MAIMON WORKING PAPERS No. 16 SEPTEMBER 2021

REJECTING BELIEF IN PSEUDOSCIENCE: A COMMITMENT IN HEALTH TECHNOLOGY ASSESSMENT TO MEET NORMAL SCIENCE STANDARDS?

Paul C. Langley, Ph.D., Adjunct Professor, College of Pharmacy, University of Minnesota, MN

Abstract

It is now accepted that the QALY is an impossible mathematical construct. This puts an end to any further discussions of its merits or otherwise in diverse populations. Just as the QALY is an analytical dead end, so is the application of the QALY in assumption driven simulation models, best characterized by the Institute for Clinical and Economic Review (ICER) in their obsession with inventing evidence to support pricing and access recommendations. Few recognize the QALYs demise; due almost entirely to a lack of appreciation of the required levels of evidence in statistical analysis and a failure to understand how. preference scores are generated. To the point it is not the concept of a quality adjusted life year that should focus our attention, but the preference score which is an ordinal measure. If we had (which we do have) a score to capture patient need as a latent construct for a target patient population that had ratio properties, then a need quality of life assessment entirely reasonable.

Unfortunately, as exemplified by a recent review in *Health Affairs****, people, through lack of understanding, cling to the preference score QALY as the Holy Grail to support health care decisions. The manifest deficiencies of preferences score are another country; one for which few have the training or experience to address. Failure to appreciate these deficiencies means that preference based value claims for therapy response and the imaginary cost-per-QALY thresholds are a waste of time. We cannot at the same time advocate the impossibility of the QALY as a construct yet proceed to argue for inapplicability on discriminatory and ethical grounds. The former precludes the latter (or should). We must abandon the preference based QALY.***

If we are to make value claims for competing products we need an entirely new approach to reporting on quality of life (QoL) among other attributes. The purpose of this commentary is to propose a set of standards that should be observed for value claims, with specific reference to QoL claims from the perspective of patients and caregivers. This is a critical step. It is one thing to maintain the QALY is an impossible construct; it is another to point to a positive way forward to next generation value measures that meet the standards of normal science and requirements for fundamental measurement.

INTRODUCTION

Formulary decisions must rest upon claims for therapy impact, notably comparative claims, that are consistent with the standards of normal science and the axioms of fundamental measurement. Instead, after 30 plus years of health technology assessment we face exactly the opposite commitment: therapy claims focus on inventing evidence, on pseudoscience, and the implicit rejection of any concern with meeting the standards of fundamental evidence [1]. This was a deliberate choice; in the early 1990s, with the focus on a single, comprehensive claim for cost-effectiveness the multiattribute preference score entered center stage with the QALY acclaimed as the only valid construct to support modelling with pricing and access recommendations. We now realize that this is arrant nonsense.

We now realize with the benefit of hindsight and a better appreciation of the measurement standards that apply in the physical sciences, education and psychology that the decision in favor of approximate invented information in lieu of hypothesis testing to generate new evidence, was a major error, setting back for decades a commitment to the standards of normal science. Unfortunately, a number of key organizations and academic modelling centers have yet to receive, let alone understand, the memo.

There is a pervasive failure to recognize the manifest limitations of health technology assessment modeling and the required standards for coherent value claims [2]. These include a belief in generic quality adjusted life year claims (QALYs) as well as disease specific PRO measures. At the same time assumption driven lifetime simulation models also fail these standards, notably those claims for pricing and access produced by the Institute of Clinical and Economic Review (ICER) [3]. We are in the unfortunate situation; after 30 years of QALYs and imaginary simulation modeling we now realize we cannot invent evidence to support formulary submissions. Our previous efforts, funded by manufacturers, are, as intended, in large part only simulated preference-based QALY models to support product marketing.

An entirely new approach to value claims is required. Fortunately, we know the standards that should apply to value claims and, in respect of quality of life (QoL) we have a successor to the imaginary or I-QALY that meets these standards. Continued insistence on developing assumption driven model simulations for imaginary value claims is an untenable situation; the I-QALY and assumption driven simulation modeling are both analytical dead ends. We must abandon pseudoscience.

