"The trouble with QALYs..."

MARTIN KNAPP and ROSHNI MANGALORE

Abstract. This paper summarises the use of QALYs in evaluating changes in mental health states, highlighting the benefits and challenges of their use in this field. The general principles underlying the QALY measure and the most common methods of measuring QALYs are discussed briefly. Evidence of the usefulness and problems of using this generic measure of health-related quality of life are provided from a sample of recent studies relating to depression, schizophrenia, attention deficit hyperactivity disorder and dementia. In each case, attempts were made to use QALYs to measure changes in health states. While in principle, the QALY is enormously attractive, its suitability for measuring changes in many mental health conditions remains open to doubt as existing tools for generating QALY scores such as the EQ-5D have tended not to perform sufficiently well in reflecting changes in many mental health states. New developmental work is needed to construct better QALY-measuring tools for use in the mental health field. Both the conceptualisation and measurement of QALYs need to be built on a valid, comprehensive model of quality of life specific to a mental health disorder, to ensure that the resultant tool is sensitive enough to pick up changes that would be expected and seen as relevant in the course of the illness.

CHOOSING BETWEEN ALTERNATIVES

Obviously, the main aims of a mental health system are to alleviate symptoms and improve quality of life, but it is widely appreciated that regard must also be paid to the economic consequences of the treatments and support arrangements that are offered. One consequence has been the growing attention paid to the cost-effectiveness as well as the effectiveness of health care interventions. The underlying reason is scarcity: there are insufficient resources to meet all of a society's needs or demands, and so difficult choices have to be made between competing uses of those resources. In particular, decision makers with responsibility for allocating resources must weigh up the outcomes achieved in those various different uses of resources and the costs of achieving them. This is quite a challenge when the outcomes are measured in different ways. How, for example, is a decision maker to compare cancer treatment with a psychological intervention for depression or a health promotion campaign to reduce coronary heart disease?

Address for correspondence: Professor M. Knapp, PSSRU, LSE, Houghton Street, London WC 2A 2AE (United Kingdom).
Fax: +44-(0)20-7955.6131
E-mail: M.Knapp@lse.ac.uk
Declaration of Interest: None.

HEALTH ECONOMIC EVALUATIONS - BROADENING OUT

When a new treatment is developed or introduced decision makers face two central questions. The clinical question is whether the intervention (say a new antidepressant) is more effective than existing interventions in alleviating depressive symptoms and generally improving health and quality of life. If the answer to the clinical question is that the new medication is no better than existing options, then there is usually no need to consider its use any further. But if it looks clinically effective, the decision maker will then want an answer to a second question: is it cost-effective? That is, does the treatment achieve the improved outcomes at a cost that is worth paying?

These two questions (Does the treatment work? Is it worth it?) sit at the heart of economic evaluation. Deciding what is or is not 'worth' the cost (which is the cost-effectiveness question) is rarely straightforward and sometimes can prove controversial.

Economists have developed a number of different types of evaluation. They differ primarily in the way they conceptualise and measure outcomes, because they are designed to address different issues. If the question is essentially clinical - what is the most appropriate treatment for someone with particular needs in particular circumstances (say a working-age woman with depression) - so that the task is to choose between two or more specific treatment packages within the or another - then information is needed on the comparative costs of the dif-

ferent treatments and the comparative outcomes measured in terms of symptom alleviation, improved functioning and so on. This is the most common form of economic evaluation in the health domain, and is usually referred to as a cost-effectiveness analysis (Drummond *et al.*, 2005).

However, the decision maker might have a broader set of responsibilities, needing to choose how to spend a budget that covers a range of diagnostic groups. For example, the question might be whether to invest available resources so as to introduce a new and more costly (but apparently very effective) antidepressant to treat more working-age adults or to spend more money on new (and again apparently effective) medications for women with breast cancer. The costs of the two options (the antidepressant and the breast cancer drug, plus the associated and knock-on service implications) can be measured in the same units (such as euros or dollars), but symptom-specific outcome measures for depression treatment will have limited usefulness in the cancer field, and vice versa. To overcome this difficulty economists have developed the quality-adjusted life year (QALY) measure.

THE QALY

Stripped down to the basics, a health care treatment or system aims to prolong life and to improve (health-related) life quality. The QALY aims to measure whether it succeeds. It looks at the quality of life enjoyed in each extra life year gained from better treatment. One year of perfect health is valued as 1, and death is valued at 0. Positions in between 1 and 0 describe states of health that are less than perfect. So, if a new treatment for breast cancer prolongs life by an additional five years, and each additional year is rated at 0.8 (a fairly high level of health-related quality of life) then the number of QALYs gained as a result of using this treatment rather than currently standard treatment would be 4 (= 5 x 0.8).

