Linear regression on a 2-entry (HAQ and EuroQol) dataset showing confidence intervals of predictive EQ5D values from HAQ values

It must be understood that this linear regression, like any other regression, is based on a certain number of technical assumptions that need to be tested. They include, among others, residual normality, non-autocorrelation, and the assumption of homoscedasticity (requirement that the predicted values have the same variance around the regression line). Assuming that these tests have been performed properly and are acceptable, based on our example, we obtain the following estimated model:

$$EQ5D\text{-}US = -0.1749HAQ + 0.9279$$

a and b are significantly different from 0 ($p < 0.0001$ for each measure). The estimated coefficient of linear correlation is 0.98. The coefficient of correlation is significantly different from 0 ($p < 0.0001$).

These properties show that the linear regression is of very good quality, since the coefficient of linear correlation is very close to 1 and the measures a and b are significantly different from 0. Knowing a health state value on the HAQ scale, we can use this linear formula to calculate a utility value for the EQ5D-US scale. For example:

For a HAQ score of 0.6, the predicted value of the EQ5D-US score is 0.82, with a 95% CI of 0.78 to 0.85.

For a HAQ score of 0.8, the predicted value of the EQ5D-US score is 0.78, with a 95% CI of 0.75 to 0.82.

Despite the quality of the linear regression, the 2 CI are superimposable. Therefore, we cannot reject the assumption of equality of predictions or conclude that they are statistically different. The 2 calculated utility values, 0.82 and 0.78, simultaneously fall within both CI. They may thus also correspond to each of the 2 HAQ scores, 0.6 or 0.8. Therefore, it is impossible to differentiate between these 2 utility values calculated from the HAQ scores of 0.6 and 0.8. This inability to discriminate can therefore compromise the interpretation of the results of a cost-utility analysis in which a scale conversion method was used.

As a practical illustration, we will assume that a reference treatment A has a HAQ-measured 1-year efficacy of 0.6 and that another treatment B has a HAQ-measured 1-year efficacy of 0.8. After converting the HAQ score to a utility value using the linear conversion formula presented above, the number of QALY will be equal to 1*0.82, or 0.82 for treatment A and to 1*0.78, or 0.78 for treatment B. Assuming that the cost differential between treatments A and B is $800, the incremental cost-utility ratio of B in relation to A (expressed as the incremental cost per QALY) will therefore be equal to:

$$(\text{Cost A} - \text{Cost B})/(\text{QALY A} - \text{QALY B})$$
or
$$\$800/(0.82 - 0.78) = \$20{,}000/\text{QALY}$$

Under this cost-QALY demonstration, product B would then be reimbursed based on an arbitrary acceptance threshold of $50,000 per QALY. However, should the utility value of treatment B be 0.81 instead of 0.78 (a very similar value within the same CI), the incremental cost-utility ratio, expressed as a cost per QALY, would then be equal to:

$$\$800/(0.82 - 0.81) = \$80{,}000/\text{QALY}$$

This time, product B would not be reimbursed, since its

incremental cost-utility ratio would be well above the $50,000/QALY used by some health technology assessment agencies.

Although these 2 utility values for product B, 0.78 and 0.81, are very similar and not statistically different, they can yield incremental cost per QALY ratios as different as $20,000 and $80,000 per QALY, with major repercussions on reimbursement decisions made by public authorities favoring cost-utility assessments. It is therefore quite important to examine the significance of estimated utility values and their potential to discriminate different alternatives. Any model carries uncertainty over calculated values, and this uncertainty must be taken into account before using these values. As scale conversion methods are characterized by a high level of uncertainty, they may compromise the practical use of predicted values in cost-utility analyses. Thus, this uncertainty inherent in prediction methods should be addressed and presented systematically, which is, unfortunately, almost never done in such analyses.

Aggregating individual utilities into a group utility. Another delicate problem arises when the preferences of interest concern not only one individual, but a group of individuals, often referred to as "society." Measuring society's preferences is a problem very familiar to public economics specialists: How does one define society's preferences from the preferences of its members?

K.J. Arrow[19], a 1972 Nobel Prize recipient in economics, showed that individual preferences could not be aggregated into a group preference without choosing to overlook at least 1 of the 4 requirements: unanimity, transitivity, independence of the nonrelevant alternatives, and absence of a "dictator" (imposing his own preference). Each aggregation procedure should be checked with these requirements. This is routinely done whenever utility or cost-utility analyses are carried out in groups of patients. Thus, most authors using utility techniques suggest that the utility measurement for society be equal to the sum (or mean) of the utilities of its members. However, this approach is not based on any justification or valid theory.

