Retaining, and Enhancing, the QALY

Joseph Lipscomb, PhD,[1] Michael Drummond, PhD,[2] Dennis Fryback, PhD,[3] Marthe Gold, MD, MPH,[4] Dennis Revicki, PhD[5]

[1]Department of Health Policy and Management, Rollins School of Public Health, Emory University, Atlanta, GA, USA; [2]University of York, York, UK; [3]University of Wisconsin, Madison, WI, USA; [4]City University of New York Medical School, New York City, NY, USA; [5]United BioSource Corp, Bethesda, MD, USA

Keywords: quality-adjusted life-years, health-related quality of life, cost-utility analysis, outcome assessment (health care), health status.

Introduction and Overview

For economic and policy analyses requiring a summary health outcome measure that integrates quantity-of-life *and* quality-of-life impacts, the quality-adjusted life-year (QALY) has been widely recognized and endorsed [1–3].

The US Panel on Cost-Effectiveness in Health and Medicine recommended the QALY for "reference case" analyses [2]. The UK National Institute for Health and Clinical Excellence (NICE) regularly uses QALYs in its technology appraisals [3]. Regulatory and purchasing agencies in several other European nations, Canada, and Australia likewise recognize the use of QALYs in analyses of whether medical products and interventions offer good value for money [4]. The enthusiasm with which analysts from academia and contract research organizations embrace the QALY in economic evaluations is evident from a quick review of the leading health policy evaluation journals or the roster of presentations at any recent ISPOR Annual International Meeting [5].

As a point of reference for much of the discussion that follows, it is useful to define at the outset what we term the conventional QALY:

$$QALY_{\text{conv}} = \sum\nolimits_{t=1\ldots T} \sum\nolimits_{s=1\ldots S} p_{\text{st}} V(H_{\text{st}})(1+r)^{t-1}$$

where p_{st} is the probability an individual will occupy health state H_{s} at time t; $V(H_{\text{st}})$ is the value (or preference) measure assigned to the individual being in state H_{s} at time t; $(1 + r)^{t-1}$ is a discount factor designed to bring $V(H_{\text{st}})$ to present value terms, with r being the selected discount rate reflecting time preference for health outcomes; S is the number of discrete health states that may be occupied; T is the time horizon relevant for decision making; and $QALY_{\text{conv}}$ is subscripted to indicate this is (some variant of) the "conventional" QALY formulation. Note that when QALYs are used to inform group-level (frequently, societal-level) decisions, the $V(.)$ are typically intended to reflect the average (usually mean) preference levels of the individuals comprising the group.

As will be seen, most of the QALY debates—and most of the opportunities for enhancing this summary outcome measure—center on either the conceptualization and construction of the health states (the H_{st}), the valuation of these states (V), or the adequacy of $QALY_{\text{conv}}$ as a guide to decision-making when ethical, distributional, or other considerations not factored into this formulation are deemed to be compelling. To be sure, there are important challenges in marshalling and analyzing the clinical and epidemiological data required for statistically sound estimates of the p_{st} parameters, but the focus of this article is on health state definition, valuation, and distributional issues (sometimes referred to as the "Q" part of the QALY setup).

The QALY thus formulated is one of the two defining ingredients in cost-utility analyses of health interventions, where the aim is to identify the candidate intervention that generates incremental gains in QALYs at the lowest economic cost (or, equivalently, the greatest incremental gain in QALYs per dollar spent) [1]. Moreover, variants of the H_{st} component of $QALY_{\text{conv}}$ are now being used selectively to track population health status at the national level [6,7] and even within some communities [8].

But should this coin of the realm really be regarded as the gold standard? In fact, several related concerns voiced over time about the QALY have their own currency.

Concern #1: There Remains Considerable Diversity of Viewpoint in the Research Community Regarding Several Key Technical and Methods Issues Pertaining to QALY$_{conv}$

Such issues, which must be addressed in any QALY application, include: 1) selection of the specific domains, or dimensions, for health viewed as a multidimensional concept; 2) selection of the specific survey items that give operational meaning to each domain; 3) psychometric approaches to eliciting individual preferences for health states; and 4) statistical modeling strategies for deriving the value weights required for computing QALY scores. In the $QALY_{\text{conv}}$ equation, issues 1 and 2 essentially concern specification of the H_{st}, and issues 3 and 4 relate to the determination of the $V(.)$ and how they "work in concert" to map a given pattern of health states for the individual to a QALY score.

Differences in how issues 1–4 are handled in practice can be seen in the major preference-based, health-related quality of life (HRQOL) measurement systems frequently used for QALY construction. These measurement systems include the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3) [9], EuroQol EQ-5D [10], Quality of Well-Being Index (QWB) [11], SF-6D [12], and the Health and Activities Limitations index (HALex) [13]. As Fryback et al. show [7], these six systems yield "similar but not identical trends" in HRQOL for older US adults. Consequently, if the systems were brought to bear concurrently in a given economic evaluation, they will yield different QALY scores and thus possibly different conclusions about the cost-utility of the intervention of interest. In practice, it has

Address correspondence to: Joseph Lipscomb, Emory University, 1518 Clifton Road, NE, Room 642, Atlanta, GA 30322, USA. E-mail: jlipsco@sph.emory.edu

10.1111/j.1524-4733.2009.00518.x

Joseph Lipscomb, Michael Drummond, Dennis Fryback, Marthe Gold, and Dennis Revicki have no conflicts to declare.

proven easier to recommend use of the QALY as a useful general approach to assessing the health payoffs of interventions than to specify *which* QALY variant should be adopted as the standard.

