MAIMON WORKING PAPERS www.maimonresearch.net

WORKING PAPER No. 13 MAY 2020

MORE UNNECESSARY IMAGINARY WORLDS: THE UNIVERSITY OF WASHINGTON MODELLED VALUE ASSESSMENT OF TARGETED IMMUNE MODULATORS IN ULCERATIVE COLITIS FOR THE INSTITUTE OF CLINICAL AND ECONOMIC REVIEW

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

Abstract

All too often organizations embrace standards for health technology assessment that fail to meet the standards of normal science. A continuing puzzle is why the axioms of fundamental measurement are ignored by researchers such as the University of Washington Model Group in constructing lifetime cost-per-QALY claims. The University of Washington Model Group is not alone; it is an accepted article of faith that multiattribute utility scales can be manipulated as if they had ratio scale properties. This commitment to pseudoscientific claims, embracing intelligent design rather than natural selection, is endorsed by professional groups such as ISPOR as well as by self-appointed arbiters of value assessment such as ICER. Perhaps the answer is peer pressure rather than ignorance of the axioms of fundamental measurement. After all, if cost-per-QALY constructs are rejected, then it is difficult to see what options there are for those attempting to model cost-effectiveness claims. If it is just ignorance of the axioms of fundamental measurement then a reasonable question is why these axioms, readily available on any number of internet sites, are ignored in health technology assessment programs. The purpose of this commentary is to review the ICER draft evidence report in ulcerative colitis. Not surprisingly, as previous commentaries have reported, it fails the standards of normal science. Not to put too fine a word on it, the modeled QALY (or imaginary I-QALY) claims are worthless. ICER has two options: first, to reject any criticisms of its modelling, in particular, cost-per-I-QALY claims, or to acknowledge the failure of previous evidence reports to meet the standards of normal science. Manufacturers whose products, their pricing and access, may be impacted by the ICER report on ulcerative colitis should ask for the report to be withdrawn. The recommendations lack scientific merit. They should not even be considered as approximate information. ICER claims and recommendations, based on the Washington group model are creations from a fantasy world.

Introduction

One of the more endearing features of health technology assessment is the belief that the axioms of fundamental measurement can be put aside. Unlike the physical sciences where measurement is taken seriously, measurement in the social sciences, notably in the development of patient reported outcomes (PROMS) instruments, fail the axioms of fundamental measurement. It is assumed, without justification, that the addition of response scores from Likert or similar scales have ratio properties. This is mistaken; the various instruments produce nothing other than manifest scores. The failure to recognize the importance of fundamental measurement not only characterizes the draft evidence report for ulcerative colitis but all previous ICER evidence reports [1]. This failure has implications not only for modeled cost-utility claims but also for the clinical assessment of competing therapies [2].

The feature that sets health technology assessment apart from the other social sciences, including mainstream economic analysis, is the commitment to the construction of imaginary worlds to support competing claims for products and devices. This is an absurd position, but one that is rigorously supported by the leaders in the field of cost-effectiveness analysis [3]. For those who have been trained in the standards of positive economics, the role that is assigned to the discovery of new facts, theory construction and hypothesis testing, this focus on imaginary lifetime incremental cost-per-quality adjusted life year (QALY) worlds, and their wholehearted embrace by organizations such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and groups such as the Institute for Clinical and Economic Review (ICER) in the US, is absurd. The active pursuit of approximate information and the rejection of hypothesis testing to support formulary decisions assessment is an analytical dead end; a feature that was been recognized 30 years ago. In this respect it is worth noting that it was not until the 4th edition of the Drummond et al textbook on economic evaluation (in 2015) that the question of fundamental measurement was raised; and then it confused interval and ratio scales[4]; a delayed, *ex post facto*, excuse for 30 years of modeled technology assessment claims that have rested on manifest scores.

The recently released ICER draft evidence report for ulcerative colitis rests on assumptions that are clearly indefensible. That the report should be rejected goes without saying; what is important are the reasons for its rejection. That is the purpose of this commentary with its focus on measurement. This is to emphasize the mathematical impossibility of creating QALYS by multiplying time spent by a manifest score the term imaginary QALY (or I-QALY) is applied throughout.

