INNOVATIONS in pharmacy

Volume 7 | Number 2 Article 14

7-8-2016

Great Expectations: Cost-Utility Models as Decision Criteria

Paul C. Langley
University of Minnesota, langley@maimonresearch.com

Great Expectations: Cost-Utility Models as Decision Criteria

Cover Page Footnote
None

Great Expectations: Cost-Utility Models as Decision Criteria

Paul C Langley, PhD
College of Pharmacy, University of Minnesota, Minneapolis MN

Abstract
One of the more puzzling features of published claims for cost-effectiveness is the popularity of claims presented in terms of quality adjusted life years (QALYs). Despite the popularity of QALYs as the 'gold standard' outcome measures among academic audiences, professional groups and a number of single payer health care systems, there is no evidence to suggest that cost-per-QALY based claims have ever been assessed, either through experimentation or observation, to support formulary decisions. In part this stems from the fact that cost-per-QALY claims are typically not expressed in evaluable terms; it also stems from the fact that, despite the plethora of QALY publications, QALYs are not collected on a regular basis by any health care system as part of administrative claims or electronic medical records. In the US QALYs have typically been ignored by health care decision makers. Given this, the continuing popularity of utility-based measures for studies published in the leading pharmacoeconomics journals is difficult to understand. One possible explanation is that those promoting QALY claims are locked into a relativist position that defends the publication of nontestable product claims. A position that is reinforced by recommendations from 'peer organizations' such as the Academy of Managed Care Pharmacy (AMCP) in their promotion of their Format for Formulary Submission standards which support the role of lifetime cost-per-QALY modeled imaginary worlds or thought experiments. Another explanation is that QALYs have been taken at face value with little though given to how they might be implemented to support both initial formulary decisions as well as ongoing disease area therapeutic class reviews. The purpose of this review is to put the case that the continued emphasis on cost-per-QALY claims has no practical benefit in formulary decision making.

Keywords: cost-effectiveness modeling, pseudoscience, credibility, imaginary worlds, scientific method

Introduction

Belief in quality adjusted life tears (QALYs) as a gold standard in modeled or simulated health technology assessment claims is widespread. Support for the QALY is found in academic groups, particularly in the US and UK, together with professional associations such as the International Society for PharmacoEconomics and Outcomes Research (ISPOR) with their recommendations for best practice modeling [1 2 3]. QALYs have also been the focus for a number of single payer health system technology assessment groups such as the National Institute for Health and Care Excellence (NICE) in the UK, with its mandatory reference case model, the Pharmaceutical Management Agency (PHARMAC) in New Zealand, the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia, the Canadian Agency for Drugs and Technology in Health (CADTH) and the Health Quality and Information Authority (HQIA) in Ireland [4 5 6 7 8]. QALYs are also a recommended endpoint in the Academy of Managed Care Pharmacy (AMCP) *Format for Formulary Submissions (Version 4.0)* [9].

Corresponding author: Paul C Langley, PhD
Adjunct Professor
College of Pharmacy, University of Minnesota
Tel: 520-577-0436
Email: langley@maimonresearch.com
Web:www.maimonresearch.com

The popularity of the QALY as an endpoint, at least for academic audiences, can be seen in the frequency with which QALY endpoints appear in published claims for cost-effectiveness in the leading technology assessment journals: *Value in Health*, *Pharmacoconomics* and the *Journal of Medical Economics.* In the US, as in a number of countries, there is what may be described as a 'life sciences industry' (a.k.a consultants), who have developed a substantial revenue stream from developing clinical cost-effective models as part of submissions for formulary evaluation. In single payer systems the submissions follow the guidelines set out by assessment agencies such as NICE in the UK and the PBAC in Australia. In the US there is a well-developed process for the development and processing of submissions following the standards recommended in the AMCP *Format*. In addition, in the US, there are independently conducted cost-effectiveness analyses and recommendations for product acceptance, such as those produced by the Institute for Clinical and Economic Review (ICER) which has modeled itself on the NICE scoping, modeling and assessment process [10].

As far as can be ascertained, given the confidentiality of submissions, the 'life sciences industry' is wedded to QALYs as the primary outcome in their cost-effectiveness models with a substantial proportion of submission adopting a lifetime cost-per-QALY framework to generate untestable claims for cost-effectiveness, prioritization and pricing based on evidence constructed within imaginary worlds.

