The use of QALY weights for QALY calculations: a review of industry submissions requesting listing on the Australian Pharmaceutical Benefits Scheme 2002 to 2004

Paul A Scuffham[1], Jennifer A Whitty[1], Andrew Mitchell[2], Rosalie Viney[3]

1. School of Medicine, Griffith University, Meadowbrook, QLD 4131, Australia

2. Pharmaceutical Benefits Division, Australian Government Department of Health and Ageing, Canberra, ACT 2601, Australia

3. Centre for Health Economics Research and Evaluation, Faculty of Business, University of Technology, Sydney, NSW 2007, Australia

Words in text: 5,832

Date: July 2007 (Revised 03 October 2007)

Running title: Use of QALYs in PBAC submissions

ACKNOWLEDGEMENTS

No specific funding was received for this study. There are no conflicts of interest. At the time the data were collected, Andrew Mitchell was the Secretary and Rosalie Viney was a member of the Economics Subcommittee of the PBAC.

Address for correspondence:

Professor Paul A Scuffham, School of Medicine, Logan Campus L03 2.43, Griffith University, Meadowbrook 4131, Queensland, Australia

Ph: +61 (0)7 338 21367

Fax: +61 (0)7 338 21338

Email: p.scuffham@griffith.edu.au

Table of Contents

ABSTRACT

Background

Quality-adjusted life years (QALYs) combine survival and health-related quality of life into a single index enabling judgements about the relative value for money of health care interventions.

Objective

The purpose of this study was to investigate the methods used for estimating QALY weights included in submissions by industry for listing on the Australian Pharmaceutical Benefits Scheme.

Study design

Retrospective descriptive review of submissions considered by the Pharmaceutical Benefits Advisory Committee (PBAC) from 2002 to 2004.

Data sources

The database of submissions considered at PBAC meetings was obtained from the Pharmaceutical Evaluation Section of the Australian Government Department of Health and Ageing. Further information on each included submission was obtained in the form of the PES commentary (expert report) on the submission.

Review methods

Submissions to the PBAC over 2002-04 presenting QALYs as an outcome measure were reviewed to identify the methods used to obtain preference-based QALY weights. Information was analysed according to the approach taken to obtain QALY weights (multi-attribute utility instrument (MAUI), health state valuation (HSV) experiment for scaling the health states, or non-preference-based approach), the population from

whom the QALY weights were obtained, the appropriateness of the population for the instrument, recommendation made by the PBAC and the main indicated category for use of the pharmaceutical.

Results

MAUIs were used in 39% of approaches reporting QALYs; the most frequently used MAUI was the EQ-5D. HSV experiments were used in 36% of the approaches and generally drawn from the published literature. Non-preference based approaches (24%) included rating scales, mapping transformations and consensus opinions. Responses from patients were used in 58% of the approaches, followed by healthcare professionals and investigators (24% and 9% respectively). Healthcare professionals and investigators' responses were frequently used in non-preference based approaches. Submissions for Nervous system, Infectious disease and Neoplasms were less likely to have presented QALY weights derived from a "more appropriate" approach. The approaches using "more appropriate" populations and techniques were associated with a 56% rejection rate of submissions by the PBAC compared with 66% of those using "less appropriate" approaches.

Conclusions

The variability in the quality of QALY weights is troubling. The PBAC Guidelines that applied over the period studied neither encouraged nor discouraged cost-utility analyses and provided only brief guidance on how QALY studies should be conducted. A consistent approach to the application of standard methods should be used when the QALY is used to inform decisions on resource allocation. The new guidelines released in 2006 provide more extensive guidance on derivation of QALY estimates and are more encouraging of the presentation of cost-utility analysis. MAUIs offer a straightforward

approach to obtaining QALY weights and ideally, should be routinely used in relevant comparative randomised trials to assess patients' health states.

INTRODUCTION

The Australian Pharmaceutical Benefits Schedule (PBS) aims to provide reliable and affordable access to a wide range of necessary medicines for Australian residents and eligible visitors through government subsidies.[1] For a pharmaceutical or vaccine to be listed on the PBS, a sponsor prepares a detailed submission according to guidelines.[2] Submissions are considered by the Pharmaceutical Benefits Advisory Committee (PBAC), an independent statutory body established to make recommendations to the Minister for Health on which pharmaceuticals should be subsidised by the Australian government. In January 1993, Australia became the first country to require an assessment of the costs and cost-effectiveness as part of the decision-making process for government subsidies of medications.[2] As such, the PBAC process is one of the most developed in the world.