Concern #2: The Conventional QALY Model Assumes Away Certain Important Issues, and Ignores Others

Arguably, such critiques sort out into three general categories, largely related to health state valuation issues. First, even proponents of the conventional QALY acknowledge that there are simplifying assumptions built into $QALY_{conv}$ that, at the very least, bear close scrutiny [14]. Second, it has been argued that $QALY_{conv}$ fails to incorporate certain fairness and distributional concerns that are important in group decision-making (e.g., in a cost-utility analysis (CUA)) [15,16]. Third, the conventional QALY model and its variants all assume that the value-component of the model should be *preference-based*, although sometimes differing on whose preferences are to be captured and how. By contrast, Dolan and Kahneman [17] and Hausman [18] challenge (though for different reasons) the contention that such health state valuations should be based on an individual's stated preferences. These issues are pursued briefly later in the article.

Concern #3: Conceptual and Methodologic Issues May Be Interfering with the Uptake of QALYs (and Cost-Effectiveness Analysis More Broadly) in the Decision-Making Contexts That They Are Intended to Serve

The QALY has played, and continues to play, an important role in regulatory and purchasing decisions in a number of jurisdictions outside the United States. Moreover, the US Office of Management and Budget (OMB) now requires that federal agencies supplement their cost-benefit analyses with cost-effectiveness analyses (CEAs) using a generalized QALY variant (the health-adjusted life-year) for "economically significant health and safety decisions" [19]. But in general, cost-effectiveness and QALYs have not been widely embraced in US health-care decision-making (see Neumann [20]). Of particular note, the US Food and Drug Administration has not encouraged preference-based variants of patient-reported outcome measures in connection with applications for drug approval [21]. The US Centers for Medicare & Medicaid Services still does not formally apply CEAs in coverage or reimbursement decisions [22]. Although QALY-based CEAs may provide a "subtle lever on policy discourses" [20], it is not clear how much of this reluctance, especially in the United States, is attributable to methodological doubts and how much to the historic and enduring political difficulties associated with the explicit rationing of health care.

Given the long-standing challenges surrounding the conventional QALY model, some may conclude it is time to deemphasize, or even abandon, this summary measure in favor of alternative approaches to valuing the health impact of interventions. We believe this would be unwise, for reasons discussed later. This is not to say the conventional QALY should be embraced uncritically as some static, unassailable gold standard. Rather, we believe that ongoing research and experimentation to improve QALYs—and health outcomes assessment generally—should build on a firm understanding of the strengths and limitations of current approaches. The most prominent—and salient—of those current approaches is the conventional QALY.

The $QALY_{conv}$ construct has proved to be a serviceable vehicle for quantifying mortality and morbidity jointly over time at both the individual and population level. Over the past three decades, there has been a large investment of public and private resources to develop, apply, and also evaluate the performance of preference-based HRQOL measures, including the six measurement systems noted earlier. In a comprehensive assessment of the performance of the EuroQol EQ-5D, QWB, and HUI 2/3 in cancer applications, Feeny [23] concluded that all three systems have generated convincing evidence of reliability, construct validity, interpretability, responsiveness, and feasibility. In the United States, a federally funded registry of published cost-utility analyses continues to chart the progress over time in the degree to which studies adhere to the recommendations of the US Panel on Cost-Effectiveness in Health and Medicine [24]. An Institute of Medicine Committee charged with providing guidance to OMB on the measurement of health benefit for regulatory decisions endorsed what we have termed the conventional QALY as "the best measure at present on which to standardize health-adjusted life-year calculations because of its widespread use, flexibility, and relative simplicity" [20].

HRQOL measures that can support the application of specific QALY models are now part of various national data sets in Canada [6], the United States [7], and several other countries [25]. Moreover, there is emerging evidence that the public at large is increasingly willing to consider including cost-effectiveness in health-care decision-making. A recent study by Bryan et al. [26] suggests that among a sample of decision-makers from public and private insurers, health plans, and employer groups in California, there is a profound understanding of the need to set priorities in coverage decisions. Focus group participants did express concerns about the potential for bias in industry-sponsored studies, and worried about litigation for organizations that were in the forefront of using CEAs for decisions-making. Still, 90% of discussants believed that CEAs and QALYs were relevant to informing coverage decisions and, further, that Medicare should take the lead in moving this agenda forward.

Over the course of hundreds of published applications of the conventional QALY model, much has been learned about the empirical methodological strengths and limitations of this preference-based approach to evaluation. To abandon the QALY model at this juncture is also to sever the link to hundreds of published studies and numerous ongoing investigations—and thus to render difficult, if not impossible, our ability to judge the extent to which alternative health measures "improve upon" $QALY_{conv}$. Likewise, the capability to do historical comparisons across economic evaluations or population health assessments would be seriously compromised.

A more productive course is to pursue a program of research that takes the conventional QALY model as a starting point for efforts to address the concerns raised earlier.

In the sections that follow, we build on this discussion to identify a number of topics that merit continuing examination. Regarding the QALY itself, there are three general issues: health state definition and description, valuation of health states, and approaches to augmenting the QALY to incorporate additional information on social equity and distributional considerations. Without question, there are also some prominent alternatives to a QALY-based approach for valuing health outcomes, but each presents its own methodological and practical challenges. Hence, we conclude that there is a need for more "comparative effectiveness" research—by which we mean here, research that critically compares alternative approaches to measuring and valuing health outcomes in service to decision-making.

Health State Definition and Description

As they have developed and evolved over time, the major health status classification systems noted earlier have played a useful role in the economic evaluation of health-care interventions and the assessment of population health.

These measurement systems are the "engines" behind the preference-based effectiveness measures in many (though by no means all) cost-utility analyses published in recent years. That the systems have been used to compute traditional QALY scores for numerous economic evaluations across a wide spectrum of diseases and interventions does not, of course, mean that further improvements in the scales should not be pursued. Indeed, as Brauer et al. [24] document, the majority of published studies using utility weighting over the 1976 to 2001 period did *not* use one of the generic health measurement systems, but relied rather on preference weights generated expressly for the study at hand. Yet on balance, these health measurement systems continue to pass a kind of market test in that many analysts (and some policymakers) are willing to rely on them for public and private sector economic evaluations.