At the same time, there appears little interest in QALYs and lifetime cost-per-QALY claims from decision makers outside the few single payer health systems which have mandated QALY standards for formulary submission. There is little evidence on which to judge whether the recipients of formulary submissions, in particular in the US, actually pay much attention to claims expressed in cost-per-QALY terms. While this may be put down to the lack of skills possessed by the myriad of formulary committees in place with the various health systems, with the number growing significantly following the implementation of the Patient Protection and Affordable Care Act from 2010, the more reasonable explanation is that the dossier submitted is seen simply as a possibly useful information source for assessing the clinical merits of competing products. The time and effort put into Markov process and event simulation models appears to be largely ignored. This is more likely not due to the 'lack of skills' on the part of recipients, but the more mundane explanation that the recipient sees no tangible or practical value in the modeled claims.

It is important to note, however, that a number of the QALY instruments, notably the instruments in the SF-36 'family' have been used extensively in clinical practice, estimating the burden of disease and monitoring health outcomes in specific disease states [11] . To this extent there is clearly scope for the application of the various instruments. It is not the purpose here to assess the value of the various QALY instruments in these applications. Rather, the purpose of this review is to consider why modeled QALY-based claims for cost-effectiveness, whether they are submitted to formulary committees in the US, or to agencies acting as gatekeepers in single payer systems, are probably irrelevant to health care decisions. This assessment focuses on two questions: (i) do modeled cost-per-QALY claims met the standards of normal science and, (ii) how do we account for their continuing popularity in published claims for the comparative cost effectiveness of pharmaceutical products and devices? In respect of the first question the focus is whether the construction of a modeled cost-utility claim represents an empirically evaluable coherent theory and whether the modeled claim facilitates the testing of hypotheses through experimentation or observation. In respect of the second question the focus is on the analytical and empirical standards accepted for the publication of cost-utility models and claims made for competing products.

It is worth noting that the terms QALY or health related quality of life (HRQOL) are used somewhat indiscriminately in the literature. In the present case the focus is on the generic QALY measures such as the EQ-5D rather than on disease specific instruments that have been labeled as HRQoL measures [12]. Putting these measures to one side does not mean, of course, that a focus on disease specific measures as appropriate metrics for patient quality of life may not be more appropriate in a clinical setting for judging the merits of competing interventions on disease progression and disease staging compared to 'generic' measures.

The Promise of the QALY

Belief in the QALY is well established. To believers, QALYs allow: (i) comparisons across a broad range of programs and disease states; (ii) decision makers to evaluate the opportunity costs of new programs; (iii) a disparate number of outcomes measures that may characterize a disease intervention to be collapsed into a single summary measure that incorporates simultaneously mortality and morbidity; and (iv) the user to attach values to outcomes with the more important outcomes more heavily weighted [13].

The technical case for the QALY rests on the measurement of utility. In economics utility theories have an established provenance. The von Neumann-Morgenstern prescriptive axioms of cardinal utility theory have attracted a substantial literature over the past 60 years on the credibility of the axioms, alternative preference measures, the impact of uncertainty and the merits of scaling versus choice response methods [14]. Aficionados of utility measurement have a choice between direct measures of health states such as the variants of the rating scale, standard gamble and time trade off as opposed to a number of multi-attribute health status classification systems with preference scores. The latter including the Quality of Well-being Scale (QWB), the various incarnations of the Health Utilities Index (HUI), the EuroQol measures (EQ-5D) and constructs drawn from the SF family of instruments (e.g., SF-6D) [15] [16].

Although there has been a general acceptance of the QALY methodology, doubts have been raised over the years concerning the accuracy and reliability of health preferences [17]. Most recently, the ECHOUTCOME project, conducted in four European countries (the UK, France, Belgium and Italy) focused on testing the validity of four key hypotheses that form the basis of multiattribute utility theory. A total of 1,361 subjects were asked to express their preferences regarding hypothetical health states. The study found that observed and calculated utility values derived from combining different health states and time durations were significantly different. The authors concluded that as the expressed preferences were not consistent with the underlying QALY assumptions, the QALY multiplicative model was an invalid measure [18]. In response NICE called the study 'limited' and maintained the measure used in its evaluations (the EQ-5D), although imperfect, was 'the best measure anyone has yet devised for this' [19].

Towards a QALY Consensus?