The PBAC considers each submission on the basis of safety, efficacy, cost-effectiveness, and the clinical role of the new pharmaceutical relative to the comparator.[2] Sponsors of pharmaceuticals are required to provide details on the pharmaceutical (action, indications and restrictions, treatment details, main comparator, clinical management), a clinical evaluation of the main indication (including a systematic review of all relevant randomised trials), an economic evaluation and the estimated impact on government health budgets. The PBAC, with the assistance of its subcommittees, appraises each submission and makes recommendation for government funding, defers the submission, or rejects the submission. Decisions to recommend are presented to the Minister of Health for approving. Decisions to defer are those where the evidence in the submission might be accepted in principle but the price of the medication requires negotiation and/or further information is required. Following rejection, a sponsor, may resubmit the drug for

consideration at a subsequent meeting; it is expected that the sponsor might present new clinical evidence, lower the price and/or undertake major revisions to decision modelling (e.g. including providing more relevant cost and/or quality of life data). Although economic efficiency is not the sole factor considered in reimbursement decisions, the importance placed on this highlights the need for a consistent approach to economic evaluation.

One type of economic evaluation, cost-utility analysis (CUA), uses the Quality-Adjusted-Life-Year (QALY) as its outcome measure. The QALY has gained popularity as the outcome measure of choice for economic evaluations of health care interventions. Underlying the QALY measure is a common assumption that health care resources should be allocated in such a way as to maximise any health gains in society as measured by additional QALYs.[3] That is, utilitarianism weights the health gains of each individual equally to maximize any increase in the level of health of society. Therefore, all else being equal, for a given budget, programs which provide more additional QALYs should be given priority for funding over programs that produce fewer additional QALYs. Since April 2004, QALYs have been required by the National Institute for Health and Clinical Excellence (NICE) in the economic evaluation of all health care interventions being appraised.[4, 5] However, the United States Food and Drug Administration appears to be avoiding the concept of QALYs and has rebadged health-related quality of life (HR-QoL) under a banner of Patient-Reported Outcomes (PROs).[6, 7] In its latest guidelines, the PBAC has described QALYs and their use in submissions;[2] a summary is provided here:

QALYs have three key characteristics:

1. They combine survival and HR-QoL into a single measure that allows comparability across health care programmes with disparate health outcomes;[8]

2. Each health state is assigned a preference-based quality of life weight, or 'QALY weight'.*[9-11] The QALY weight indicates how much of a life year a person would be prepared to sacrifice to improve their quality of life from a given health state to full health. It is measured on a cardinal scale such that a year of life in perfect health has a score of one and death has a score of zero. Scores less than zero, to reflect states worse than death, are possible.
3. QALY weights are derived from individuals in stated preference tasks which are designed to reflect the individuals' trade-offs between quality and quantity of life.[2]

As such, there is some validity to QALYs representing social trade-offs and social values, which are desirable properties in a measure of health outcomes for making decisions around the funding of health care interventions.[3] HR-QoL scores which do not indicate strength of preference have less intrinsic validity.

To calculate QALYs is a straightforward process. The duration a person is in a health state is multiplied by the QALY weight for that health state. People may experience many health states throughout life; as such, the products of QALY weights and (expected) duration for each health state are summed to give QALYs. The two main approaches to obtaining QALY weights to calculate QALYs are:

1. Using a multi-attribute utility instrument (MAUI), and
2. Creating scenarios to elicit QALY weights through health state valuation (HSV) experiments.

MAUIs have three key elements: A generic quality-of-life instrument, a scaling technique (based on a stated preference task, such as a time trade-off (TTO) and standard gamble (SG)) reflecting trade-offs that individuals representative of society are willing to

make between health outcomes (i.e. preference-based rankings of health states), and a mathematical or statistical model that provides a scoring algorithm to generate QALY weights for any health state that can be described by the instrument. The quality of life instrument consists of a questionnaire which breaks down a health state into (multiple) attributes each of which corresponds with one or more questions and responses in the instrument. Responses are scaled and the mathematical model combines these attribute scores into an overall instrument score.

MAUIs allow health state experiences to be described by patients using a common HR-QoL instrument, in an environment that minimises bias in the reporting of that experience; MAUIs also assign "strength of preference" from the general population (i.e. a valuation) associated with the health states. In this way, MAUIs are an appropriate and practical method to obtain preference-based valuations for HR-QoL. MAUIs, along with disease-specific measures, can be used in clinical trials to collect information at baseline and at various time points during follow-up.

There are several well known MAUIs available that possess the three key elements; the most recent is the SF-6D developed by Brazier et al. to transform SF-36 or SF-12 scores into QALY weights.[12, 13] The most widely used is the EQ-5D;[14, 15] others include the Health Utilities Index (HUI2 and HUI3)[16, 17] and the Assessment of Quality of Life (AQoL).[18, 19] These MAUIs are all based on acceptable scaling techniques using the SG or TTO.[8, 20] The comparative merits of the SG and TTO techniques have been the subject of some debate; however, there appears to be no established compelling normative or empirical reason to favour the SG over the TTO method or vice versa.[8, 21] Similarly, no single MAUI has demonstrated unequivocal superiority or universal

acceptance.[2] Moreover, there is variation between scores derived from MAUIs;[18] this variation may be due to the preferences of the population valuing the health state, cultural differences and differences in the construction and domains of the instruments.