In addition, the instrument items from several of these measurement systems have been embedded in major national health surveys, providing a means to derive preference-based measures for population health over time. Prominent examples include:

1. US Medical Expenditure Panel Survey [27], which in the past has included both the EuroQol EQ-5D and the SF-12. Currently only the SF-12 items are asked, but recently analyses provide prediction models for mapping SF-12 responses to EQ-5D preference scores [28–30].
2. Both the Joint Canada–US Survey of Health [31] and the Canadian National Population Health Survey [32], which include the HUI3.
3. US Centers for Medicare and Medicaid Services Health Outcomes Survey [33], which has been using the SF-36v1 and is now adopting the Veterans Affairs SF-12, which can support the derivation of (preference-based) SF-6D scores, as well as be mapped potentially to the EQ-5D.
4. In the UK over the past 15 years or so, government agencies have periodically conducted population surveys that include the EQ-5D to inform decision-making, although individual researchers have also carried out national-level studies assessing population health status, e.g., Kind's application of the EQ-5D [34].
5. In the United States, a federally supported project to generate US-based preference weights for the EQ-5D has led to a series of population-based studies comparing the resulting preference scores with those obtained from the EQ-5D with UK-based weights and with scores from the HUI2 and HUI3 [35].
6. Fryback and colleagues [7] published representative US preference scores for six health measurement systems (EQ-5D, SF-36, HUI2, HUI3, QWB (self-administered version), and the HALex), as part of the federally supported National Health Measurement Study.
7. The Patient Reported Outcome Measurement Information System (PROMIS) Network included the EQ-5D index in this large project aimed at developing item banks for assessing pain, fatigue, physical functioning, emotional distress, and social functioning. Prediction models have recently been developed using PROMIS global items and domain item banks for estimating health preference scores [36].

To the extent these health state "descriptive systems" are radically altered or abandoned in pursuit of a wholly reconceptualized QALY or other evaluation metric, our ability to link past, current, and future analyses is jeopardized and possibly destroyed. On the other hand, if incremental changes are made to these systems with due attention to such linkages, we will retain the capacity to compare findings over time.

Are there aspects of health state definition and description that merit further attention by researchers? The following issues continue to provoke discussion:

> *The major health state measurement systems have notably different domain structures, some of which seem to be "getting at the same thing," though not exactly in the same way. This likely has adverse implications for the comparability of health status (and QALY) calculations across measurement systems.*

To illustrate, the domains for four well-known measurement systems are labeled (by their developers) as follows: EQ-5D (Mobility, Self-Care, Usual Activities, Pain/Discomfort, Anxiety/Depression); HUI 3 (Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition, Pain); QWB (Mobility, Physical Activity, Social Activity, a set of Symptom-Problem Complexes); and SF-6D (Physical Functioning, Role Limitations, Social Functioning, Pain, Mental Health, Vitality). Fryback et al. [7] confirm that when these measurement systems are applied to the same population, they yield significantly different estimates of health status. Hence, the systems will neither, in general, produce the same QALY estimates nor the same cost-utility ratios when applied to a given sample.

On the other hand, the variety of options does provide the analyst with the opportunity to tailor the choice of instrument to the particular health problem being analyzed. Indeed, it is difficult to render convincing a priori judgments about what constitutes the "correct" (that is, necessary and sufficient) set of health domains in the absence of some kind of external criterion, such as the desire to match the domain structure of the instrument with the dimensions of health most likely affected by the intervention. Such judgments will also likely depend on the scope and nature of the application, e.g., whether the focus is on interventions for a specific disease, or on multiple interventions across several diseases.

By contrast, there are well-defined psychometric techniques for appraising and improving the item content along each selected domain—a good thing, in light of the next issue.

There is a tendency of health status instruments to suffer from ceiling and floor effects and, in general, to have item content that is "too sparse" to provide adequate coverage along the full continuum of outcomes associated with each heath domain. To the extent such problems exist, they have direct implications for QALY scores.

This matter of adequate item content has been recognized by several instrument developers. For example, the evolution from the SF-36v1 to SF-36v2 involved a number of changes [37], including enriching the item content in some places (inserting a five-item response choice in place of a dichotomous choice for seven items in the two role function scales) and simplifying the item content elsewhere (moving from a six-item to a five-item response category set for the Mental Health and Vitality scales). The HUI3 was developed to address particular concerns about the HUI2, with both items and dimensions being altered to improve applicability to both clinical and population-level studies and to enhance the structural independence of domains (which improves the validity of scoring algorithms) [38]. In an application of both instrument variants to diabetes, Maddigan et al. [39] find that the HUI3 yields a greater range of possible health status scores and better discriminates between individuals according to their clinically reported degree of impairment.

Most recently, Pickard et al. [40] have provided a striking example of how to use modern psychometric techniques to assess whether incremental changes in the structure of a health status instrument improve the measurement properties of the instrument. Specifically, they examine the impact of moving from the standard three-level EQ-5D scaling format to a five-level format, for each of the instrument's five domains. Applying Rasch modeling—a variant of item response theory (IRT) modeling—to Dutch and US samples, Pickard et al. are able to identify the item levels on the 5-L instrument and 3-L instrument that are statistically equivalent (thus securing the link between the two instruments) and also to confirm that the 5-L instrument broadens the coverage of the underlying health continuum of each domain.

The successful incremental modifications of the SF-36, HUI, and EQ-5D (still in progress) strongly suggest that the health status measurement component of $QALY_{\text{conv}}$ can evolve in ways that improve the scientific properties of the instruments although maintaining the ability to link the older original instruments to newer versions of these instruments. Likewise, this linking using IRT or other techniques allows the old and new measures to be placed on the same measurement metric.