In 2009 ISPOR held a Development Workshop on "Moving the QALY forward: Building a Pragmatic Road". Following the presentation of a number of papers ranging from concerns and detailing challenges to the QALY metric to the benefits (or otherwise) of QALYs as a resource allocation tool, Drummond et al reported on the deliberations of a workshop consensus group to identify common ground on key issues [20]. Putting on one side the questions of how QALYs should be measured or which instruments should be used, the committee decided that the best way forward was to reach agreement on high level principles and issues for further research.

The consensus issues were:

- QALYs are only one health-based input to health and health-care decisions
- QALYs can be used at various levels in the health care system
- QALYs do not encompass all aspects of well-being
- Both ex ante preferences for health states and experience of health states may count
- Distributive issues need to be addressed using QALYs in some settings (e.g., the US)
- Developing a reference method of estimating QALYs

In conclusion, it was noted that it was not possible to agree on a single way forward yet the majority opinion was that for the QALY to gain wider acceptance by decision makers then efforts had to be made to develop a reference method for estimating QALYs.

If this remote possibility is to occur then we need to agree on the choice of QALY. This is a virtually impossible task: In single payer systems such as the UK, New Zealand and Ireland, the choice is straightforward: the instrument is mandated by the assessment agency (which in these cases comes down to the EQ-5D or a local variant). If the choice is between the various multi-attribute instruments the potential user must be aware that the individual instrument constructs are quite different. First, they vary in the number of health dimensions included, the number of levels captured in each health dimension and their severity. Second, they differ in the population surveyed to elicit preferences for health states and in how the preference score is derived. Finally, they vary in how the preference data were translated into a preference score. Not surprisingly, similar patient groups can generate quite different scores depending on the instrument used.

There is little guidance as to which instrument should be considered most appropriate in target disease states for specific populations. While there is evidence that some measures 'travel well' in addressing different populations, at least within countries such as the US, considerable effort has been put into developing 'country specific' measures where the local population is surveyed to elicit local preferences for health states.

The result is that we now have a number of instruments to choose from. Each instrument has its supporters who argue for its relevance and its ability to capture clinically meaningful differences in health state interventions and quality adjusted life years saved. Considerable efforts have also gone into crosswalking instruments with little apparent success or agreement on the appropriate techniques.

Since the release of the Drummond et al consensus report little appears to have been accomplished, at least from the ISPOR perspective, on constructing 'the pragmatic best road forward'. There has been little effort towards agreeing on the need for, let alone the construction, of a reference method for QALYs. The obstacles are formidable (and understandable). First, there is no agreement on whether or not it is even possible to agree on a reference standard given the diversity and incompatibility of the various QALY measures; a situation that is no different to the plethora of patient reported outcomes instruments that characterize various disease states. Second, it is doubtful, even within a health care system whether agreement could be reached on the choice of QALY instrument and the preferences for the defined health states. Third, unless a QALY metric is established as a process or outcomes measure for quality assessment, there is little chance that any health care system would invest resources in capturing QALYs from a specific instrument on a regular basis. Fourth, the chance that a QALY quality metric would be mandated in the US is effectively zero. This is made abundantly clear in the *Affordable Care and Patient Protection Act (2010)* which requires that the Patient Centered Outcomes Research Group (PCORI) exclude discounted cost-per-QALY or similar discounting measures and threshold values for priority setting in health care, although the Act does not define a QALY [21]. Although QALYS, as Adler points out, have been used by many administrative agencies in the US in making regulatory decisions, the ACA objection is that it disadvantages people with disabilities [22]. Irrespective of the merits, and potential for ambiguity, in the ACA case, the point worth noting is that nowhere in the ACA is the issue of measurement and the potential for evaluating QALY-based claims raised. Finally, there is the issue of the anchoring on the QALY in death vs. perfect health. For the majority of disease states death seems a remote possibility with the relevant metric the clinical improvement toward 'perfect' health.

Imaginary Worlds and the Reference Case

As issue that was not addressed in the ISPOR Development Workshop, and continues to be put to one side at the present is the status of modeled claims, in particular those involving

cost-per-QALY comparative effectiveness claims. In the last 10 to 15 years there have been hundreds of modeled claims published which present constructed claims for the relative merits of one product over another. The question which has been ignored is whether the particular QALY-based claims meet the standard of normal science: are they testable and are they replicable.