The purpose of this brief commentary is to make explicit the standards that must apply to value claims, with particular reference to disease specific and patient centric need-based quality of life (N-QoL).

WHERE IGNORANCE IS BLISS

It is important not to understate the extent to which belief in assumption driven simulation modeling and the absence of any apparent understanding of the limitations imposed by the axioms of fundamental measurement are held in health technology assessment. Previous commentaries have considered this belief system as a meme with high transmission fidelity. For over 30 years this meme, central to beliefs held by groups such as ICER and the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), academic centers of excellence devoted to modelling imaginary worlds and by journal editors (and peer reviewers) in their rejection of attempts to raise or even recognize critical questions, has held center stage.

This last point is exemplified by a recent systematic review and blog in the journal *Health Affairs* that entirely misses the point that the preference scores (and the ICER/ISPOR QALY) is an analytical dead end and should never have been developed in the first place [4]. What stands out is the apparent lack of understanding of the standards of normal science and the limitations imposed by the axioms of fundamental measurement. It is not as though these standards have just been recognized; they were proposed over a century ago with a number of papers published over the past 40 years or more pointing to the implications of an ignorance of measurement theory. This literature appears entirely foreign to the authors of thus systematic review. Indeed, one of the most egregious omissions is absence of any thought given to the implications of negative preference scores for health states. A reference to a recent valuation of the EQ-5D-5L for the US would have been instructive where out of 3,125 health states (5 symptoms, 5 levels of response), 625 (20%) had negative scores (preference range -0.575 to 1.0 [5]. This alone would inform the knowledgeable reader that preference scales cannot have ratio properties as there is no true zero. This is not a new observation; it led to a significant debate in the early 1990s and was brushed under the carpet. You cannot adjust an ordinal weighted scale algorithm to create a ratio scale *ex post facto*. There will inevitably be the possibility of negative scores for health states, particularly if the response levels are extended as with the EQ-5D-5L.

Also overlooked, as the preference score is an ordinal score, is that the distance between score rankings has no meaning; it is just a ranking. An ordinal scale can only support non-parametric statistics. Ordinal preference scores cannot support arithmetic operations for addition, subtraction, multiplication and division. Claims for an average preference score are therefore invalid. This not only invalidates entirely ICER models (and the thousands of other QALY models developed over the past 30 years) but raises the prospect that if negative preferences are present in the distribution or preferences for the target population, the average will be 'deflated' and, for the true believer,

understate QALY gains. It is perhaps fortunate, again for the true believer, that studies seldom report on the distribution of these ordinal scores.

In the discussion of preference scores the authors take the QALY as a given; there is no consideration of the undeniable fact that the various direct and indirect preference measures all produce ordinal scores. This is not surprising as the various authors and groups had no appreciation (or vague awareness) of the axioms of fundamental evidence. If you require a ratio scale this has to be built into the study design. Otherwise, you just fit numbers to data and try to approximate a 0 = death and 1 = perfect health scale. Invariably you fail (as detailed below). The QALY is a mathematically impossible construct because you cannot multiply time by an ordinal score; for multiplication to be possible you require a ratio score. Add to this the lack of appreciation of the importance a focusing on single attributes (as in the physical sciences) and not on dimensionally heterogeneous (multiattribute) scores that lack construct validity. The notion of health states defined by a bundle of clinical attributes, destroys any claim to meeting evidence standards; you fail at the first hurdle.

More puzzling is the apparent willingness to accept assumption driven simulation claims where the discredited QALY takes center stage; the Holy Grail in health technology decisions. This framework, with ICER in the US the chief malefactor, fails the standards for normal science, as well as a simple application of logic. Inventing evidence by assumption, where the assumptions drawn from literature extend decades into future, accompanied by the future family of QALYs, fail Hume's paradox of induction: *it cannot be established by observation, since we cannot observe future events… it cannot be established by logical argument , since from the fact that all past futures have resembled past pasts it does not follow that all future futures will resemble future pasts* [6]

As Bertrand Russel noted over a century ago, you cannot defend induction in probabilistic terms [7]. If you want to base claims on past observations to support assumptions as to future truth then the role of induction has to be assumed; as a principle it cannot be proved or disproved. Prior experiences does not give support to any future claims irrespective of whether or not it appears 'reasonable'; you cannot argue to choose between prior experiences. An ICER model cannot be defended on the grounds that, of the multitude of all possible models, it is the most realistic.