A particularly promising avenue for pursuing such improvement is IRT modeling. The application by Pickard et al. [40] provides insights into what can be accomplished by exposing the candidate items for a given health domain to rigorous psychometric scrutiny using modern measurement approaches. For each scale in the multidimensional set of scales comprising the health measurement system, one can test jointly for ceiling and floor effects, for whether the items are sufficient in number and in content to adequately "cover" the underlying domain continuum, and for whether the scale itself is (sufficiently) unidimensional, as implicitly posited in these measurement systems. In concept, the domain-specific scales for each of the current measurement systems could undergo such examination.

In the future, such IRT-based analyses might also provide a pathway for creating new, stronger item sets for each posited health domain, just as is being done now for nonpreference-based health status assessment in the NIH-sponsored project to develop the PROMIS [41]. Once the item content for each and every domain has been calibrated, one would proceed to obtain representative preference values for the items, the scales, and ultimately the composite health index using standard preference assessment approaches. Revicki et al [36] has developed a prediction equation that can use either PROMIS global items or selected domain banks for estimating EQ-5D index scores. A recent application of IRT modeling to cross-walk health status scores across preference-based instruments is discussed in the next section.

Valuation of Health States

Each of the major health state measurement systems has a scoring process for mapping its particular multidimensional characterizations of health status to scalar summary scores, essentially the $V(H_{\text{st}})$ in the $QALY_{\text{conv}}$ equation. But the scoring processes differ among the measurement systems in potentially important ways. Moreover, certain aspects of health state valuation methodology common to all the systems have been challenged, as noted in the first section.

In our view, these health state valuation issues can be most effectively addressed by taking the conventional QALY as the starting point for further investigation. Exploratory investigations can be conducted either to 1) better understand and cope with differences among the current health state systems; or 2) pursue incremental—or perhaps more than incremental—improvements in the QALY while still preserving the capability to link to the large literature of $QALY_{\text{conv}}$ applications. In this way, we gain valuable perspective regarding the impact of any changes in QALY valuation procedures on such bottom-line matters as cost-utility ratios and assessments of population health status.

Recognizing and Handling Differences within the Conventional QALY Model

Valuation issues relating to $V(H_{\text{st}})$ in the conventional QALY model that merit further inquiry pertain both to the measurement and aggregation of the weights *within* any given health state system and to the examination and possible reconciliation of differences in summary scores *across* health state systems. We note these issues in rapid succession now, recognizing that each could merit its own article-length discussion.

Alternative methods for eliciting preferences. Values for the QWB were obtained through rating scale procedures [11]; for the EQ-5D *via* the time-trade-off (TTO) approach [10]; for the HUI2/3 using, alternatively, both the standard gamble (SG) and visual analog scale (VAS) procedures [9]; and for the SF-6D, using the SG [12]. The time perspective used in the preference measurement questions varies significantly; for example, respondents for the QWB were asked to imagine each candidate health state enduring for 1 day, while respondents to the EQ-5D (both the UK and US samples) were working within a 10-year time frame when valuing state through the TTO technique. All measurement systems base their preference weights on samples of respondents drawn from the general community (as opposed, say, to subpopulations with particular illnesses or disabilities); but the specific communities and the eras for data collection vary markedly [1,9–12].

Alternative approaches for deriving an aggregate score for a (multidimensional) health state. In the HUI2/3, the aggregate (point-in-time) health status score for an individual with an assigned position along each of the system's health dimensions is derived through multi-attribute utility theory (MAUT) modeling, alternatively using both multiplicative and multilinear functional forms. For the EQ-5D (whether based on the UK or US samples), the QWB, and the SF-6D, aggregate scores are derived through econometric modeling: sample-rendered health state scores are regressed against health state levels (attributes) to develop models for predicting the composite score associated with any observed combination of health state levels. Petrillo and Cairns [42] provide a useful summary of the methodological issues arising with the different approaches to deriving aggregate health preference scores.

States worse than death. Importantly, the major systems differ on whether they recognize and assign values to states judged to be worse than death (or, more precisely, being dead). Both the EQ-5D and HUI2/3 are constructed to permit such negatively valued health states (with death still anchored at 0), while neither the QWB nor the SF-6D (nor the HALex) do this. Clearly, some difficult philosophical and even moral issues arise here. The more immediate point is that the major measurement systems can reach quite different conclusions, from a societal perspective, about the health status of individuals residing in especially serious or dire conditions.

These cross-system differences in valuation procedures, coupled with the cross-system differences in health state definition and construction noted in the previous section, mean the systems will generally assign different summary scalars (different values of $QALY_{\text{conv}}$) in a given application, whether a CUA or an assessment of population health status.

There are at least three responses to this state of affairs.

First, one could applaud the multiplicity of instrumentation and valuation approaches as shedding multiple points of light on the genuinely tough problem of *quality-adjusting* life-years. There is also the option, readily available in principle, of conducting sensitivity analyses to determine whether substituting one measurement system for another significantly impacts the CUA or population health calculation. Some recent examples (out of many that could be cited) include the study by Franks et al. [43] on the impact of choice of measurement system on incremental cost-effectiveness calculations for a given health problem and also across problems; by Fryback et al. [7] examining how calculated population health status varies across measurement systems; by Janssen et al. [44] comparing the EQ-5D and HUI2/3 on the basis of information-theoretic measures of performance in the same sample; and by Stevens et al. [45] investigating differences in the predictive validity of the HUI2 depending on whether state level scores are aggregated by MAUT or regression modeling.

Second, one can attempt to cross-walk the scores from one measurement system to another, or (more expansively) from each measurement system to all others. In essence, this involves mapping the $V(H_{\text{st}})$ from one system (e.g., the SF-6D) to another system (e.g., EQ-5D). To the extent this can be done successfully, the multiplicity of QALY scores emerging from the various measurement systems can be "reconciled" and placed on the same metric, or at least compared directly on identical samples of respondents. Cross-walking may prove to be a very practical, constructive response to the current reality of multiple competing health status measurement systems, and some recent articles and conference presentations suggest how it might proceed.