An alternative to using a MAUI is the use of HSV experiments. In HSV, stated preference experiments around health state scenarios are undertaken to elicit QALY weights. The scenarios are often hypothetical, and the respondents may be patients or drawn from a community sample. The stated preference experiments generally involve an explicit trade-off between health profiles described by HR-QoL and survival durations. The TTO and SG have been the most commonly used approaches for HSV experiments. Such experiments generally involve face-to-face interviews requiring numerous decisions by participants to identify their preferences. These can be relatively labour-intensive and sometimes cognitively challenging. Rating scales have been appealing because they are easy (and inexpensive) to use, but they do not involve a choice or a trade-off to be made between HR-QoL and survival, and therefore, the relative strength of preference cannot be inferred validly.

While there continues to be debate about methods for deriving QALYs, there is general agreement that QALYs, as a measure of health benefit, should have a sound basis in measurement of individuals' relative preferences for health states such that they capture the values of society as a whole.[22-24] Recognising the need for consistency, NICE and the U.S. Panel on Cost-Effectiveness in Health and Medicine have produced reference cases for undertaking economic evaluation for pharmaceuticals and health care interventions.[4, 25] Both reference cases suggest the use of choice-based techniques to elicit preferences, the use of generic MAUIs to describe and value health states, and the

use of a representative sample of the public as the most appropriate source for preferences. This approach is now also encouraged in the most recent version of the PBAC Guidelines (2006).[2]

This descriptive study investigates the methods used for estimating QALY weights included in submissions from industry to the PBAC for listing on the PBS prepared according to previous versions of the PBAC Guidelines which neither encouraged not discouraged cost-utility analysis. We review the submissions that have reported QALYs as a measure of health outcomes with the purpose of identifying the approach taken to obtain QALY weights, assessing whether this approach involved use of a MAUI or HSV experiments for scaling the health states, and identifying the population from whom the QALY weights were obtained. We do not assess the strengths and weakness each of of the various MAUIs or the different HSV techniques to elicit preferences (e.g. we do not compare TTO with SG). In addition, we assess whether the approach used has any bearing on the PBAC's decision to recommend or reject the submission. Of note is that submissions may include both a cost-effectiveness analysis (CEA) and CUA to support their case. Thus, the PBAC may place greater weight on the CEA if the quality of the CUA is low, or may assess the CEA results as being less uncertain than the CUA results.

METHOD

The database of submissions considered at PBAC meetings was obtained from the Pharmaceutical Evaluation Section (PES) of the Australian Government Department of Health and Ageing (DoHA). All submissions considered by the PBAC at their meetings between January 2002 and December 2004 that reported QALYs as an outcome measure were included for this analysis. Further information on each included submission was obtained in the form of the PES commentary on the submission. These commentaries are prepared by expert analysts who prepare detailed critical reports on each submission.

Included submissions were sorted according to which of the following approaches were used to obtain QALY weights:

1. Use of a MAUI
2. Use of a HSV experiment in response to health state scenarios
3. Non-preference based approaches.

We extracted data from the commentaries on the submissions focussing on the approach used to obtain QALY weights. When a MAUI was used, details of the instrument and the population completing the questionnaire were recorded. Where an approach involved HSV experiments around health state scenarios, the technique and population from whom preferences were elicited were recorded. Finally, details were noted for non-preference based approaches, including rating scales and mapping transformations. We have included mapping transformations in this category where the mapping function involved mapping a disease-specific or generic HR-QoL scale onto a 0-1 scale that did not involve trade-offs for preference elicitation (e.g. mapping SF-36 scores onto the Quality of

Well-being index [26]). Likewise, rating scales (e.g. visual analogue scales, VAS) do not require a trade-off or choice decision to be made. As such, scores from these approaches are not consistent with the QALY approach, which explicitly trades-off quality of life and survival.

Populations from whom QALY weights were obtained were categorised into:

1. Target patient group
2. Healthcare professionals
3. Investigators
4. General population.

The target patient group are living with or being treated for the disease, and includes those actively taking the study medication or those in a clinical trial. Healthcare professionals included clinicians, nurses, and allied health professionals; and the general population are a random (and potentially representative) sample of society.

Where the submission relied on data from published literature, the primary study was retrieved. Where more than one of the above approaches to estimate QALY weights was used in a submission, the approaches were analysed separately. For example, if a submission used a MAUI and a HSV experiment to derive QALY weights, both were recorded and considered independently. Similarly, when an approach was used in more than one population group (e.g. a HSV experiment was undertaken by the general population and a group of patients), each population group was recorded and considered separately.

The approach and the population were then classified as “more appropriate” and “less appropriate”. The “more appropriate” approaches were where a MAUI was administered to patients who were currently experiencing the health states being valued, or when a HSV experiment (such as a TTO or SG) was undertaken in either the general population to value a health state derived from clinical and quality of life studies or a population of patients to value their own health state. All other approaches were considered “less appropriate”. We also sorted data into recommend, reject, or defer according to the outcome from the PBAC meeting.

Because data submitted to the PBAC was bound by secrecy provisions in the *National Health Act 1953*, it is not possible to name the pharmaceuticals or indications requested in the submissions. However, we do disaggregate results according to a general category indicated for the use of each medication.