If this level of ignorance is to drive policy recommendations to support therapy choice and resource allocation in health care systems then we have a long way to go; as a first step, the ICER QALY should be dismissed out of hand [8]. As a final point it might be noted that the PORTAL group in the Harvard Medical School are supported financially by Arnold Ventures; Arnold Ventures are also a supporter of ICER, also in Boston. Arnold Ventures are aware of these criticisms of ICER, as is ICER who hold to a belief in a mystical ratio without a true zero. The authors make no mention of their affiliations and funding although they are associated with the PORTAL group.

RASCH MEASUREMENT AND STATISTICAL MODELLING

In the social sciences statistical modelling is the dominant analytical techniques to describe a data set. The object is to fit the model to the data, if necessary by the rejection of potential explanatory variables. This stands in contrast to the physical sciences where the measurement task is to obtain data that fit the model. The requirements of the model, construct or trait that is to be measured drive data collection and item selection. The distinction Is between exploratory/descriptive models, fitted to the data (e.g., econometric modelling), and confirmatory/predictive models utilizing probabilistic conjoint measurement, where the requirement is for the data to fit the model. It is this latter approach that drives the Rasch model. This leads to the consideration of the measurement properties of the required instrument. In human subject research where the objective is to measure latent traits we must start with a substantive theory about what it is we are trying to measure. Item development and selection must be driven by our knowledge of the latent trait. It may turn out that the latent trait is not actually quantitative. At the moment the Rasch model is the only one available to test the hypothesis that we are measuring a quantitative latent trait. Two questions are central to this: (i) how well do the empirical data fit the measurement model requirements and (ii) Does the instrument yield invariant interval-level measures for the intended purposes? As Wright states: *Rasch models are the only laws of quantification that define objective evidence, determine what is measurable, decide which data are useful, and expose which data are not* [9].

The attractive features of measurement in Rasch modeling – unidimensionality with linear, additive, invariant values on an interval-level measurement scale – exist only to the extent that the data fit the Rasch model requirements; guided, of course, by an understanding of how the latent trait will be captured in practice. To this should be added the recent development that allows us, under certain conditions, to transform the interval-level measurements to a bounded ratio scale. No other patient reported outcome instrument, whether generic or disease specific can meet these requirements for fundamental measurement. They are locked into a paradigm that dismisses (or is unaware of) the required axioms of fundamental measurement, relying on their belief in the primacy of data over substantive theory; the notion of quality control in the selected data does not arise. We have to use all the data regardless of quality and measurement properties.

Rasch measurement supports strong inferences that measured behaviors reflect the underlying latent trait or construct. The first step in Rasch modeling must be to measure the construct and evaluate construct validity as an umbrella concept to encompass content validity, face validity and concurrent validity. This can only occur at the disease or target patient group level if the resulting value claims are to have any merit in meeting the standards in normal science for credibility, empirical evaluation and replication. We must reject in their entirety value claims based on direct or indirect generic preference measures; claims from time trade off (TTO) and standard gamble (SG) direct measures as well as the EQ-5D-3L/5L, the HUI Mk2/3, the SF-36/6D and the AQoL among others; add to this the majority of patient reported outcome measures (PROs) which similarly fail the required measurement standards.

NEXT GENERATION QUALITY OF LIFE

Rejecting invented evidence also means rejecting ordinal multiattribute preference scores and the QALY. Both are well past their use by date; indeed, if they ever had one in the first place [10]. Fortunately, in respect of quality of life defined from the patient perspective, the latent construct of need fulfillment, we have a measure that meets the required measurement standards with instruments meeting Rasch measurement standards in some 30 disease states. These are not preference measures, community valuations of multiattribute health states are not involved. Indeed, as noted, attempts to value health states are absurd. This is typically overlooked with the result we have a smorgasbord of unfit preference scores which lack dimensional homogeneity and construct validity. After some 80 years of agreement on the levels of measurement this is a ridiculous situation.