The difficulty of obtaining group utilities is seldom discussed in the medicoeconomic literature, as it would invalidate most cost-utility analyses.

Discussion

Not only are direct utility measurement methods difficult to use in clinical trials, but their limitations compromise response reproducibility. Moreover, Bansback, *et al*[1] report that utility measurements provided by direct methods (standard gamble and time tradeoff) correlate weakly with the clinical outcomes observed in RA.

It was in an effort to address these difficulties that indirect methods were developed. Unfortunately, it is relatively easy to validate a questionnaire but much more difficult to validate a scoring procedure for arriving at a utility value. This is why these questionnaires draw much criticism with regard to the validity of the utilities generated and to their lack of sensitivity in chronic diseases, such as RA. A number of articles have also shown significant differences in utility values according to the type of indirect method used[16,17]. Similarly, Suarez-Almazor and Conner-Spady[20] and our own team[14] found that, upon using direct and indirect methods in several populations (general population, a population of patients with RA, a population of physicians), the results differed significantly according to the method used and the test population. When used in cost-utility analyses, the results led to figures ranging from $40,000 US to $220,000 US per QALY, which may result in significantly different reimbursement decisions from health authorities. In addition, given the methodological challenges they represent, few authors have tested direct utility assessment methods in patients with RA. Ariza-Ariza, *et al*[21] tested a direct method (time tradeoff) on 300 patients with RA, in comparison with an indirect questionnaire-type method (EuroQol), and found that the results were not convergent.

Lastly, Jorstad and colleagues[22] found that, despite a good relative correlation between different utility values obtained from 4 indirect questionnaires (15D, EQ5D, SF6D, and EQ VAS) tested in a population of 1,041 patients with RA, the utility values were significantly different for the same health state. They concluded that when these differences are incorporated into cost-utility analyses, they can lead to divergent results and thus have major consequences in terms of potential reimbursement decisions pertaining to RA treatments.

Given the difficulties in gathering and interpreting data in direct and indirect methods, an increasing number of authors propose, as a practical solution, the development and use of tables for converting from health state questionnaires (HAQ, SF-36, etc.) to utility values. We have seen that, even in the presence of a potential, very strong correlation between 2 scales, the need to take into account the CI for predicted values compromises discrimination between utility values, which makes this approach very insensitive. For instance, Wong and colleagues[23] determined that major improvements on the HAQ scale were necessary in order to generate a QALY benefit.

Regardless of the utility assessment method used, measured utility values are almost always used to calculate a QALY indicator in a cost per QALY-type analysis. As the QALY indicator is the product of survival (expressed as a number of life-years gained) times utility gained, not only do QALY results depend directly on the utility measurement techniques, but the very construction of the QALY indicator penalizes all non-lethal diseases, i.e., most chronic diseases (including RA), for which treatments have little or no influence on survival. It is therefore not surprising that most innovative products for RA yield results that seem high when expressed as a cost per QALY. This is due to the lack of sensitivity of and the dissimilarities between utility measurement techniques, and to the fact that there is little or no effect on survival.

Conclusion

Methods for evaluating patient preferences address the laudable concern of being attentive to the patient's point of view when taking a treatment's efficacy into consideration. However, the large number of proposed utility measurement techniques and the very wide differences in the results that they yield invite extreme caution when interpreting cost-utility results.

All the methodological problems in utility assessment reside in the fact that a numerical value is assigned to a preference and that this value is used as if it were a "real number" free of uncertainty. Adapting some part of the game theory to evaluations for the purpose of calculating utility values to assist decision-making in healthcare is probably very interesting from a theoretical and research standpoint. However, the behavior of patients is not the same as that of gamblers in a casino and does not necessarily lend itself to the assumptions underlying these theories. In fact, these techniques such as standard gamble do not correspond to the typical decision-making task in health, where multiple potential outcomes are possible and the choice of 2 options as certain as death and perfect health are not scenarios that typically confront people.