The University of Washington Model

The University of Washington model framework follows the standards established by ISPOR for the creation of approximate information. This is important, as it clearly puts to one side the standards of normal science, hypothesis testing and the discovery of new evidence, in favor of a framework that is designed to provide simulated approximate information for decision makers. In this case the purpose of the exercise is to create evidence for the cost-effectiveness of targeted immune modifiers (TIMs) for moderate to severe ulcerative colitis in biologic-naïve and biologic-experienced sub-populations. A total of eight products are assessed within this imaginary simulated framework: adalimumab (Humira: AbbVie); golimumab (Simponi; Janssen Biotech); infliximab (Remicade: Janssen Biotech); infliximab-dyyb (Inflectra: Pfizer); infliximab-abda (Renflexis; Merck); tofacitinib (Xeljanz: Pfizer); ustekinumab (Stelara: Janssen Biotech); and vedolizumab (Entyvio IV: Takeda) The interventions are compared to each other and to conventional treatment defined as induction with corticosteroids followed by azathioprine or mercapotopurine.

The base-case analysis takes a health care sector perspective (i.e., focused on direct medical care costs only), over a lifetime time horizon. Due to uncertainty of treatment patterns over this timeframe shorter time horizons of two, five, and 10 years were explored as additional scenario analyses. The model was structured as a Markov model with eight-week cycles, based on a common point of assessment in clinical trials to mark the end of induction and beginning of maintenance treatment. Costs and outcomes were discounted at 3% per year. The model health states were active UC, clinical response without remission, clinical remission, post-colectomy (with and without complications), and death. The model structure and health states were chosen based on the disease course, the impact of treatment, and prior economic models in ulcerative colitis. The model takes a lifetime horizon (with scenarios for shorter periods). In the base case model patients remain in until death.

The primary purpose of the model is to create imaginary lifetime incremental cost-per-I-QALY claims for the various products. This is achieved by estimating time spent in each of the four health states. Utility scores are applied to each of these states and an aggregate lifetime I-QALY count created by multiplying and aggregating time spent in each health state adjusted by the 0 – 1 manifest utility score. Utility scores are from a systematic literature review and meta-analysis of a mix of utility instruments to create a synthetic amalgam of manifest scores for ulcerative colitis disease states (active disease, clinical response without remission and clinical remission) together with post-colectomy EQ-5D utilities from a cross section survey of patients (Evidence Report Table 5.12) [5]. The model proceeds to add one 'realistic' assumption after another to create the vision of the ulcerative therapy impact future to populate the Markov framework.

The base case outcomes for the Washington model are presented in cost-per-I-QALY terms for the two subgroups: biologically naïve and biologically experienced. Comparisons with conventional treatment yield scenarios for the biologically-naïve yield I-QALY gains for TIM therapies ranging from 15.80 to 16.04 I-QALYS. Total model lifetime imaginary costs range from $379,000 (conventional treatment) to $502,000 for ustekinumab. The result is an imaginary cost per I-QALY gained, compared to conventional treatment, of $229,000 to $584,000. For the biologically naïve the incremental I-QALYs (measure unspecified) range from 15.80 to 16.04 I-QALY years. Similar results are presented for the biologically experienced population with estimates of costs per I-QALY gained ranging from $553,000 to $382,000 with conventional treatment as the reference case. Incremental lifetime IALYs gained range from 15.60 to 15.78 years.

The Washington model then proceeds to present imaginary probabilistic sensitivity analyses; the likelihood that a product will be cost-effective at selected cost-per-I-QALY thresholds. For TIMs versus conventional treatment in the hypothetical biologic-naive population at a threshold of $250,000 per I-QALY the imaginary likelihood is 60% for infliximab-abda and 59% for infliximab in the base case lifetime model. In the case of the hypothetical biologic experienced population, at $250,000, the highest likelihood was for tofactinib at 8%.