In cross-walking scores, there would appear to be at least two general avenues of attack. The most straightforward approach is to work within one or more "training samples" to develop a statistical mapping relationship (through correlational or regression analysis) between the health status scores for any and all pairs of competing instruments. Ideally, one would check the predictive validity of these statistical models in validation samples. Using data from their National Health Measurement Study, Fryback et al. [46] demonstrate the feasibility of this approach, producing predictive models to support pair-wise comparisons of five preference-based indexes: EQ-5D, HUI2, HUI3, QWB-SA, and SF-6D. Note also that published work establishing predictive relationships between SF-12 and EQ-5D scores [28–30] *could* support similar cross-walk analysis (in two steps), because SF-6D scores can be imputed directly from the SF-12. The second promising approach to cross-walking, developed and presented initially by Fryback et al. [46], uses hierarchical IRT modeling jointly involving all five of the indexes to establish statistical linkages that facilitate pair-wise instrument score cross-walks. In essence, an individual's score on some index (say the HUI3) can be mapped onto the underlying IRT continuum for multidimensional health status scores (θ, in common IRT terminology), and a predicted score on some other index of interest (say the SF-6D) corresponding to *that* particular value of θ can be inferred directly.

A third general response to the multiplicity of health status instruments supporting $QALY_{\text{conv}}$ is to initiate some form of consensus process to identify, among candidate health status measurement systems, a "reference case" health status measure (in support of a reference case QALY). This would be in the spirit of the approach adopted by the US Panel on Cost-Effectiveness in Health and Medicine to establish a standard set of methodological practices in CEA to promote improvement in the technical quality and comparability of studies [2]. It is beyond the scope of this article to consider the benefits, costs, and feasibility of such a strategy. But any such deliberations should be informed by a comprehensive, dispassionate evaluation of measurement systems, with assessment criteria well specified in advance. Possible criteria, which have been applied successfully in the evaluation of nonpreference-based HRQOL measures in cancer and elsewhere, include those published by the Medical Outcomes Trust [47,48].

Exploring Additional Issues in Health State Valuation

Much of the discussion and critique of the traditional QALY model has focused on issues of health state valuation, broadly construed, and we highlight two of those issues as follows.

First, even proponents of the conventional QALY acknowledge there are certain simplifying operational assumptions built into $QALY_{\text{conv}}$ that, at the very least, require further analysis. For example, it is assumed that the value associated with being in health state s' for 2 years is twice the value of being in s' for 1 year (sometimes called the constant quantity effect)—save for the application of an exponential discount factor to bring these value computations to present value. One prominent, much debated formal response to such concerns about whether preferences at the individual level are being properly reflected in the conventional QALY calculus is the healthy-year equivalent approach (see Mehrez and Gafni [14]). Recently, Salamon and Murray [49] developed and applied a multimethod approach to analyze jointly and compare the major preference elicitation approaches (SG, TTO, VAS, and the person trade-off (PTO)) in terms of their ability to yield health state weights while accounting for such influences as risk attitude, time preference, and distributional issues.

In general, additional empirical evidence is needed on whether individual preferences over health profiles are well approximated by the time-preference-adjusted sum of preferences for health state components, as posited in $QALY_{\text{conv}}$. Despite the technical and cognitive challenges in such holistic assessments of health outcomes, there are potentially important payoffs. These experimental analyses could shed light on the reasonableness of assumptions about preferences over time and health states built into $QALY_{\text{conv}}$, e.g., the constant quantity effect noted previously, which implies the marginal utility of health state occupancy is a constant and independent of length of stay.

To ensure that holistic valuations of such profiles can be compared with valuations based on $QALY_{\text{conv}}$, the health states comprising the building blocks of each profile could be drawn from the states as defined in one of the current health state systems (e.g., from the HUI, EQ-5D, QWB). Valid comparisons of the holistic and $QALY_{\text{conv}}$ approaches require that they be applied to profiles with the same overall time frame. To see how such comparison can be carried out in practice, using health states drawn from the QWB, see Lipscomb [50].

A second general critique is that all of the approaches discussed so far assume that the value-component of the model should be *preference-based*, though differing in whose preferences are captured how. Practitioners of the conventional QALY model generally argue that the $V(H_{st})$ should be community-based, reflecting the ex ante preferences of a representative sample of the general population. For "saved-young-life-equivalents" (SAVEs), Nord has argued that preferences should be derived from individuals who have experienced the health states at issue (not simply random community samples).

Nevertheless, Dolan and Kahneman [16] and Hausman [17] have challenged the contention that such individual-level valuations should be preference-based. Dolan and Kahneman argue that whether captured from an ex ante or ex post perspective, individual preferences will tend to yield biased estimates of the value the individual attaches to a health state *at the moment the individual is experiencing the state*. From their perspective, the appropriate analytical task is to capture the value of these moments, not as they are contemplated or remembered, but as they are experienced (hence, their emphasis on "experience utility" rather than "decision utility"). Hausman, on the other hand, rejects any of these approaches to valuing health in terms of its contribution to "well-being." Instead, he advocates the development of strategies for health resource allocation that would optimize the opportunity for individuals to pursue productive and rewarding lives. In essence, health is viewed as an intermediary for the capacity of the individual to pursue life's important projects, while the QALY model essentially serves to capture the individual's health-related well-being.

For these and other important critiques of the conventional QALY, the clear and present challenge is to develop alternative operational models that, in the end, can inform societal decision making in a more useful and compelling way that the conventional model.

At the same time, work on these issues should proceed in a way that permits the resulting modified QALY models—whatever form they may take—to be linked back to the conventional QALY model. This would facilitate subsequent sensitivity analyses to pinpoint how such changes in the health state valuation process might influence the construction of health measures for CEA, for monitoring population health status, and for other practical applications. In this way, we become positioned to judge what difference a difference in the QALY might make.

Addressing Equity and Distributional Considerations

There is a broad consensus that equity and distributional considerations matter (or should matter) in health resource allocation decisions, and that the conventional QALY model does not explicitly incorporate such considerations in CEAs or other forms of economic evaluation. Nevertheless, there are divergent views about the best way to address this issue analytically, with at least two possible pathways forward.