Avoiding community preferences for health states defined in terms of a bundle of symptoms and functions, does not mean that the next generation measures ignore clinical symptoms and functional status. The potential contribution of these attributes is seen through the lens of the patient (or caregiver) as elements in a broader holistic framework. As the patient (or caregiver) is the ultimate beneficiary of a therapy intervention the value claim focuses on the need of the patient and the extent to which that need is fulfilled. It is the benefit a patient derives from an intervention specific to a disease state defined in the patient's own terms.

Patient focused QoL measures are not new; they have just been ignored in favor of ordinal multiattribute preference measures [11]. Developed over the past 25 years for specific chronic disease states there are now some 30 disease states covered (including: atopic dermatitis, psoriasis, growth hormone deficiency, Crohn's disease, depression, asthma, COPD, sickle cell disease, herpes, ulcerative colitis). These various measures are based on a coherent outcome model. They determine the extent to which respondents can meet their fundamental human needs. Items or statements are presented (with a binary True/Not true response) derived directly from relevant patients (and caregivers) and provide data on the value these groups derive from alternative interventions. This ability is clearly related to the symptoms and functional limitations they experience. However, in contrast to clinician determined HRQoL quality of life measures such as the EQ-5D, these new measures generate a basis for a single index of patient value or QoL rather than adding together (inappropriately) a basket of clinical outcomes defined as ordinal scales.

As disease specific measures they identify the overall impact of living with a particular disease from the patient's perspective. This provides the framework for evaluating the extent to which patient (or caregiver) need is met with competing therapy interventions. The items selected for each instrument are based upon intensive interviews with patients and an extended process of item selection through the application of Rasch Measurement Theory [12]. Items finally selected are ranked in terms of difficulty of a need being met and the ability of the respondent to meet that need expressed in

probabilistic terms. The number of items selected is relatively small, typically in the range 25-30. The instrument can be completed in 4 or 5 minutes.

This single index of patient value is transformed to a bounded ratio scale that is unique to each instrument [13]. This creates the Need-QOL (or N-QOL) measure, which is robust and accurate, meeting all the required standards detailed above. As the N-QOL is on a bounded ratio scale in the range 0 =no needs are met to 1 = all needs are met It can be used to create need-based quality of life claims by multiplying time in a disease stage by the N-QOL score to create the N-QAL. By design, negative values are impossible; scores for different instruments across disease states can be compared.

MINIMUM STANDARDS FOR VALUE CLAIMS

There are six standards that must be met for credible and evaluable value claims, including clinical endpoints, patient reported outcomes (PROs), QoL and, and resource allocation. Failing to meet any one of these standards means the value claim must be rejected. In many cases claims will have already ratio properties based on agreed clinical measurement together with measurable (e.g., CPT code) claims for resource allocation impact; costs are not an acceptable claim. The focus of Rash measurement, as the only acceptable analytical framework is, of course, on latent constructs.

1. MEETING THE STANDARDS OF NORMAL SCIENCE

The single most important standard is to meet the requirements of normal science: *All value claims must be credible, evaluable and replicable.* If not, like the QALY, the claim is nothing more than pseudoscience and must be rejected. Invented value claims have been the mainstay of health technology assessment for 30 years; to overcome this will be difficult.

2. SUBMITTING VALUE CLAIM PROTOCOLS

Manufacturers and others submitting value claims must demonstrate how the claim can be evaluated: *All value claims must be accompanied by an evaluation protocol.* Failure to provide a claims evaluation protocol must lead to a rejection of the claim.

3. RECOGNIZING THE AXIOMS OF FUNDAMENTAL MEASUREMENT

All value claims must conform to fundamental measurement standards; this means that *the claim submitted must have ratio measurement properties* with a true zero and invariance of comparisons.

SUBMITTING SINGLE ATTRIBUTE CLAIMS

Following the standards of measurement of the physical science, all value claims should be for a single attribute whether this is for clinical, outcomes, PROs, QoL or resource utilization: *value claims must be*

for single attributes defined by a ratio scale meeting requirements for construct validity, content validity and unidimensionality.