The final step in the ICER imaginary incremental cost-per-I-QALY construct is to introduce cost-per-I-QALY thresholds. In order to create recommendations for prices to achieve the imaginary thresholds, the model has to be adjusted to accommodate a scenario that allows pricing recommendations. Using net prices, modeled incremental cost-effectiveness ratios were found to be above commonly cited thresholds for cost-effectiveness for all TIMS in the base case analysis.

Deconstructing the Imaginary Washington Model

The details of these thresholds and recommendations for price discounting from WAC are immaterial; the key point is the absurd lifetime value assessment framework. As detailed below, the I-QALYS are mathematically impossible constructs and, as a result, threshold based price discounting claims are nonsensical. This commentary on the ICER evidence report for ulcerative colitis will focus on four aspects of the University of Washington's ulcerative colitis imaginary model. Issues addressed are: (i) standards of normal science; (ii) assumptions in models; (iii) approximate information and (iv) fundamental measurement and the construction of impossible I-QALYS. All are ignored by the ICER model builders. This is not to single out this particular university group but to point out that in common with other groups contracted by ICER to build models, they are woefully unaware of the standards of normal science, notably the axioms of fundamental measurement.

Standards of Normal Science

It is important for ICER and its contracted modelling groups to understand the basis on which new evidence is provisionally discovered (not proved). The paradigm that supports discovery in the development of pharmaceutical products through the phases of drug development should apply equally to claims for the impact of products in treating populations. We don't ask manufacturers to create evidence from assumptions; the evidence will emerge from a process of conjecture and refutation or hypothesis testing. If the evidence to support claims is not available at product launch then instead of creating imaginary cost-utility constructs to generate ersatz evidence claims, the focus should be on evidence platforms to support models with credible and evaluable claims.

The requirement for testable hypotheses in the evaluation and provisional acceptance of claims made for pharmaceutical products and devices is unexceptional. Since the 17th century, it has been accepted that if a research agenda is to advance, if there is to be an accretion of knowledge, there has to be a process of discovering new facts. ICER is opposed to this. By the 1660s, the scientific method, following the seminal contributions of Bacon, Galileo, Huygens and Boyle, had been clearly articulated by associations such as the Academia del Cimento in Florence (1657) and the Royal Society in England (founded 1660; Royal Charter 1662) with their respective mottos *Provando e Riprovando* (prove and again prove) and *nullius in verba* (take no man's word for it) [6].

By the early 20th century, standards for empirical assessment were put on a sound methodological basis by Popper (Sir Karl Popper 1902-1994) in his advocacy of a process of 'conjecture and refutation [7] [8] . Hypotheses or claims must be capable of falsification; indeed, they should be framed in such a way that makes falsification likely. Although Popper's view on what demarcates science (e.g., natural selection) from pseudoscience (e.g., intelligent design) is now seen as an oversimplification involving more than just the criteria of falsification, the demarcation problem remains [9]. Certainly, there are different ways of doing science but what all scientific inquiry has in common is the 'construction of empirically verifiable theories and hypotheses'. Empirical testability is the 'one major characteristic distinguishing science from pseudoscience'; theories must be tested against data. Hence pivotal clinical trials; not simulated imaginary worlds with selected data inputs from pivotal trial data to recycle old (and imagined) facts. We can only justify our preference for a theory by continued evaluation and replication of claims. Constructing imaginary worlds, even if the justification is that they are 'for information' is, to use Bentham's (Jeremy Bentham 1748-1832) memorable phrase 'nonsense on stilts'. If there is a belief, as subscribed to by ICER and its contractors, in the sure and certain hope of constructing imaginary worlds, to drive formulary and pricing decisions, then it needs to be made clear that this is a belief that lacks scientific merit. It fails the demarcation test; it is pseudoscience (i.e., pure bunk).