One approach is to modify the preference weighting scheme used in the QALY model so that the value weights—rather than reflecting community-wide, average values for the states of health—are allowed to vary according to the characteristics of individuals who might occupy the states. Thus, $V(HS_{st})$ in the conventional QALY model would now be written as $V(H_{stx})$, where x represents certain characteristics of the individual. Specifically, Nord has taken issue with the $QALY_{conv}$ assumption that the *societal* value of a health outcome for an individual is proportional to the size of the value-gain for the individual (that is, $\Delta QALY_{conv}$)—irrespective of the severity of the individual's initial condition, age, or other factors. In response, he has advocated the PTO method as a way to elicit value weights that take such considerations into account, generating SAVEs as an alternative coin of the realm [15,51]. Ubel et al. [16] discuss approaches to basing cost-effectiveness measurement on directly solicited societal values rather than QALYs.

An alternative approach is to incorporate fairness concerns into the resource allocation decision process in a hierarchical fashion: equity enhancements (or constraints) are applied to the conventional QALY model in a second, clearly definable step, rather than being woven into model's preference weights. This affords the opportunity to determine the impact on efficiency (that is, the total quantity of QALYs produced) and equity (their distribution across the relevant population) attendant to a given resource allocation decision. It allows one to appraise the efficiency–equity trade-offs associated with the pursuit of specific, operationally defined rules of fairness or just distribution. Inputs for the prioritization of equity considerations could be gathered through community-based deliberative processes as described at the end of this section.

We believe there is much merit in this second general approach because it can be seen as a natural extension or augmentation of the conventional QALY model. One is positioned to appraise the equity gains, as well as calculate the possible efficiency (pure QALY) losses, associated with moving from $QALY_{conv}$ to some alternative formulation that accommodates distributional concerns. Hence, the opportunity cost (in forgone aggregate health improvement) of seeking greater fairness can be assessed.

How might such a hierarchical approach building directly on the conventional QALY model be pursued analytically? At least two broad strategies exist.

Equity Weighting

Among the several contributions in this area is the recent work by Bleichrodt et al. [52] to develop "rank-dependent" QALY models that allow the application of equity weights to each possible QALY profile that may be experienced by the assumed N members of society. Such a profile here is an ordered vector, from highest to lowest, of $QALY_{conv}$ scores for these N individuals, conditional on interventions and other assumptions. Hence, the health-related outcomes anticipated from any two competing interventions would be compared from a societal perspective in terms of equity-weighted QALYS. Special cases and potential variants of this formulation include "QALY utilitarianism" (what we have when the conventional QALY model is applied in a standard CEA); a Rawlsian social welfare function that assigns all equity weight to the worst off individual; and Williams "fair innings" approach to priority setting.

Wagstaff's has proposed to measure the *degree* of society's aversion to health inequalities through an estimable parameter that would indicate the rate with which society is willing to forgo $QALY_{conv}$ to achieve certain equity gains [53].

In response to claims by some that conventional QALY maximization models discriminate against the disabled and chronically ill, Johannesson [54] has discussed an alternative formulation in which the relative change in QALYs, rather an absolute change, is maximized in the conduct of a CEA. The "relative change" parameter for a patient group of some age and sex is computed as the average expected QALYs for the population of that age and sex, divided by the average expected QALYs for the patient group at baseline. Two patient groups with the same relative change parameter would have equal equity weighting in a CEA, irrespective of the absolute changes in QALYs expected from the interventions.

Note also that Nord et al. [51] use a multistage approach involving equity weights (specifically, weights reflecting the relative severity of disease and the relative potential for health improvement) elicited from the general population by the PTO technique. But the value weights assigned to health states in this "cost-value analysis" formulation (see also [15]) are to come via the time trade-off valuations from individuals who have experienced these states, not from the general population. Hence, and perhaps by design, there is not a clear link back to the conventional QALY model.

Constrained Optimization Modeling

It is well known that any CEA can be reformulated as a mathematical programming problem in which one solves for the intervention set that maximizes improvement in health status, subject to a budget constraint. When the objective function is specified in terms of $QALY_{conv}$, we end up with a linear or integer programming problem that corresponds precisely to the standard cost-utility model. What has also been recognized periodically through the years is that one can bring additional constraints into the programming setup to reflect equity and distributional considerations. For example: If subpopulation B has an expected QALY score at baseline of less than *X*, assign intervention *Y* to B, regardless of the expected gain in QALYs. Indeed, one has the flexibility to impose literally any set of equity rules that can be translated into the algebraic language of a model constraint. (And one might contend that if an equity rule cannot be stated unambiguously in algebraic terms, perhaps the rule itself is ambiguous.)

Studies that have examined the implications of this programming approach to building fairness considerations into the analytical framework for CEA include Epstein et al. [55], Stinnett and Paltiel [56], and Chen and Bush [57]. In all these applications, the conventional QALY is essentially preserved, so that one can readily calculate the difference between the maximum QALY improvement attainable when equity constraints are not applied and the maximum attainable corresponding to specific constraints. This appears to be an alternative and comparatively practical way to derive Wagstaff's proposed equity–efficiency frontier [53].

Finally, a less formulaic and more interactive approach to identifying equity considerations could be pursued through community-based *deliberative processes* where the implications of simple CUA ratios could be examined for their fidelity to the consensual preferences of the members of the public. Such processes would bring together individuals (e.g., general citizens, or perhaps designated community representatives) to work with facilitators who would present the relevant health program and outcome data, encourage dialogue, and guide the individuals toward recommendation(s) that, in principle, could be either qualitative or quantitative in nature. One prominent example of such activity at the national level is the UK NICE Citizens Council, a 30-member group comprising a broad cross section of the public. The Council's deliberations, focusing on factors that should influence the distribution of QALYs within the National Health Service, are drawn upon to inform the decisions of NICE's appraisals committees [58,59].