RESULTS

The PES database contains 230 unique submission numbers which were considered by the PBAC at meetings during 2002 to 2004. A total of 49 (21%) submissions were identified that included QALYs (Table I). These 49 submissions were included in this study for further consideration. There was no apparent trend in the proportion of submissions reporting QALYs, with 18%, 27% and 19% in 2002, 2003 and 2004 respectively.

The 49 submissions with QALYs consisted of new submissions (27%), resubmissions (51%)**, submissions for a change to listing (12%), and other submissions such as deferred PBAC considerations and submissions in response to pricing matters (10%) (Table I).

TABLE I ABOUT HERE

Five of the 49 submissions identified had insufficient information available to allow analysis of the approaches used to derive QALY weights and were excluded from further analysis. Of the preference source(s) and methodology used in the 44 submissions with data available, 11 submissions used more than one approach and two submissions presented four approaches. Overall, there were 66 approaches reported from 44 submissions. Of the 66 approaches used in the submissions, 47 (71%) approaches were taken from studies previously published in academic journals; the remaining 19 (29%) presented studies specifically undertaken for the purposes of preparing a cost-utility analysis to include in submissions to reimbursement agencies. Of these 66 approaches

used in the submissions, 16 (24%) resulted in a positive recommendation for the drug, 10 (15%) were deferred and 40 (61%) were rejected.

The approaches used

MAUIs and HSV experiments were the most frequently employed approaches accounting for 26 (39%) and 24 (36%) of the approaches respectively. Non-preference based approaches accounted for 16 (24%) of all approaches used. When a specific study was undertaken for the purposes of a submission, MAUIs were most frequently used (12/19; 63%) followed by non-preference approaches (5/19 26%) and two (11%) undertook a HSV experiment.

Of the MAUIs used (n=26), the EQ-5D was the most common (15/26; 58%), followed by the AQoL (6/26; 23%), and the relatively new SF-6D (3/26; 12%) (Table II). Of the HSV experiments, the TTO was used in 14/24 (58%), SG in 5/24 (21%), and both TTO and SG were used in 5/24 (21%) of the approaches. The five approaches referring to combined use of the TTO and SG all referred to one paper in the published literature;[27] however, the details of how the TTO and SG were implemented and combined was not reported in that paper.

TABLE II ABOUT HERE

Of the non-preference based approaches, mapping functions and rating scales were each employed in 7/16 (44%) approaches, with consensus opinion and an unvalidated instrument in the remaining 2/16 (12%). All mapping functions were developed from regressions to transform a quality of life measure onto a 0-1 scale. A

published algorithm was used twice[26] and two approaches developed algorithms specifically for the submissions. Use of non-preference approaches, including global quality-of-life scores and visual analogue scales, were relatively more common in specific studies compared with approaches drawn from the literature (i.e. 8/19 (42%) vs 8/47 (17%) respectively).

Submissions employing MAUIs had the lowest rejection rate in PBAC decisions (14/26; 54%) whereas 15/24 (62.5%) submissions with HSVs had rejections; non-preference approaches had the highest rejection rate (11/16; 69%) but surprisingly, 3/16 (19%) non-preference approaches were associated with positive recommendations (Table II).

The participants

Of the 66 approaches used, six sampled from more than one population group. When a MAUI was used, patients with the relevant condition completed the instrument based on their current health state in 18/26 (69%) occasions (Table III); 8/26 (31%) were completed by proxy (or a combination of patients and proxy) based on an assessment of each health state. These proxies were healthcare professionals or the investigators; in 6/26 (23%) of occasions, scenarios describing health states had been developed by the investigator and the proxy asked to complete the questionnaire. In another 2/26 (7%) of occasions, the investigator or health professional completed the questionnaire based on their impression of the patient's health state.

TABLE III ABOUT HERE

For those submissions that included a HSV experiment, preferences were elicited from patients on 12/24 (50%) occasions and from healthcare professionals on 7/24 (29%) occasions. The general population was used in 2/24 (8%) and were included in a mixed sample in a further 2/24 (8%) of occurrences. Most notable was that there was only one specific study which undertook a HSV experiment; all other HSV experiments were sourced from published studies.

In the non-preference based approaches, patients were the most common group reporting "QALY weights" (8/16; 50%) followed by investigators (4/16; 25%), healthcare professionals (3/16; 19%) and investigators plus healthcare professionals combined (1/16; 6%).

Overall, QALY weights derived using "more appropriate" approaches were used in the majority of approaches (34/66; 52%). The remainder used non-preference approaches (16/66; 24%) or "less appropriate" populations (16/66; 24%) responding to MAUIs or HSVs. Of the 34 "more appropriate" approaches, 56% were in submissions for drugs rejected by the PBAC. Rejection rates were slightly higher for submissions with HSV's (10/16; 63%) compared with MAUIs (9/18; 50%). This may reflect greater potential for uncertainty about the methods for HSV approaches.