If the NICE reference case is taken as the exemplar reference case then there is little if any chance that claims based on a reference case would meet these standards [23]. The NICE reference case is now well established and has been applied to a substantial number of pharmaceutical products, typically through single technology assessments (STAs). Submissions from manufacturers are required to meet reference case standards in modeling or simulating the natural course of a disease in chronic disease states. This involves a modeling framework, with Markov processes the most popular, projecting direct medical costs and QALY outcomes over the lifetime of the patient (unless an earlier resolution can be justified). Both costs and QALYs are discounted with the results presented in incremental lifetime cost outcomes terms (ICERs). The models presented are defended in terms of their input assumptions and their core mechanism, with account taken of uncertainty in the input parameters and the choice of modeled health states. The submitted model is then reviewed by an Evidence Review Group (ERG), often university based, who can accept, modify or create their own reference model. The final 'good housekeeping' seal of approval for the modeled claim is given by the NICE Advisory Group.

One interpretation of the NICE commitment to constructing reference models is that while they have taken on board the arguments of academic advisors and accepted the central place of cost-per-QALY claims and willingness-to-pay thresholds as driving resource allocation decisions, they also recognize that it is not only impractical, but inadvisable to actually attempt to validate cost-per-QALY claims. The reference case is designed to support this dichotomy. The truth is constructed. The process of assessing the modeled claims of manufacturers through the ERG process and the final seal of approval by the Advisory Committee are designed to demonstrate the 'scientific' foundation for decision making yet at the same time putting to one side any future attempts to evaluate and challenge claims.

The time horizons recommended for cost-per-QALY modeled or simulated claims place a further barrier to claims evaluation. While lifetime QALY models may have a visceral appeal to model builders and decision makers who have attended ISPOR workshops and other presentations supported by manufacturers, the fact is that such claims are imaginary. The lifetime cost-per-QALY requirement excludes the possibility of generating evaluable predictions, let alone trying to replicate those claims. Indeed, there is no evidence to suggest that it was ever seriously intended that the reference case would support evaluable predictions. The reference case is best seen, therefore, as a blueprint for constructing evidence in the framework of an imaginary world.

From the decision perspective the reference case gives the impression of a well thought through and scientifically justified approach to the complex issue of formulary decisions. The truth, as defined by the reference case, is constructed. The credibility of the claims rest on the correspondence of the model or simulation with 'reality' as initially perceived by the manufacturer, modified or challenged by the contracted ERG to generate their 'reality' and finally reviewed by the NICE Advisory Board to give their 'final' version of reality. At no stage is there any consideration given to evaluable claims. This would, quite obviously, be impossible given the focus of the reference case on the natural course of a disease and the discounted cost-per-QALY requirements. Although the preferred utility instrument is the EQ-5D, there is no evidence to suggest an interest in actually collecting EQ-5D preferences from target populations to potentially evaluate claims as opposed to filling in evidence gaps in constructing the imaginary worlds.

This does not mean that a modified, short-term version of a modified reference case type model could not generate evaluable cost-per-QALY claims in, for example, end of life therapy interventions. The point is that NICE is not interested in evaluable claims. The reference case is not intended to set the stage for generating new evidence but to provide a framework for price negotiations. Manufacturers are quite aware of the reference case as a rite of passage pricing model. If they can provide a believable modeled or simulated claim for a unit price that generates sub-threshold cost-per-QALY estimates then they have a good chance of a successful price negotiation. If they are unsuccessful then they can negotiate alternative pricing and discount arrangements. At no stage are either party charged with actually evaluating claims; they have no independent check or interest in knowing whether the clinical or cost-utility claims are right or even if they are wrong.

Imaginary Worlds, QALYS and Pseudoscience

Against this background, it might not come as a surprise to raise the question of whether the construction of modeled or simulated reference-type claims for cost-effectiveness should be best described as a pseudoscience, or at most a 'soft science' as opposed to a 'hard' science. If we consider the difference between 'hard' and 'soft' science the distinction is blurred. Science, as Pigliucci points out, is a heterogeneous group of disciplines where the degree of precision in

predicting events can vary from highly controlled experiments in the 'queen of sciences' physics to non-laboratory assessments where the ability to explain and predict is much reduced [24] . However, the common thread linking these disparate environments is the ability (or the commitment) to 'produce and test hypotheses based on systematically collected empirical data [via experiments or observation]'.