Approximate Information (or Disinformation)

It is worth emphasizing that ISPOR, as noted above, ICER's methodological mentor, explicitly disavows hypothesis testing as a core activity in health technology assessment. ICER presumably concurs. The primary role of health technology assessment for ISPOR (and ICER) is to create 'approximate information'. It is not clear what this means (presumably it can be distinguished from 'approximate disinformation') as there is not, in the imaginary world of ISPOR/ICER modeling, any known reference point for 'true information' to judge approximation. How close are we? It is difficult to be approximate to the 'truth' when the context is imaginary and the 'truth' will only be revealed 10, 20 or 30 years or more ahead if all the assumptions in the model are realized. The OED definition may relate approximate to a 'known' truth but in the construction of imaginary worlds then can be no such reference point.

It is difficult to judge how formulary committees would react to ICER saying it supported the construction of approximate and unevaluable 'approximate information' in decisions. Does imaginary evidence (or claims) constructed from lifetime models trump (apologies!) evidence based medicine where claims are provisional and where a research program could be proposed to capture evidence that was unavailable at product launch. Constructing imaginary evidence is, of course, a lot easier and more cost-effective. With the opportunity, as noted by ICER to revisit imaginary claims if data become available to modify assumptions, producing a new round of imaginary and non-evaluable claims. The great advantage non-evaluable claims have is precisely that; no one can be held accountable for the claim. To quote Wolfgang Pauli (physicist, 1900 – 1958): *It is not only not right, it is even not wrong.* But perhaps this was always the intent in the ICER reference case.

Choice of Assumptions

One of the more intriguing elements in the Washington model is the insistence on 'realistic assumptions'. But what does this mean? Is there an accepted distinction, a criterion for categorizing assumptions as 'realistic'? Is it possible to be unequivocal as to the realism of a set of assumptions that might hold over the lifetime of modelled target patient populations? The number of assumptions that have to be captured to support the various simulations and their scenario progeny in ulcerative colitis is truly awesome; some come from the literature, others are pure guesswork. This does not mean there is only one possible model; there is presumably scope for a multiverse of models each with their own family of scenarios, each producing claims which can never be evaluated. Indeed, were never meant to be capable of evaluation. That is the great advantage of building assumption driven imaginary worlds; only the assumptions can be challenged (which seems a fruitless endeavor).

Unfortunately, even if an assumption driving the imaginary value assessment framework is defended by appealing to the literature (including pivotal clinical trials) the effort is wasted. The point, and this goes back to Hume's (David Hume 1711 – 1776) induction problem, is that we cannot ask clients in health care to believe in models constructed on the belief that prior assumptions will hold into the future. It is logically indefensible: 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*' [10].

Fundamental Measurement

There are four main types of measurement scale; putting to one side conjoint simultaneous measurement which underpins Rasch Measurement Theory (RMT)[11]. These are: nominal, ordinal, interval and ratio. Each satisfies one or more of the properties of: (i) identity, where each value has a unique meaning; (ii) magnitude, where each value has an ordered relationship to other values; (iii) interval, where scale units are equal to one another; and (iv) ratio, where there is a 'true zero' below which no value exists. Nominal scales are purely descriptive and have no inherent value in terms of magnitude. Ordinal scales have both identity and magnitude in an ordered relation but the unknown distances between the ranks means the scale is capable only of generating medians and modes; it is a manifest scale. The interval scale has identity, magnitude and equal intervals. It supports the mathematical operations of addition and subtraction. A ratio scale satisfies all properties, supporting the additional mathematical operations of multiplication and division. Recognition and adherence to these fundamental axioms of measurement theory is critical if an instrument is to have any credibility. In the physical sciences this has been long recognized; accurate measurement is the key to hypothesis testing and the discovery of new facts. The same arguments apply to the social sciences. Unfortunately, they appear all too often to be absent in health technology assessment.