The clinical indications

Results were disaggregated into general categories based on the indicated use; "other" was used to represent all categories with less than five submissions in the category.

Submissions for drugs to treat Neoplastic disorders tended to use a MAUI but with responses from healthcare professionals (Table IV); most of these submissions were rejected. HSVs were frequently used in submissions for Infectious diseases; however, the response population was split between healthcare professionals and patients. MAUIs and HSVs were both used in submissions for drugs to treat Mental disorders with responses more often from patients (3 HSVs and 3 MAUIs were completed by patients); the PBAC made one positive recommendation and one deferred decision (both of these involved MAUIs completed by patients) and rejected all other submissions in this category. Musculoskeletal system disorders had the greatest number of submissions over the 2002-04 period; almost half employed a MAUI and a quarter used a non-preference approach to QALYs. In this indication, most submissions used a "more appropriate" approach (14/23; 61%) of which 10/14 (71%) were in submissions rejected by the PBAC.

Half of the submissions for drugs to treat Nervous system disorders (3/6; 50%) employed MAUIs but 2/3 (67%) used responses from healthcare professionals; positive recommendations were made for the HSV with responses from patients and for a non-preference approach with responses from clinicians. Other diseases and disorders all used responses from patients; 7/10 (70%) of these were "more appropriate" approaches of which only 2/7 (29%) were in rejected submissions.

TABLE IV ABOUT HERE

DISCUSSION

The 21% of submissions to the PBAC using QALYs represents a considerable increase over the 2.5% found by George et al. in the early 1990's.[28] Given the increased support for QALYs as the primary outcome measure, particularly by the NICE,[4] and now by the PBAC,[2] it is expected this proportion will continue to rise in the future. Therefore, the comparability and appropriateness of the approaches and methodology used to derive QALY weights is an important consideration.

The findings of this study suggest that, in the recent period studied, there was little consistency in the approaches and methods used to estimate QALY weights in submissions considered by the PBAC. MAUIs and HSV experiments were most widely used, accounting for 39% and 36% of the approaches used respectively. Surprisingly, non-preference based approaches were used in almost 24% of approaches. The submissions considered in this study were prepared using the guidance in the 1995 PBAC Guidelines (and 2002 revision), which did not require CUA, and provided limited guidance on the derivation of QALYs.[29] Typically, CUA was used as a supplementary analysis to augment a base case CEA, which may have reduced the sponsor's perceived necessity to obtain high quality QALY weights as the PBAC may have primarily relied on the CEA as the analysis guiding their decision. The revised PBAC guidelines released in 2006 provided more detailed guidance on QALY approaches and include a more explicit preference for CUA in most situations (although the use of CUA is not mandatory).

Approaches reporting use of a MAUI had the lowest rate of drugs rejected by the PBAC (54%) whereas approaches using a HSV or non-preference approach were more likely to be in rejected submissions (63% and 69% respectively). When a submission used a "more appropriate" approach, 56% were rejected by the PBAC. It is emphasised here that a "more appropriate" approach does not mean the recommendation will be positive; however, a "less appropriate" approach and/or poorly selected population may increase uncertainty and be a strong contributing factor to a rejection decision. The magnitude of the cost-utility ratio and the uncertainty are only two factors considered by the PBAC in their decision-making process. Other factors include safety, efficacy, and the clinical role of the new pharmaceutical (or vaccine) relative to the comparator. Moreover, a CUA was not required by the PBAC during the study period, and only 21% of submissions included a CUA. In many cases CUA may be used as a supplementary analysis to augment a base case cost-effectiveness analysis, and therefore, the PBAC may, when there is uncertainty arising from the CUA approach, place greater emphasis on the CEA results. As such, it is possible that obtaining QALY weights or using a "more appropriate" approach was not given sufficient priority.

Of those submissions that did include a CUA, many studies to generate QALY weights were of low quality. For example, proxies such as investigators and healthcare professionals are not representative of society; likewise, non-preference based approaches cannot reflect society's preferences or be used to provide QALY weights. The inconsistencies found provide challenges for the decision-making process and increase uncertainty. To some extent, the inconsistencies and inappropriate methods may be related to a limited amount of data being available to the sponsors at the time of submission. This lack of appropriate data is due to a lack of foresight in designing clinical

trials without regard for the concurrent collection of data necessary for a good economic evaluation (i.e. use of a MAUI and resource use data). Although large multinational trials are often run from a centre in the USA or Europe, there is a need to allow and encourage the inclusion of local country-specific data in the data collection process. Without local data, additional uncertainties are introduced when transferring results to Australia. Ideally, country-specific data on resource use, costs and health outcomes would be available in adequately powered multinational trials; this would also allow differences between the aggregate and country-specific results to be assessed for generalisability. In future, it is recommended that validated, generic MAUIs (with the relevant country-specific scoring algorithm) are routinely included in all relevant comparative randomised trials.