Pseudoscience is distinguished from the continuum of disciplines that fit into the 'hard' to 'soft' categorization of science by the absence of the three core elements of 'science': (i) an investigation of nature; (ii) the construction of empirically verifiable theories; and (iii) testing through observation of experimentation. If we consider the NICE reference case against these standards then it could be argued that while the question of the relative merits of competing health care interventions might meet the first criteria and while, for the second standard, we could possibly argue that the reference case is a creative 'coherent conceptual construct' but with an unrealized potential for generating evaluable claims. Unfortunately, the reference case fails the third, and most critical standard, of evaluation through experimentation or observation. In short, theories are not enough. The failure of the reference case is that it excludes empirical testability. Models or simulations that conform to the reference case standards, lifetime or long-term cost-per-QALY claims (appropriately discounted) are not science.

A point that is worth considering is whether or not health technology assessment was initially conceived as a soft science or pseudoscience or whether it has, in Pigliucci's felicitous expression 'fallen into that category'. It is with the emergence from the late 1980s of cost-effectiveness analysis that the commitment to hypothesis testing appears to drop off the radar screen. Perhaps the key word is 'validation'. Modeled claims are, if we follow the ISPOR standards for modeling, to be judged by their correspondence to reality. Predictive validation is seen as the strongest form of evaluation, yet it can be put to one side with correspondence assured through face, internal, cross and external validity checks. If the correspondence of the constructed model is judged to be significant then it follows that the claims are necessarily entailed. Concerns with predictive experimentation or systematic observation can be put to one side. The fact that the modeled outcomes are expressed in terms that are incapable of meeting the standards for hypothesis testing adds additional weight.

Perhaps the catalyst for the fall was the publication in 1990 by the Pharmaceutical Benefits Committee, of the draft version of the Australian standards for formulary submissions, which were formally released in 1992 [25]. The standards received widespread publicity and were quickly emulated in other single payer health care jurisdictions. The guidelines made no mention of the need to present claims in evaluable terms. The modeled claim, subject to review, was accepted at face value. This position has remained unchanged throughout the subsequent versions of the guidelines. A similar position was taken in guidelines developed subsequently for other health jurisdictions. There is no evidence of any single payer health system considering the merits of evaluable claims in formulary submissions, including NICE in the UK and the adoption of the reference case.

A charitable interpretation of this 'lapse' might be that, in the 1990s, those advocating modeled or simulated claims recognized the difficulties of evaluating claims in a limited data environment. A less charitable interpretation might be that there was no pressing interest in evaluation, either from agencies such as the NICE and the PBAC or from manufacturers who were enthusiastic supports of modeled cost-effectiveness claims, yet showing less than enthusiastic support for the replication of phase 3 results. At the same time there was, as far as can be judged, little interest from health systems, notably in the US, in pursuing cost-effectiveness claims in target patient population. The notion of an outcomes based formulary did not emerge until much later [26].

A more substantive reason for the continued commitment to the construction of imaginary worlds is that there is, perhaps inadvertently, no commitment to meeting the standards of normal science. In effect, standards for modeled claims are to be understood from a relativist perspective. The relativist believes that all perspectives are equally valid [27]. In their advocacy of the equivalence or symmetry principle, from a relativist perspective health care decisions are to be understood sociologically. No one body of evidence is superior to another. There is no need to set standards for evaluable claims. For the relativist, the success of a scientific research program, in this case one built on models and simulations, rests not on its ability to generate new knowledge but on its ability to mobilize the support of a community committed to developing models and simulations. Basing decisions on models and simulations reinforces the consensus view that evidence is constructed, never discovered. Instead of coming to grips with reality science is about rhetoric, persuasion and authority. Truth is consensus and truth is constructed.

Cost-Utility an Irrelevant Metric

Given these professed advantages in a preference-based summary score that captures both morbidity and mortality across disease states, why have health systems, particularly those in the US, been reluctant to embrace this measure in evaluating the impact of competing treatment interventions?

The academic enthusiasm for developing and comparing preference measures, including somewhat inconclusive debates on the relative merits of QALYs and DALYs and whether QALYs should be discounted, does not seem to be matched by an equally enthusiastic embrace of this single valued outcome measure by health care decision makers. Indeed, the academic and health decision making audiences, with a few notable exceptions, appear to be talking past one another. The academics are busily publishing non-testable comparative product claims couched in cost per QALY terms (more often than not funded by pharmaceutical manufacturers) while health care decision makers are focused on triangulating efficiency, clinical outcomes, and the total cost of care to support value claims.