Conjoint simultaneous measurement was proposed by Luce and Tukey in the early 1960s as a new type of fundamental measurement, which subsumed the other types. A specific application was to provide a framework for detecting measurement structures in non-physical attributes (e.g., quality of life). In a slightly modified form this is now applied as RMT. Without going into the technical details of matrix construction to establish whether it is possible to identify measurement structures, Rasch measurement standards, following those of the physical sciences are designed to create instruments that have interval measurement properties. In terms of item selection and order, RMT provides a framework where the probability of successfully responding to an item depends on the difference between the ability of the person and the difficulty of the item; hence the focus on item selection and the ordering of items to meet RMT standards. The more difficult an item, to give an affirmative response, the greater is the required ability of the patient to respond affirmatively.

As detailed in previous commentaries RMT is not compatible with either classical test theory (CTT) or item response theory (IRT). They are, as Bond and Cox point out, competing paradigms [8]. RMT takes the perspective that if the instrument is to meet fundamental measurement standards then we should adopt the Rasch *data-to-model* paradigm. If we are not concerned with, or are happy to ignore, questions of fundamental measurement, then we can follow the CTT or IRT *model-to-data* paradigm. The key distinction is that *RMT uses the measurement procedures of the physical sciences as the reference point* [8]. We can aim for the standards in the physical sciences by, as Stevens pointed out in the 1940s, allocating numbers to events *according to certain rules* [12]. It is these rules that comprise RMT. To reiterate: RMT is designed to construct fundamental measures. CTT and IRT focus on the observed data, these data have primacy and the results describe those data. As Bond and Cox emphasize: In general, CTT and IRT are *exploratory* and *descriptive* models; the Rasch model is *confirmatory* and *predictive* [8] . If RMT is ignored then, by default, instruments utilizing Likert scales or similar frameworks will fail to meet the required axioms of fundamental measurement and remain ordinal or manifest scales.

This failure to meet the axioms of fundamental measurement is a characteristic of virtually all patient reported outcomes measures (PROMs). This is most obvious in the application of Likert scales in developing responses and the treatment of overall scores (the 'add 'em up' school favored by all too many clinicians in PROMS development). In ulcerative colitis the classic example of an instrument that fails to meet required standards for fundamental measurement is the multiattribute Inflammatory Bowel Disease Questionnaire (IBDQ). With development stretching over some 20 years the long version of the IBDQ comprises 32 questions, each item represented on a Likert scale, with integer values attached from 1 = worst to 7 = best). The IBQD has four domains: bowel function, systemic function, social and emotional. The range of possible scores is from 32 to 224, with scores allowed for each of the four domains.

Unfortunately, this scoring system fails the standards required for fundamental measurement. Apart from the fact that, even though RMT had been applied since the 1960s, there was no intent (or recognition) that if you want to develop an instrument to assess response to therapy then it needs to be constructed to meet the required fundamental measurement standards. Simply adding up Likert integers is unacceptable. The usual method for analyzing Likert scale data is to disregard the implicit subjectivity of individual responses, while making unwarranted assumptions about the meaning attached to the integer values. The scoring assumes that the scale is interval level with an integer value of 1 indicating a higher degree of agreement than 2, with integer 2 higher than integer 3 and so on. While this may seem a trivial point, it relies not on an assumption of a ranked response but of a ranked response where the distance between the integer responses is invariant. That is, an integer value of 2 means that the respondent is feeling, for fatigue as an example, twice as fatigued as a response with integer value 4; or someone with an integer value of 1 is feeling five times as fatigued as someone who never feels fatigued and scores an integer value of 5. By assigning integer values the user falls into the trap of assuming that the responses are on a ratio scale where the integer value of 1 is treated as a true zero.

But that is not all. It is also assumed that the responses for each of the items are equivalent. Each item contributes the same amount to the total score. This assumes that the respondent finds it equally easy or difficult to respond to each item. This is at variance to the Rasch measurement model where, following the axioms of conjoint simultaneous measurement, it assumes that the probability of affirming an item depends on two factors: the ability of the respondent and the difficulty of the item. In short, the traditional summation of Likert scale data is based on the assumption that all of the items are of equal difficulty for all respondents and that the threshold between steps is of equal distance or equal value. Unless we are entitled, by assumption, to reject the axioms of fundamental measurement it is illogical to add the integer items across the 32-items for a total score; or for the various sub-domains. Any claims for response to therapy are nonsense as we have no idea what the intervals mean between the item integers.