Concluding Remarks

The overarching theme of this article—to retain, and enhance, the conventional QALY model—can be viewed as a means to a larger set of ends. These include improving our ability to measure and value the health of individuals and populations, and to evaluate the health impact of competing interventions.

We have discussed in some detail the elements of a research agenda to enhance the scientific soundness and usefulness of the QALY approach by improving 1) health state descriptive systems; 2) valuation methods; and 3) the capacity to account for equity and distributional concerns. In proceeding, there are important scientific and policy-related reasons to treat the conventional QALY model as a point-of-departure for the development and testing of alternative preference-based models—whether the proposed changes to $QALY_{conv}$ are incremental or more substantial. Doing so capitalizes on what has been learned across many years and research studies about the science and art of measuring *and* valuing health outcomes. Doing so also serves to maintain continuity and promote comparability in the tracking of trends in population health and in CEAs to identify interventions that offer good value for money.

Clearly, there are alternative approaches to valuing health outcomes and the health benefits of interventions, including cost-benefit analysis based on willingness-to-pay (WTP) assessments and discrete choice experiments (DCE) using conjoint analysis to establish a ranking of interventions. As research and experimentation continue in these areas, investigators will not only be wrestling with their own methodological and empirical challenges, but may (or should) want to have benchmarks for comparisons. The conventional QALY model is a natural benchmark, allowing one to determine whether, in a given resource allocation problem, a WTP-based cost-benefit analysis, a DCE evaluation, and a QALY-based CEA yield quite similar or quite different recommendations.

For multiple reasons, we should retain the QALY, enhance it, and work more broadly to improve the scientific soundness and usefulness of preference-based approaches to health measurement.

Source of financial support: Funding for the ISPOR "Building a Pragmatic Road: Moving the QALY Forward" Consensus Development Workshop was made possible in part by grant 1R13 HS016841-01 from the Agency for Healthcare Research and Quality. The views expressed in written workshop materials or publications and by speakers and moderators neither necessarily reflect the official policies of the Department of Health and Human Services; nor does mention of trade names, commercial practices, or organizations imply endorsement by the US government. Funding for this *Value in Health* special issue, "Moving the QALY Forward: Building a Pragmatic Road" was made possible in part by Contract No. HHSN261200800148P from the National Cancer Institute.

References

1 Drummond M, Sculpher TG, O'Brien BS. Methods for the Economic Evaluation of Health Care Programmes (3rd ed.). New York: Oxford University Press, 2005.

2 Gold M, Siegel R, Weinstein M. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press, 1996.

3 National Institute for Health and Clinical Excellence (NICE). Available from: http://www.nice.org.uk [Accessed January 24, 2009].

4 O'Donnell J, Pham S, Pashos C, Miller D. Health technology assessment: Lessons learned from around the world. Value Health 2009;12(Suppl.) in press.

5 ISPOR. ISPOR Thirteenth Annual International Meeting Abstracts. Value Health 2008;3:A1–311.

6 Statistics Canada. Canadian Community Health Survey. September 2000 and Ongoing. Available from: http://www.statcan.gc.ca [Accessed January 24, 2009].

7 Fryback DG, Dunham NC, Palta M, et al. Norms for six generic health-related quality-of-life indexes from the national health measurement study. Med Care 2007;45:1162–70.

8 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. Med Decis Making 1997;17:1–9.

9 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;40:113–28.

10 Brooks R, Rabin R, de Charro F. The Measurement and Valuation of Health Status Using EQ-5D: a European Perspective. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003.

11 Andresen EM, Rothenberg BM, Kaplan RM. Performance of a self-administered mail version of the Quality of Well-Being (QWB-SA) Questionnaire among older adults. Med Care 1998; 36:1349–60.

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

13 Erickson P. Evaluation of a population-based measure of quality of life: the Health and Activity Limitations Index (HALex). Qual Life Res 1998;7:101–14.

14 Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy years equivalents. Med Decis Making 1989;9:142–9.

15 Nord E. Cost-Value Analysis in Health Care: Making Sense out of QALYs. New York: Oxford University Press, 1999.

16 Ubel P, Nord E, Gold M, et al. Improving value measurement in cost-effectiveness analysis. Med Care 2000;38:892–901.

17 Dolan P, Kahneman D. Interpretations of utility and their implications for the valuation of health. Econ J 2008;118:215–34.

18 Hausman DM. Valuing health properly. Health Econ Policy Law 2008;3:79–83.

19 Miller W, Robinson LA, Lawrence RS, eds. Valuing Health for Regulatory Cost-Effectiveness Analysis. Washington DC: Institute of Medicine, National Academy Press, 2006.

20 Neumann PJ. Using Cost-Effectiveness Analysis to Improve Health Care. New York: Oxford University Press, 2005.

21 US Food and Drug Administration. Guidance for industry—patient-reported outcome measures: use in medical product development to support labeling claims. 2006. Available from: http://www.fda.gov/cber/gdlns/prolbl.pdf [Accessed January 24, 2009].

22 U.S. Federal Registry. Medicare program: criteria and procedures for extending coverage decisions that relate to health care technology. US Federal Register 1989;54:4302–18.

23 Feeny D. The roles of preference-based measures in support of cancer research and policy. In: Lipscomb J, Gotay CC, Snyder C, eds. Outcomes Assessment in Cancer: Measures, Methods, Applications. Cambridge: Cambridge University Press, 2005.

24 Brauer CA, Rosen AB, Greenberg D, Neumann PJ. Trends in the measurement of health utilities in published cost-utility analyses. Value Health 2006;9:213–8.

25 Szende A, Williams A, eds. Measuring Self-Reported Population Health: an International Perspective Based on the EQ-5D. Budapest: SpringMed Publishing, 2004.

26 Bryan S, Sofaer S, Siegelberg T, Gold MR. Has the time come for CEA in U.S. Health Care? J Health Econ Policy Law in press.