The most appropriate use of a MAUI for obtaining QALYs involves completion of the instrument by blinded participants in a randomised trial to rate their own health at baseline and various points during follow-up using the instrument. This study found that no MAUI was used in this way. Most MAUIs were administered to (non-trial) patients with the condition under study, which raises some uncertainty in the validity of the matching of their results to the sets of patients taking the therapies being compared in the analysis.

In some uses of a MAUI, health state scenarios were described for completion of the MAUI by a healthcare professional or investigator. The use of a MAUI to generate QALY weights from scenarios is inappropriate. In an extreme case, the investigator could effectively nominate the QALY weight expected based on his or her own opinion and then construct scenarios aligned to the text of the MAUI. That is, it would be near impossible to describe a health state to a respondent without actually telling them how to rate it on the MAUI. However, when a MAUI is used appropriately such as in a group of patients in a

randomised trial, there are numerous advantages of MAUIs including more accurate and unbiased measurement of health states, comparability across studies (and internationally), and efficiency for respondents and analysts (because no MAUI takes longer than 5-8 minutes to complete and analysis is well developed).[2]

When a HSV experiment is undertaken, a choice-based technique, such as the SG or TTO, should be used to elicit preferences for the described health states. To avoid the introduction of bias and self-interest, preferences should be obtained from a representative sample of society.[2] Preferences elicited from representative samples of society may be inconsistent with those of patients who may be elderly, young, from a particular social group, and may have their preferences altered by the disease. However, preferences were rarely elicited from the general population with only 2/24 using the general population exclusively, and 2/24 using a mix of patients and the general population. Instead, subgroups of patients or healthcare professionals were commonly used for preference elicitation. This reflects difficulty in ensuring that respondents adequately understand the health states that they are requested to choose between. Furthermore, because there are less decision points for the analyst in a MAUI compared with HSV and therefore less potential for bias and uncertainty to be introduced by the methods, MAUIs are preferred to HSV.

Approaches based on mapping from quality of life questionnaires were often inappropriate and/or lacking validity, and frequently did not involve choice techniques. Of particular concern were attempts to derive QALY weights based on expert consensus or on an assumed linear transformation of a global quality of life or a rating scale to a QALY weight index. Weights estimated in this way are not preference based and do not conform

to the theory underpinning economic evaluation; as such they cannot be considered appropriate for use in QALY calculations. Best practice would ensure that any transformation function would be based on data collected concurrently for both instruments in the same population; unless this approach is taken the validity of any transformation is severely compromised and increases uncertainty. In one submission, QALY weights were transformed from the EQ-5D to the AQoL. This was unnecessary and also created greater uncertainty as the rationale for doing so was unfounded and a sensitivity analysis using a different set of weights would have provided greater elucidation of the effect of changes in those weights.

When more than one approach was used, it was often used to obtain a QALY weight for the same health state. A second approach was either the same technique applied to a different population, or a different technique used in the same population. The additional effort needed to use two approaches could be better spent by finding and using one technique with a larger relevant population. The concept of efficiency suggests that greater effort should go into obtaining greater accuracy for the main drivers, and less robust estimates may be used for those states that have little effect (such as when there is a low probability of going into that health state, or the duration in that state is short). As such, we recommend that any post-trial marginal effort in obtaining QALY weights should be directed to collecting data from patients using a MAUI or from the general population using HSV.

The 2006 PBAC Guidelines moderates the preference for obtaining QALY weights from the general population with an understanding of the information asymmetry relating to the health state being valued.[2] In the HSV context, the issues are (a) how the analyst

described the key features of the health state(s) actually experienced and (b) how the respondent understands the health state from the analyst's description before then proceeding to generate the valuation via SG or TTO. In the MAUI context, the respondent completing the questionnaire actually understands the health state, so the issues are (a) if the respondents are not in the randomised trial(s), how well the patients completing the MAUI "match" the patients receiving the alternative therapies being compared, (b) how well the questionnaire captures the elements which "drive" the health state and (c) how well the separate set of respondents in the general population understand each health state assigned when completing the SG or TTO to value the health states assigned and (d) how well the scoring algorithm aggregates these. Because of the difficulty of minimising investigator bias, there are greater concerns with HSV experiments. These factors also explain the 2006 PBAC Guidelines preference for MAUIs to be completed in the relevant double-blind randomised trials.

The limitations of the current study include the assumption that all data entered in the PES database are accurate and complete, and that the analysis is based on the PES commentaries of submissions. The database has been independently checked and validated against the commentaries as part of this study. Whilst there were instances when insufficient data were available in a commentary to analyse a submission, it is considered unlikely that either the database or commentaries would be sufficiently inaccurate to alter the findings of this study. While we may have double-counted some approaches in resubmissions where the approach used did not change, we believe this is a reasonable approach to take as sponsors have the opportunity to amend their approach to estimating QALY weights; in some cases the feedback from the PBAC on the original submission may have suggested that the approach should be changed. Despite these

limitations, the data provides a reliable trend of the approaches and methodologies used to estimate QALY weights in support of submissions.