One strength of the QALY - so long as it is validly measured, of course - is that it allows comparisons to be made between different disease areas. If the same measure of effectiveness is being used to evaluate a breast cancer drug and an antidepressant, then the cost-effectiveness of one can also be compared with the other. This gives decision makers a chance to look across a wide expanse of options to see which would offer the greatest health improvements from available budgets. Of course, this might not be their only aim, for they might also want to ensure that there is fair access to treatment across all income groups, or that people with conditions for which there is currently no effective treatment are supported with dignity. But the pursuit of better value for money is always likely to be one of the objectives of a health system. For example, the National Institute for Health and Clinical Excellence (NICE) which has responsibility in England and Wales for providing national guidance on promoting good health and preventing and treating ill health, is a firm advocate of QALY-based technology appraisals, even (surprisingly) when there is no evidence in a particular area that the approach has any validity.

Using the QALY as the measure of effectiveness in an economic evaluation is often called a *cost-utility analysis*. The bottom line of such an evaluation is an estimate of the cost per quality-adjusted life year gained from using one intervention rather than another. But, as the old expression says, the proof of the pudding is in the eating - that is, it is results that count - and the disarmingly simple QALY is only ever going to be as good and as valid as the tools that are used to construct it.

HOW ARE QALYs MEASURED?

The QALY is a preference-based measure. A QALY combines gains from reduced morbidity (health-related quality of life gains) and reduced mortality (quantity gains) into a single measure, using a measure of the preference for observed health status. In other words, the QALY is based on the relative desirability (called 'utility' by economists) of the different outcomes. Different methods are available for estimating these utility weights to be attached to health states. The three most widely used techniques for establishing preferences (measuring utilities) are the rating scale, the standard gamble and the time trade-off. Detailed explanation of these techniques can be found in Drummond et al. (2005). Briefly:

- Rating scale or the *Visual Analogue Scale* (VAS) is a method where the respondents are asked to rate a state of ill health on a scale from 0 to 100, with 0 representing worst health state (or death) and 100 representing perfect health.
- *Standard Gamble* (SG) is the classical method of measuring utilities. This method has a firm theoretical basis as it is based on von Neumann and Morganstern (1953) theory of preference measurement. Respondents are asked to choose between remaining in a state of ill health (i) for a certain period of time (t), or choosing a medical intervention which involves a gamble: a chance of either restoring them to perfect health (probability p), or leaving them in the worst health state (or

death) (probability *1-p*). Probability *p* is varied until the respondent has no preference for one option over the other, at which point the required utility value for state i is equal to the value of *p*.

Time-Trade-Off (TTO) is an alternative method of generating preference values where the respondents are asked to choose between remaining in a state of ill health (*i*) for a period of time (*t*), or being restored to perfect health but having a shorter life expectancy of, say, *x* years which is less than *t*. Time *x* is varied until the respondent has no preference for one alternative over the other. The required preference value for the state *i* then given by *x/t*.

Since the measurement of preferences using these techniques is quite complex, most studies bypass this measurement task by using one of the pre-scored multi-attribute health status classification systems found in literature.

Two of the most commonly used tools for generating QALY measures are the Euroqol (EQ-5D), and *Health Utilities Index* (HUI).

Euroqol (EQ-5D)

The EuroQol-5D (usually called the EQ-5D) is a standardised instrument for use as a measure of health outcome and applicable to a wide range of health conditions and treatments. It is a popular preference-based non-disease-specific instrument for describing and quantifying health-related QoL (Kind, 1996). EQ-5D contains a questionnaire that is used to classify health states that can be assigned a preference or utility value using scoring weights derived from responses to the instrument made by (say) the general public. The EuroQol assesses problems on five dimensions: mobility, self-care, daily activities (work, study, housework, family, leisure), pain/discomfort and anxiety/depression. Responses on each dimension are recorded on a 3-point scale indicating levels of severity, corresponding to 'no problems', 'some problems' and 'severe/extreme problems'. Combinations of scores define a total of 243 theoretically possible health states. Using the time-trade-off method in which valuations of a UK general population survey were used (Dolan & Phil, 1997), a single health index was generated: the TTO score. A TTO score of 1.00 indicates perfect health, a score of 0 or negative values represent an evaluation of health worse than death.

In addition there is EQVAS (visual analogue scale) where patients rank their health on scale of 1-100.