The point that is overlooked is that, irrespective of the confidence that we have in the methodological foundations of a single measure, the measure is of little interest to decision makers if utility measures such as the EQ-5D are never systematically collected or have no prospect of ever being collected. Can we seriously entertain the prospect of a commitment by health administrators to ensure QALYs are generated from representative samples of patients going forward on an annual or semi-annual basis across disease states and clusters of comorbidities? If not, then claims based on QALYs are a waste of time unless there is a commitment by those making the claim to follow through on data collection. A commitment which might be questioned if the claims are in long-term or lifetime cost-per-QALY terms with a time horizon that matches the course of the chronic disease and, as such, are incapable of being evaluated let alone replicated.

As far as can be ascertained, there is not a single health care system that regularly collects utilities as an outcome measure. Putting on one side the issue of which utility measure is best suited to the patient population, the absence of such measures embedded in electronic medical record (EMRs) or other process measures such as administrative claims, means that any cost-per-QALY claim has to be assessed through a prospective observational study. Rather than being able to access EMRs where clinical endpoints are reported on a regular basis, a study has to be designed to capture the QALY endpoints. In the US, the accountable care organizations (ACOs) are a case in point. The process and outcomes measures that are in place to support quality metric standards make no mention of QALYs. Rather, the focus is on disease specific clinical and process measures that can be readily captured.

For those who see the abandonment of the QALY as a major step backward the answer apparently lies in 'improving' the measure. Even if an 'improvement' were to occur, the question still remains: is there any likelihood of the measure having a practical role in the allocation of healthcare resources outside of the construction of cost-per-QALY imaginary worlds? [28] While there are certainly a number of initiatives that have been considered including the equity of values put on different QALYs by different groups of people and attempts to improve the sensitivity of generic instruments such as the EQ-5D-5L [29 30]. Other groups, such as the National Institute for Quality and Efficiency in Health Care (IQWiG) in Germany employ a modeled efficiency frontier (EF) framework rather than QALYs and willingness to pay thresholds to facilitate setting maximum reimbursable prices for new interventions [31]. In common with the NICE reference cases there is no intention that any of the claims made from the modeling should be evaluable.

If QALYS are irrelevant for health decision making in the US and other countries, why do manufacturers and others continue to support modeled claims for comparative cost-effectiveness expressed in utility terms? Apart from the obvious response that they don't know any better, one possibility is that they (or at least their in-house pharmacoeconomics group) have bought intro the consensus view of the role of non-testable modeled or simulated claims as a sustainable marketing strategy.

Transitioning To Evaluable Claims

Perhaps the most obvious point to make is that if it is possible to construct complex lifetime cost-per-QALY models relying on Markov processes or discreet event simulations, then it should not be difficult to reorient our thinking towards the construction of models that generate short-term evaluable predictions. Given the effort that apparently goes into the construction of lifetime imaginary worlds, as evidenced by the effort put into examining the minutiae of modeled claims by the ERGs in the UK and their equivalent in Australia in reviewing submission to the Pharmaceutical Benefits Advisory Committee (PBAC), the demand for a potentially less complex and less demanding short-term extrapolation from, say, randomized clinical trials (RCTs) does not seem an unreasonably onerous or, indeed, challenging request.

It might also seem obvious that rather than relying on the relativistic belief systems of academics advising single payer health system decision makers, it would be useful to ask health system decision makers in the US what information they consider critical in evaluating competing comparative clinical and cost-effectiveness claims. On the clinical side are they primarily interested in phase 3 results, focusing on quality, adverse events and efficacy? Or are they more focused on direct or indirect comparative effectiveness claims? As direct head-to-head claims are unlikely to capture all comparator therapies in the disease or therapeutic area, the focus is then on the type of indirect claim and proposals for how these indirect claims, as provisional hypotheses for

anticipated comparative effectiveness in target patient populations, are to be evaluated, and over what time horizon?