Quantifying qualitative notions is indeed a real problem in decision theory and mathematical economics. While these disciplines can make a very significant contribution to the evolution of medical sciences, the theories that they advance are applicable only if their assumptions are validated beforehand, which is practically never done in published cost-utility analyses. Consequently, for chronic disease such as RA, it appears that real cost-effectiveness analysis based on observed clinical outcomes (success rates, such as achieving ACRn or remission) are methodologically more robust and reliable to assist decision-making. Compared to calculated utility values, clinical outcomes reflect true treatment outcomes with evident "face-validity" for clinicians. The quality-of-life dimension can certainly be studied by itself using validated generic or specific instruments, and presented separately.

The non-universality of the clinical indicators does not seem to be a problem as such, as physicians rarely need to compare diseases as different as RA and Alzheimer's disease. Within a given specialty, such as rheumatology, the variety, choice, and quality of the clinical indicators matter most in their ability to be used to compare new products, new medical practices, or clinical studies. Sound cost-effectiveness analyses can thus be especially relevant from this standpoint, without any need to deal with uncertain utility values.

REFERENCES

1. Bansback NJ, Regier DA, Ara R, et al. An overview of economic evaluations for drugs used in rheumatoid arthritis. Focus on tumour necrosis factor antagonists. Drugs 2005;65:473-96.
2. von Neumann J, Morgenstern O. Theory of games and economic behavior. 3rd ed. Princeton: Princeton University Press; 1953.
3. Gerard K, Mooney G. QALY league table: handle with care. Health Eco 1993;2:59-64.
4. Duru G, Auray JP, Beresniak A, Lamure M, Paine A, Nicoloyannis N. Limitations of the methods used for calculating quality-adjusted life years values. Pharmacoeconomics 2002;20:463-73.
5. Torrance GW, Feeny D, Furlong W. Visual analog scales: Do they have a role in the measurement of preferences for health states? Med Decis Making 2001;2:329-34.
6. Carr-Hill RA. Health related quality of life measurement – Euro style. Health Policy 1990;16:199-208.
7. EuroQol, a new facility for the measurement of health related quality of life. The EuroQol group. Health Policy 1992;20:321-8,329-32.
8. Nord E. EuroQol: health related quality of life measurement. Valuations of health states by the general public in Norway. Health Policy 1991;18:25-36.
9. Brooks RG, Jendteg, Lindgren B, Persson U, Bjork S. EuroQol:health related quality of life measurement. Results of the Swedish questionnaire exercise. Health Policy 1991;18:37-48.
10. Gafni A, Birch S. Searching for a common currency: critical appraisal of the scientific basis underlying European harmonization of the measurement of health related quality of life (EuroQol). Health Policy 1994;28:67-9.
11. Feeny D, Furlong W, Boyle M, Torrance GH. Multi-attribute status classification systems. Health Utilities Index. Pharmacoeconomics 1995;7:490-502.
12. Boyle MH, Torrance GW. Developing multiattribute health indexes. Medical Care 1984;22:1045-57.
13. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998;51:1115-28.
14. Russell AS, Conner-Spady B, Mintz A, Mallon C, Maksymowych W. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. J Rheumatol 2003;30:941-7.
15. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. Operation Research 1980;28:206, 223.
16. Marra CA, Woolcott JC, Kopec JA, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571-82.
17. Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality adjusted life-years by different preference-based instruments. Med Care 2003;41:791-801.
18. Michaud K, Wolfe F. EQ5D changes rheumatoid arthritis quality of life in United States: A retrospective study of 11,289 patients [abstract]. Arthritis Rheum 2005;52 Suppl:S400.
19. Arrow KJ. Social choice and individual values. New Haven: Yale University Press; 1970.
20. Suarez-Almazor ME, Conner-Spady B. Rating of arthritis health states by patients, physicians, and the general public. Implications for cost-utility analyses. J Rheumatol 2001;28:648-55.
21. Ariza-Ariza R, Hernandez-Cruz B, Carmona L, Ruiz-Montesinos D, Ballina J, Navarro-Sarabia F. Assessing utility values in rheumatoid arthritis patients: a comparison between time-trade-off and the EuroQol. Arthritis Rheum 2006;55:751-6.
22. Jorstad IC, Kristiansen IS, Uhlig T, Kvien TK. Performance of four utility measures in 1041 patients with RA: Well correlated but differing widely in valuing health states. Arthritis Rheum 2005;52:S660-1.
23. Wong JB, Singh G, Kavanaugh A. Estimating the cost-effectiveness of 54 weeks of infliximab for rheumatoid arthritis. Am J Med 2002;113:400-8.