In a previous review of submissions for inclusion on the PBS (1994 to 1997), Hill and colleagues reported finding examples of inappropriate questionnaire design and inadequate sample size in time trade-off analyses.[30] Other studies based on published CUAs have found extensive variation in the derivation of preferences.[31-33] The findings of the current study suggest there is a general improvement with increased use of generic MAUIs and a reduction in the use of rating scales. For example, in a review of 228 CUAs involving 949 health state descriptions published prior to 1997, Bell and colleagues found that 20% of health state values were derived through MAUIs.[32] This compares to the current study which found 43% of approaches used a MAUI.

In contrast to earlier studies, which found the Rosser Index, HUI (and its subsequent variants) and QWB to be the most frequently employed instruments,[31, 32] a review of 23 CUAs undertaken alongside clinical trials (based on a MAUI questionnaire completed by participants in the trial) and published between 1995 and 2002 found that the EQ-5D was the most common MAUI employed (70%) followed by the HUI (and its variants) (26%).[34] This is similar to the findings of the current study where the EQ-5D was used in 58% of all MAUIs used.

The EQ-5D is a generic, validated instrument using choice-based methods for preference elicitation, originally based on the preferences of a UK population; more recently, a scoring algorithm for the USA population has been developed.[15] There are substantial differences between the USA and UK QALY weights for the same health state.

For example, for the health state described in the EQ-5D as 11223, the QALY weights are 0.506 and 0.255 using the USA and UK preferences respectively. Significant differences in QALY weights between UK and Japan from this instrument were evident;[35] however, there were no significant differences across six European countries.[36] These differences in preferences across populations create uncertainties around the applicability of the QALY weights to populations other than from whom the preferences were drawn. QALY weights for the EQ-5D and the SF-6D with preferences elicited from the Australian population are currently being developed. This will allow greater consistency and applicability for calculating QALYs in Australia.

Because of the above problems (and others), some analysts advocate abandoning QALYs and reverting to clinical outcomes (e.g. cases, life-years etc).[37] However, this would reintroduce a lack of outcome comparability across health care programmes and pragmatically remove our ability to place a societal value on disparate health outcomes inherent with a single preference-based measure. Moreover, this is equivalent to giving a QALY weight of 1 to all life years and is likely to be an even less accurate representation of health outcomes than from the existing instruments. The problems need to be addressed and methods for eliciting preferences and the use of MAUIs need to be further developed and used appropriately.[38] Factors to consider when choosing an MAUI include evidence of its validity, reliability, responsiveness/sensitivity (both between groups of individuals and in the same individual at different points in time), and feasibility of use in the relevant health condition or population.

The source of the preferences used should also be applicable to the society in which the decision-making occurs. The preferences from representative samples of

society continue to be uncommonly elicited. We estimate responses were obtained from an appropriate population in 52% of QALY approaches submitted to the PBAC (i.e. where patients responded to a MAUI or the general population or patients responded to a HSV experiment); this is an improvement on the 24% reported by Bell and colleagues and the 22% reported by Stein from a review of NICE assessments.[32, 33]

CONCLUSIONS

The quality of QALY weights is troubling and the methodological issues described impact on the measurement of health outcomes. A consistent approach to the application of standard methods should be used when the QALY is used to inform decisions on resource allocation. The new guidelines released in 2006 provide more extensive guidance on derivation of QALY estimates and are more encouraging of the presentation of cost-utility analysis. MAUIs offer a straightforward approach to obtaining QALY weights and should be routinely used in relevant comparative randomised trials to assess patients' health state.

FOOTNOTES

* We use the term "QALY weight" to distinguish the weights used to estimate QALYs from the scoring systems commonly used in quality of life instruments, such as the SF-36. The scores for the latter are not based on any direct measurement of individual preferences, and hence there is no indication of strength of preference for different health states, or for the trade-offs individuals may make between dimensions of HR-QoL that contribute to the instrument. Elsewhere QALY weights are sometimes referred to as utilities or utility weights, and claimed to have a basis in Von Neumann Morgenstern expected utility. However, QALYs are only a measure of utility if additional strong restrictions are imposed on the utility function.[9-11]

** A resubmission is a submission that was previously rejected by the PBAC; the sponsor may choose to resubmit to provide additional information and/or make other appropriate amendments. The main reasons for rejections typically include insufficient evidence on the clinical benefit (e.g. from a different target population to that described in the evidence provided), too much residual uncertainty around the clinical benefit or cost-effectiveness estimate, or unacceptable cost-effectiveness ratio.

REFERENCES

1. Australian Government Department of Health and Ageing. Schedule of pharmaceutical benefits for approved pharmacists and medical practitioners. Canberra: Commonwealth of Australia; 2006.

2. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.1). 2006 [cited 06 Jan 2007]; Available from: www.health.gov.au/internet/wcms/publishing.nsf/Content/pbacguidelines-index

3. Bleichrodt H, Diecidue E, Quiggin J. Equity weights in the allocation of health care: the rank-dependent QALY model. Journal of Health Economics. 2004;23(1):157-71.