When it comes to comparative cost-effectiveness claims, the focus is then the cost side. Is the health system interested primarily on claims for resource utilization over the same time horizon as the clinical assessment or is the focus on total costs? If it is the latter, then a decision has to be made as the appropriate unit costs. Is the health system also interested in the impact of competing therapies on absenteeism and presenteeism? Is the health system interested in the comparative impact on adherence and persistence? Is the health system interested in the potential impact of socio-economic status of target populations on clinical outcomes?

If there is agreement that the focus of cost-effectiveness claims should be on those claims that are evaluable rather than on constructing imaginary worlds, then one way forward may well be to consider, as has been proposed on a number of occasions, that a submission to a formulary committee should be accompanied by a protocol proposing how the claims are to be evaluated in the target patient population [32 33 34]. This ensures a commitment from the manufacturer to standing behind the clinical and cost-effective claims made for the product. It also ensures that the claims made can be evaluated, either from identified data vendors utilizing retrospective or future tracking data or, if necessary from a prospective observational or experimental study. Agreement between the health system and the manufacturer who is underwriting the study can set the stage both for interim and final reports on claims assessment, but also for risk sharing agreements and discounting.

To support this proposed transition to evaluable claims, it is worth noting recently introduced guidelines for formulary evaluation released by the College of Pharmacy, University of Minnesota [35]. These guideless reject untestable modeled or simulated claims, emphasizing the importance of testable claims, experimentation and replication. A key feature is the requirement for a claims assessment protocol to accompany formulary submissions. Manufacturers are asked to underwrite value claims assessment in a timeframe relevant to formulary decisions.

Conclusions

The commitment to the construction of imaginary worlds, together with the commitment to QALYs as the 'gold standard' endpoint, is not a sound base for promoting the contribution of cost-effectiveness models and simulations to formulary decision making. This is not to deny, of course, the expedient role of non-evaluable reference QALY frameworks to justify resource allocation decisions in single payer health systems. Until questions are raised as to the empirically evaluable basis on which decisions are made, those advocates of QALY endpoints should be asked to defend their commitment to constructed evidence.

A commitment to QALYs as a gold standard should be abandoned. Apart from the obvious issues of the inherent credibility of assumptions supporting multi-attribute preference models and the absence of agreed criteria for selecting a 'superior' measure (the ultimate gold standard QALY), the fact is that health care systems are either unwilling or uninterested in capturing utilities and QALY estimates as part of their data set. The result is that we have spent 30 years in developing QALYs with the belief that, at some time a health system such as the UK NHS will commit to the regular collection of utilities from patients. There seems little prospect that this will occur. Of course, QALYs can be generated from national health surveys, but the sample size is typically too limited to capture the level of detail required to report on disease specific QALY benchmarks, let alone the impact of specific therapy interventions.

Under what circumstances will an appeal to the standards for cost-effectiveness claims put to one side the relativist claim for equivalence? Under what circumstances will those supporting modeled cost-per-QALY and similar claims attempt to escape from the sanctuary of pseudoscience? Given the consensus view of the contribution of modeled or simulated claims, the likelihood of an acceptance of the standards of normal science seems remote. There is too much sunk capital by manufacturers, consultants and 'leading' academic groups for them to abandon modeled or simulated, yet untestable, claims. The only impetus for evaluation standards will be from health care decision makers. Recognition of the fact that submissions from manufacturers have been driven by comparative product claims which lack credibility may be the catalyst for change.

In retrospect, it is doubtful, that the great expectations for QALYs could ever be realized outside of reference case imaginary worlds, or the willingness of decision makers to suspend belief in the standards of normal science, and accepted lifetime cost-per-QALY claims as decision criteria. Unless, therefore, a case can be made for short-term and evaluable QALY claims, there seems little scope for QALYS, and associated cost-per-QALY claims, as inputs to formulary decision making. Perhaps, as Pip says to Estella, it has been 'a vain hope and an idle pursuit'[36]. After over 30 years perhaps we can put QALYs to one side and return to clinically and quality specific endpoints in comparative claims for pharmaceutical products in disease and therapeutic areas.

References

[1] Caro JJ, Briggs A, Siebert U et al. Modeling good research practices – Overview: A report of the ISPOR-SMDM Modeling good practices task force – 1. *Value Health*. 2012 (15:796-803

[2] Eddy DM, Hollingworth W, Caro JJ et al. Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices Task Force – 7. *Value Health*. 2012;15:843-850

[3] Caro JJ, Eddy DM, Kan H et al. Questionnaire to assess relevance and credibility of modeled studies for informing health care decision making: An ISPOR-AMCP-NPC good practice task force report. *Value Health*. 2014;(17):174-182

[4] National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal*. London: NICE, April 2013

[5] New Zealand. Pharmaceutical Management Agency. Prescription for Pharmacoeconomic Analysis: Methods for cost-utility analysis (v 2.1). PHARMAC, 2012.