Health Utility Index

The *Health Utility Index* (HUI) generates scores that can be used to adjust survival duration by reduced quality of life. HUI assesses four major concepts of HRQL: physical function which includes mobility and physical activity; role function which includes self-care and role activity; social-emotional function which includes well-being and social activity; and health problems (Horsman *et al.*, 2003). The most recent version (HUI3) contains eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain. Each of these attributes has five or six levels. Preference weights for members of the general public are available. Preferences are measured using a visual analogue scale and standard gamble instruments. Questionnaires are available in three formats: face-to-face interview, telephone interview and self-administration. HUI is a widely used and well-validated measure. However, as far as we are aware, it has not been used in studies of patients with schizophrenia.

USING QALYS IN MENTAL HEALTH RESEARCH

QALYs are now being used quite regularly in mental health studies, although - as we argue below - there are considerable challenges with both measurement and interpretation. Here we give a few examples of mental health studies using QALYs, some which worked well and some which did not.

An open, pragmatic randomised trial by Peveler *et al.* (2005) compared three treatments for depression in primary care: a tricyclic antidepressant (TCA, choosing one of three options), a selective serotonin reuptake inhibitor (SSRI, choosing again one of three options) and lofepramine. Among the outcome indicators used was the EQ-5D, alongside more familiar depression scales. The EQ-5D appeared to perform satisfactorily with this patient group. The evaluation found no significant differences between patient groups in terms of depression free weeks, but the economic evaluation suggested that, if decision makers valued each additional QALY at £5000 or more then SSRIs would likely be the most cost-effective strategy, followed by lofepramine and then TCAs. However, the probability of this occurring was relatively low, leading the authors to conclude that choice of antidepressant should be based on doctor and patient preferences, which they think would lead to less switching of medication.

A different approach was taken in an evaluation of computerised cognitive behavioural therapy (McCrone *et*

al., 2004). As originally designed, the trial used the Beck Depression Inventory as the primary outcome, and also collected data using the *Beck Anxiety Inventory* and the *Work and Social Adjustment Scale*. However, after data collection, QALYs were estimated using a method described by Lave *et al.* (1998). QALY values were obtained by cross-walking from the estimate of the number of depression-free days, with a depression-free day scored as 1, and a day with depression scored as 0.59. This is clearly quite a crude approach, but gave a reasonable indication of what QALYs might follow from the intervention. The trial in fact found that computerised CBT looked highly cost-effective by comparison to other interventions.

Sometimes the patients being treated are not able to complete ratings themselves and proxy ratings are used. Matza *et al.* (2005) tested the EQ-5D in relation to a sample of children with attention-deficit hyperactivity disorder (ADHD) in the US and UK. The parents of 126 children with ADHD participated, completing a proxy version of the EQ-5D, as well as a measure of ADHD symptoms based on standard diagnostic criteria (DSM) and a generic multi-dimensional paediatric health-related quality of life questionnaire. The weights attached to generate the utility scores were from the UK general public (societal weights). The parent-proxy version of the EQ-5D performed modestly well, with correlations with the paediatric scores ranging between 0.13 and 0.35, depending on measure and country, most of these being statistically significant. Correlations were lower in the US than in the UK.

Two schizophrenia studies illustrate how the EQ-5D can sometimes appear to perform satisfactorily and sometimes does not. The CUtLASS trial comparing atypical (second-generation) and typical (first-generation) antipsychotics used the QALY as the primary variable in the economic evaluation (Davies *et al.*, 2007). In contrast, de Willige *et al.* (2005) were not positive about the performance of the EQ-5D after assessing the sensitivity and validity of EQ-5D for measuring changes in the quality of life of patients with chronic schizophrenia. Subjective changes in quality of life were measured and compared to objective changes in psychopathology and social functioning for a sample of 76 patients from various inpatient and outpatient mental health service in the Netherlands who participated in a randomised controlled trial in which two treatment conditions were compared. Patient self-rating were used. The authors found that the weighted TTO-score of EQ-5D did not correspond with the changes in positive psychotic symptoms which appeared to be the most important factor in improving quality of life, which indicates that it is less sensitive to changes in social and psychological well-being. They concluded that the use of EQ-5D as the core measure of health state evaluation in the field of psychiatry seems less than fully convincing or appropriate.

Coucill *et al.* (2001) examined the patient and proxy ratings of the EQ-5D instrument in assessing health-related quality-of-life of 64 patients suffering from dementia with a range of severity. The visual analogue scale was used to obtain the ratings. The authors found that responses to EQ-5D questions were highly variable across the raters - patients, carers and physician. They highlighted the serious concerns regarding the validity of patient self-rated values due to variations in their cognitive ability, degree of insight and capacity to make judgements. Results from proxy ratings raised further questions as to whom the appropriate proxy should be as different groups provide different ratings.