4. National Institute for Clinical Excellence. Guide to the methods of technology appraisal. London: National Institute for Clinical Excellence; 2004 April.

5. Gold MR, Patrick DL, Torrance GW, et al. Identifying and valuing outcomes. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, editors. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 82-134.

6. Kind P. US FDA Guidance - Apropros of PROs. Pharmacoeconomics. 2006;24(9):833-6.

7. US Department of Health and Human Services Food and Drug Administration. Draft guidance for industry of patient-reported outcome measures: use in medical product development to support labeling claims [online]. 2006 [cited 08 Dec 2006]; Available from: www.fda.gov/CDER/guidance/5460dft.pdf

8. Dolan P. Output measures and valuation in health. In: Drummond M, McGuire A, editors. Economic evaluation in health care: merging theory with practice. Oxford: Oxford University Press; 2001. p. 46-67.

9. Bleichrodt H, Quiggin J. Life-cycle preferences over consumption and health: when is cost-effectiveness analysis equivalent to cost-benefit analysis? Journal of Health Economics. 1999;18:681-708.

10. Bleichrodt H, Wakker P, Johannesson M. Characterizing QALYs by Risk Neutrality. Journal of Risk and Uncertainty. 1997;15:107-14.

11. Pliskin J, Shephard D, Weinstein M. Utility functions for life years and health status. Operations Research. 1980;28:206-24.

12. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002;21:271-92.

13. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004 Sep;42(9):851-9.

14. Kind P, Dolan P, Gudex C, et al. Variations in population health status: results from a United Kingdom national questionnaire survey. Brit Med J. 1998;316(7133):736-41.

15. Shaw R, Johnson J, Coons SJ. US valuation of the EQ-5D health states - development and testing the D1 valuation model. Medical Care. 2005;43(3):203-20.

16. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002 Feb;40(2):113-28.

17. Torrance GW, Feeny DH, Furlong WJ, et al. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care. 1996 Jul;34(7):702-22.

18. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Annals of Medicine. 2001;33:358-70.

19. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of Health Related Quality of Life. Quality of Life Research. 1999;8:209-24.

20. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life. Health Economics. 1999;8:41-51.

21. Bleichrodt H. A New Explanation for the Difference between Time Trade-Off Utilities and Standard Gamble Utilities. Health Economics. 2002 July;11(5):447-56.

22. Dolan P. The measurement of health related quality of life for use in resource allocation decisions in health care. In: Culyer A, Newhouse J, editors. Handbook of Health Economics. Amsterdam: Elsevier; 2000.

23. Torrance G. Measurement of health state utilities for economic appraisal: a review. Journal of Health Economics. 1986;5(1):1-30.

24. Torrance G, Furlong W, Feeny D, et al. Multi-attribute preference functions: Health Utilities Index. PharmacoEconomics. 1995;7:503-20.

25. Gold MR, Siegel JE, Russell LB, et al., editors. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press; 1996.

26. Fryback DG, Lawrence WF, Martin PA, et al. Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study. Medical Decision Making. 1997;17:1-9.

27. Wong J, Bennett W, Koff R, et al. Pretreatment evaluation of chronic hepatitis C: Risks, benefits, and costs. Journal of the American Medical Association. 1998;280(24):2088-93.

28. George B, Harris A, Mitchell A. Cost-effectiveness analysis and the consistency of decision making: Evidence from pharmaceutical reimbursement in Australia (1991 to 1996). Pharmacoeconomics. 2001;19(11):1103-9.

29. Scuffham P. The assessment of pharmaceuticals for government subsidy in Australia: recent developments. Journal of Medical Economics. 2007;10:163-69.

30. Hill SR, Mitchell AS, Henry DA. Problems with the interpretation of pharmacoeconomic analyses, A review of submissions to the Australian Pharmaceutical Benefits Scheme. JAMA. 2000 Apr 26;283(16):2116-21.

31. Neumann PJ, Zinner DE, Wright JC. Are methods for estimating QALYs in cost-effectiveness analyses improving? Medical Decision Making. 1997;17:402-8.

32. Bell CM, Chapman RH, Stone PW, et al. An off-the-shelf help list: a comprehensive catalog of preference scores from published cost-utility analysis. Medical Decision Making. 2001;21:288-94.

33. Stein K, Fry A, Round A, et al. What Value Health? A review of health state values used in early technology assessments for NICE. Applied Health Economics and Health Policy. 2006;4(4):219-28.

34. Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. Health Economics. 2004;13:1203-10.

35. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: The case of Japan. Health Economics. 2002;11(4):341-53.

36. Greiner W, Weijnen T, Nieuwenhuizen M, et al. A single European currency for EQ-5D health states: Results from a six-country study. European Journal of Health Economics. 2003;4:222-31.

37. McGregor M, Caro J. QALYs Are they helpful to decision makers? Pharmacoeconomics. 2006;24(10):947-52.

38. Vijan S. Should we Abandon QALYs as a Resource Allocation Tool? Pharmacoeconomics. 2006;24(10):953-4.