4. SUBMITTING DISEASE SPECIFIC CLAIMS

As the patient (or caregiver) is the presumed beneficiary of therapy intervention, value claims to support that intervention must be *specific to a target patient population within a disease area.*

5. REPORTING VALUE CLAIM EVALUATIONS

Value claims must, in the case of formulary submissions, be *evaluated and reported to the formulary committee* or other health system decision makers in a reasonable or meaningful time frame.

OVERVIEW: A PARADIGM SHIFT?

Optimism is a desirable quality; in the case of health technology assessment and the patent ignorance by so many of required measurement standards, notably in respect of latent constructs, there is no doubt there will be pushback and a retreat to the obvious defense, exemplified by ICER when challenged, that 'everyone does it'. Perhaps, as noted in previous commentaries, belief is strongest when the object of that belief is clearly impossible: *Certum est quia impossibile est* [14]. If this is true then we can look forward to a protracted defense as the wagons are circled and the cool aid is distributed.

We must abandon the search for a single value Holy Grail to drive formulary decisions, with acceptance or denial of care. We must base decisions on attributes specific to a disease state and established by formulary committees. On the assumption that decision makers can walk and chew gum at the same time, factoring in a range of attributes with required measurement properties together with input from patients themselves should be sufficient to negotiate an acceptable provisional price and conditions for access to care that can be modified over time as new data become available as part of ongoing disease area and therapeutic class reviews.

The advent of the disease specific N-QOL means the end of multiattribute ordinal preference scores and the impossible QALY. This provides an assured basis for value claims that represent he need of patients (and caregivers) and a measure of the extent to which that need is met. The key development that has made this possible is the ability, recently developed, to transform single index of patient value from these instrument to a bounded ratio scale with all necessary properties to evaluate need and its determinants as well as robust and accurate measures of therapy response.

The next step, given the number of instruments already developed, is to initiate a research program to evaluate need in these diverse disease areas, supported by trials and observational studies to create value claims for therapy interventions. There is no longer any need to invent evidence for non-evaluable QALY claims.

REFERENCES

[1] Langley P. Peter Rabbit is a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(2): No 22
https://pubs.lib.umn.edu/index.php/innovations/article/view/3992/2855

[2] McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting Patient Reported Outcome Instruments and Interpreting the Data they Produce*. InovPharm*. 2021;12(2): No. 17
https://pubs.lib.umn.edu/index.php/innovations/article/view/3911/2764

[3] Langley PC. Nonsense on Stilts –Part 1: The ICER 2020-20234 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12
https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348

[4] Rand L, Kesselheim A. Controversy over using Quality-Adjusted Life-Years and Cost-Effectiveness Analyses: A systematic review. *Health Affairs*. 2021;40(9)

[5] Pickard A, Law E, Jiang R et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*.2019; 22(8):931-41

[6] Magee B. Popper. London: Fontana, 1974

[7] Russell B. The Problems of Philosophy. 1012.

[8] Langley P. Abandoning Eugenics and the QALY. *InovPharm*. 2021;12(3): No.20
https://pubs.lib.umn.edu/index.php/innovations/article/view/4291/2939

[9] Wright B. Fundamental measurement for psychology. In Embretson S and Hershberger S (Eds). The new rules of measurement: What every educator and psychologist should know. Mahnah NJ: Lawrence Erlbaum Associates, 1999

[10] Langley PC and McKenna SP. Measurement, modeling and QALYs.*F1000Research.* 2020; 9: 1048 https://doi.org/10.12688/f1000research.25039.1

[11] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7
https://pubs.lib.umn.edu/index.php/innovations/article/view/3359/2517

[12] Bond T, Fox C. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 3rd Ed. New York: Routledge, 2015

[13] Langley P, McKenna S. Fundamental Measurement: The Need Fulfillment Quality of Life (N-QOL). *InovPharm*. 2021;12(2):6
https://pubs.lib.umn.edu/index.php/innovations/article/view/3798/2697

[14] Dawkins R. A Devil's Chaplain. New York: Houghton Miflin, 2004