[6] Ireland. Health Information and Quality Authority. Guidelines for the Economic Evaluation of Health Technologies in Ireland. Dublin: HIQA; 2014

[7] Canadian Agency for Drugs and Technologies in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada (3rd Ed). CADTH: Ottawa, 2006

[8] Australian Government. Department of Health. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (version 4.4).* Canberra: July 2013

[9] Academy of Managed Care Pharmacy. AMCP Format for Formulary Submissions, Version 4.0, April 2016

[10] www.icer-review.org

[11] www.optum.com

[12] www.euroqol.org

[13] Drummond MF, Sculpher MJ, Torrance GW et al. Methods for the Economic Evaluation of Health Care Programmes. 3rd Ed. Oxford University press, 2005

[14] Neumann, John von, Morgenstern, Oskar, *Theory of Games and Economic Behavior*. Princeton, NJ. Princeton University Press, 1953.

[15] https://hoap.ucsd.edu/qwb-info/QWB-Manual.pdf

[16] http://www.healthutilities.com/

[17] McGregor M, Caro J. QALYs: are they helpful to decision makers? Pharmacoeconomics. 2016;24(10):947-52

[18] Beresniak A, Medina-Lara A, Auray JP et al. Validation of the underlying assumptions of the quality-adjusted life-years outcome: results from the ECHOUTCOME European project. *PharmacoEconomics.* 2015;33:61-69

[19] NICE hits back at limited EC study 28-1-2013 http://www.pharmafile.com/news/177270/nice-hits-back-limited-ec-study

[20] Drummond M, Brixner D, Gold M et al. Towards a consensus on the QALY. *Value Health*. 2009;12(Supp 1):S31-S35

[21] Persad G. Priority setting, cost-effectiveness and The Affordable Care Act. *Am J Law Med*. 2015;41:119-166

[22] Adler MD. QALYs and policy evaluation: A new perspective. *Yale J Health Policy & Ethics*;6(1):4-6

[23] Langley PC. Sunlit Uplands: The genius of the NICE Reference Case. *INNOVATIONS in Pharmacy*. 2016;7(2)): Article 12

[24] Pigliucci M. Nonsense on Stilts: How to tell science from bunk. Chicago: University of Chicago Press, 2010

[25] Commonwealth of Australia, *Draft Guidelines for the Pharmaceutical Industry om Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee: Including Submissions involving Economic Analysis* (Canberra: Department of Health, Housing, and Community Services, 1990)

[26] Sweet B, Tadlock CG, Waugh W et al. The WellPoint outcomes based formulary: enhancing the health technology assessment process. J Med Econ. 2005;8:13-25

[27] Wootton D. *The Invention of Science*. New York: Harper, 2015.

[28] Vijan S. Should we abandon QALYs as a resource allocation tool? *PharmacoEconomics*. 2006;24(10):953-4

[29] van Excel J, Baker R, Mason H et al. Public views on principles for health care priority setting. Findings of a European cross-country study using Q methodology. *Soc Sci Med*. 2015;126:128-37

[30] Herdman M, Gudex C, Lloyd A et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;126:128-37

[31] Institute for Quality and Efficiency in Health Care (IQWIG). General Methods (V 4.2). April 2015. https://www.iqwig.de/download/IQWiG_General_Methods_Version_%204-2.pdf

[32] http://www.healthutilities.com/

[33] Langley PC. Recent developments in the health technology assessment process in Fulda TR, Wertheimer AI (eds). Handbook of Pharmaceutical Public Policy. New York: Pharmaceutical Products Press. 2007.

[34] Langley P. Validation of modeled pharmacoeconomic claims in formulary submissions. *J Med Econ.* 2015;18(12):993-999

[35] Langley PC. Nullius in verba. The University of Minnesota Social and Administrative Pharmacy Program proposed Guidelines for Formulary Evaluations. *INNOVATIONS in Pharmacy*. 2016;7(2): Article 10

[36] Dickens C. Great Expectations. London, 1861