27 US Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey (MEPS). Available from: http://www.meps.ahrq.gov [Accessed January 24, 2009].

28 Sullivan PW, Ghushchyan V. Mapping the EQ-5D Index from the SF-12: U.S. general population preferences in a nationally representative sample. Med Decis Making 2006;26:401–9.

29 Lawrence WF, Fleishman JA. Predicting EuroQol EQ-5D preference scores from the SF-12 Health Survey in a nationally representative sample. Med Decis Making 2004;24:160–9.

30 Franks P, Lubetkin EI, Gold MR, et al. Mapping the SF-12 to the EuroQol EQ-5D Index in a national U.S. sample. Med Decis Making 2004;24:247–54.

31 Statistics Canada and the US Centers for Disease Control and Prevention. Joint Canada–United States Survey of Health (JCUSH). Available from: http://www.cdc.gov/nchs/about/major/nhis/Canada_US.htm [Accessed January 24, 2009].

32 Statistics Canada. National Population Health Survey. Available at: http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=82-618-M [Accessed January 24, 2009].

33 US Centers for Medicare and Medicaid Services. Medicare Health Outcomes Survey. Available from: http://www.cms.hhs.gov/hos/ [Accessed January 24, 2009].

34 Kind P, Dolan P, Gudex C, Williams A. Variations in population health: results from a United Kingdom national questionnaire survey. BMJ 1998;316:736–41.

35 Shaw JW, Johnson JA, Coons SJ. U.S. Valuation of the EQ-5D Health States: development and testing of the D1 valuation model. Medical Care 2005;43:203–20.

36 Revicki DA, Kawata A, Harnam N, et al. Predicting EuroQol (EQ-5D) scores from the Patient Reported Outcomes Measurement Information System (PROMIS) global items and domain item banks in a United States representative sample. UnitedBioSource Corp. working Paper, November 2008.

37 Ware JE Jr. SF-36® Health Survey Update. Available from: http://www.sf-36.org/tools/sf36.shtml [Accessed January 24, 2009].

38 Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties, and applications. Health Qual Life Outcomes 2003;1:54–66.

39 Maddigan SL, Feeny DH, Johnson JA, for the DOVE Investigators. A comparison of the Health Utilities Indices Mark 2 and Mark 3 in type 2 diabetes. Med Decis Making 2003;23:489–501.

40 Pickard A, Kohlmann T, Janssen M, et al. Evaluating equivalency between response systems: application of the Rasch Model to 3-Level and 5-Level EQ-5D. Med Care 2007;45:812–9.

41 National Institutes of Health. Patient-reported outcomes measurement information system: dynamic tools to measure health outcomes from the patient perspective. Available from: http://www.nihpromis.org [Accessed January 24, 2009].

42 Petrillo J, Cairns J. Converting condition-specific measures into preference-based outcomes for use in economic evaluation. Exp Rev Pharmacoeconom Res 2008;8:453–6.

43 Franks P, Hanmer J, Fryback DG. Relative disutilities of 47 risk factors and conditions assessed with seven preference-based health status measures in a national U.S. sample: toward consistency in cost-effectiveness analysis. Med Care 2006;44:478–85.

44 Janssen MF, Birnie E, Bonsel GJ. Evaluating the discriminatory power of EQ-5D, HUI2 and HUI3 in a U.S. general population survey using Shannon's indices. Qual Life Res 2007;16:895–904.

45 Stevens K, McCabe C, Brazier J, Roberts J. Multi-attribute utility functions or statistical inference models: a comparison of health state valuation models using the HUI2 health state classification system. J Health Econ 2006;26:992–1002.

46 Fryback D, Palta M, Cherepanov D, et al. for the Health Measurement Research Group. Cross-walks among five self-reported summary health utility indexes: progress and prospects. Presented at the Annual Meeting of the Society for Medical Making, Pittsburgh, PA, Oct 24, 2007.

47 Scientific Advisory Committee for the Medical Outcomes Trust (Lohr K, et al.) Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.

48 Lipscomb J, Snyder CF, Gotay C. Cancer outcomes measurement through the lens of the medical outcomes trust framework. Qual Life Res 2007;16:143–64.

49 Salomon JA, Murray CJ. A multi-method approach to measuring health-state valuations. Health Econ 2006;13:281–90.

50 Lipscomb J. Time preference for health in cost-effectiveness analysis. Med Care 1989;27:S233–53.

51 Nord E, Pinto JL, Richardson J, et al. Societal concerns for fairness in numerical valuations of health programmes. Health Econ 1999;8:25–39.

52 Bleichrodt H, Doctor J, Stolk E. A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. J Health Econ 2005;24:655–78.

53 Wagstaff A. QALYs and the equity-efficiency trade-off. J Health Econ 1991;10:21–41.

54 Johannesson M. Should we aggregate relative or absolute changes in QALYs? Health Econ 2001;10:573–7.

55 Epstein DM, Chalabi Z, Claxton K, Sculpher M. Efficiency, equity, and budgetary priorities. Med Decis Making 2007;27: 128–37.

56 Stinnett AA, Paltiel AD. Mathematical programming for the efficient allocation of health care resources. J Health Econ 1996;15: 641–53.

57 Chen MM, Bush JW. Maximizing health system output with political and administrative constraints using mathematical programming. Inquiry 1977;13:215–27.

58 NICE Citizens Council. NICE Citizens Council report: ultra orphan drugs. London, November 2004. Available from: http://www.nice.org.uk/ [Accessed January 24, 2009]. [Alternatively or together, cite: Culyer AJ. NICE's Use of Cost-Effectiveness as an Exemplar of a Deliberative Process. Health Econ Policy Law 2006;1:299–318. and/or a critique of the process.].

59 Gold MR, Sofaer S, Siegelberg T. Medicare and cost-effectiveness analysis: time to ask the taxpayer. Health Aff (Millwood) 2007;26:1399–406.