CHALLENGES AND OPPORTUNITIES

The QALY approach offers a number of distinct advantages. It is a unidimensional measure of impact, it is generic and so allows comparisons to be made across diagnostic or clinical groups (for example, comparing psychiatry with oncology or cardiology), and the tools to generate it are brief and easy to use. It is also built on a fully explicit, transparent methodology for weighting preferences and valuing health states. Moreover, the measure itself has cardinal properties, which means it can be used in statistical tests and prediction analyses in a way that is not possible with other scales (even though some of them are misused in that way).

But some of these same features have sometimes been seen as disadvantages: the measure may be seen as too reductionist; the generic quality of life indicator may be insufficiently sensitive to the kinds of change expected in schizophrenia or other mental health treatment; and a transparent approach to scale construction paradoxically opens the approach to criticism from those who question the values thereby obtained (see early reservations of this kind expressed by Chisholm *et al.*, 1997). Of most concern, as shown by some of the uses of QALYs in mental health studies, there are worries that the generic QALY-generating instruments such as the EQ-5D and HUI are just not sufficiently sensitive to the kinds of symptoms, functioning and quality of life change important for people with mental health problems. And within the mental health spectrum, the approach has generally not been found to work well in the areas of psychosis, personality disorder, child and adolescent problems and dementia.

THE WAY FORWARD

In principle, the QALY is enormously attractive, and QALYs *as currently measured* generally work well in evaluations of treatments for people with acute physical health problems. However, their suitability elsewhere remains open to doubt, not so much at the level of principle, but in their operational measurement, for existing tools for generating QALY scores (such as the EQ-5D) have tended not to perform sufficiently well in reflecting the underlying effectiveness changes in, for example, many mental health applications.

New developmental work is needed to construct QALY measures anew for use in the mental health field, and hopefully some such work will get underway quite soon in England. Both the conceptualisation and measurement of QALYs need to be squarely built on a valid, comprehensive model of quality of life specific to schizophrenia or dementia or whatever the area, to ensure that the resultant tool is sensitive enough to pick up changes that would be expected and seen as relevant in the course of the illness.

REFERENCES

Chisholm D., Healy A. & Knapp M. (1997). QALYs and mental health care. *Social Psychiatry and Psychiatric Epidemiology* 32, 68-75.

Coucill W., Bryan S., Bentham P., Buckley A. & Laight A. (2001). EQ-5D in patients with dementia. *Medical Care* 39, 760-771.

Davies L., Lewis S., Jones P., Barnes T.R., Gaughran F., Hayhurst K., Markwick A., Lloyd H. & CUtLASS team (2007). Cost-effectiveness of first-v. second-generation antipsychotic drugs: results from a randomised controlled trial in schizophrenia responding poorly to previous therapy. *British Journal of Psychiatry* 191, 14-22.

de Willige G., Wiersma D., Nienhuis F. & Jenner J. (2005). Changes in quality of life in chronic psychiatric pateints: a comparison between EuroQol(EQ-5D) and WHOQoL. *Quality of Life Research* 14, 441-451.

Dolan P. & Phil D. (1997). Modelling valuations for EuroQol health states. *Medical Care* 35, 1095-1108.

Drummond M., Sculpher M., Torrance G., O'Brien B. & Stoddart G. (2005). *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press: Oxford.

Horsman J., Furlong W., Feeny D. & Torrance G. (2003). The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health and Quality of Life Outcomes* 1, 54.

Kind P. (1996). The EuroQol instrument: an index of HRQOL. In *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd ed. (ed. B. Spilker). Lippincott-Raven: Philadelphia, PA.

Lave J., Frank R., Schulberg H. & Kamlet M.S. (1998). Cost-effectiveness of treatment for major depression in primary care patients. *Archives of General Psychiatry* 55, 645-651.

Matza L., Secnik K., Mannix S. & Sallee F. (2005). Parent-proxy EQ-5D ratings of children with attention deficit hyperactivity disorder in the US and the UK. *Pharmacoeconomics* 23, 777-790.

McCrone P., Knapp M., Proudfoot J., Ryden C., Cavanagh K., Shapiro D.A., Ilson S., Gray J.A., Goldberg D., Mann A., Marks I., Everitt B. & Tylee A. (2004). Cost-effectiveness of computerised cognitive behavioural therapy for anxiety and depression in primary care: randomised controlled trial. *British Journal of Psychiatry* 185, 55-62.

Peveler R., Kendrick T., Buxton M. Longworth L., Baldwin D., Moore M., Chatwin J., Goddard J., Thornett A., Smith H., Campbell M. & Thompson C. (2005). A randomised controlled trial to compare the cost-effectiveness of tricyclic antidepressants, selective serotonin reuptake inhibitors and Lofepramine. *Health Technology Assessment* 9 (16), 1-134.