Utilities and QALYs

Responses to the position taken by ICER regarding their advocacy of imaginary worlds suggests either than ICER is unaware (along with the various university modeling groups) of the axioms of fundamental measurement or prefers to duck behind the defense that constructing I-QALYs and then imaginary worlds is the 'gold-standard' in health technology assessment. If it is then, to extend the metaphor, it is becoming somewhat tarnished. There is no escape: utilities are manifest scores. They are constructed to have this property. The Washington group's apparent ignorance of the axioms of fundamental measurement is seen, as an example, their approach to alternative utilities where they propose to average over EQ-5D baseline scores without realizing that manifest scores cannot be added and averaged (pg. 94).

I-QALYS are the Achilles heel of the ISPOR and ICER model universe. Exeunt I-QALYs and the fantasy cost-per-QALY house of cards collapses. Apart from their use in the ISPOR and ICER contribution to the science fiction literature (ICER receives a Hugo award for imaginary constructs), QALYs can only survive if the measure is credible, evaluable and replicable. The concept of a QALY is not new; it goes back some 40 plus years with the notion of combining time spent in a disease state with some multiplicative 'score' on a required ratio scale of 0 to 1 (death to perfect health). Combining the two, multiplying time by utility is assumed to produce a 'reaL' QALY (undefined). In the University of Washington imaginary world these are combined to produce I-QALYs for the modeled lifetime as the hypothetical target population as the TIM populations proceed through the hypothetical disease stages with imaginary time spent in each of these stages.

The case presented here is that the EQ-5D-3L, as a generic multiattribute instrument, generates ordinal or manifest scores [13]. It does not have interval properties (i.e., invariance of comparisons) and it certainly does not have ratio properties as the EQ-5D-3L 'score' lacks a true zero (i.e., distance from zero). Unfortunately, the EQ-5D-3L scale has no demonstrable interval measurement properties (with odd ceiling and floor effects) as well as allowing negative utilities (below a true zero). The same is true of other instruments to include the HUI Mk3, Australian Quality of life Instrument (AQoL) and the time-trade off (TTO) measure [14] . Of course, if the EQ-5D-3L fails to demonstrate interval properties, then it is a waste of time to consider whether it has ratio properties. The actual range for the EQ-5D-3L is not from 0 = death to 1 = perfect health, but from -0.59 = experiencing five severe symptoms (aggregate score -1.59) to 1.0 = perfect health; the algorithm to compute utilities allows negative values (and always will). The fact that the EQ-5D-3L has ordinal properties is easily demonstrated: the symptom elements that comprise the EQ-5D-3L attributes are on an ordinal scale. Simply applying community preference weights and adding these up results in an ordinal scale.

There is the further question of unidimensionality. Measurement scales should have the property of unidimensionality. The focus should be on one attribute at a time. We must avoid confusing a number of attributes into a single score. Mutiattribute scales such as the EQ-5D-3L reduce confidence in predictions and the score is a less useful manifest summary. In Rasch modeling, estimates of item difficulty and person ability are meaningful if every question item contributes to the measurement of a single underlying attribute. Our analytical procedures, if we are to meet the property of unidimensionality, must incorporate indicators of the extent to which the persons and items fit our concept of an ideal unidimensional line. Items should contribute in a meaningful way to the construct/concept being investigated.

In the case of the EQ-5D-3L and other multiattribute scales, the notion of unidimensionality is absent. While it is claimed to capture health related quality of life (HRQoL), there is no single attribute or latent construct. It comprises 5 symptoms (mobility, self-care, usual activity, pain/discomfort, anxiety depression) with three ordinal response levels (no problem, some problems and major problems); creating a multiattribute scale with ordinal properties. Each of the symptoms is an attribute that could be the foundation for its own unidimensional scale. While ISPOR/ICER apparently believes the EQ-5D-3L has ratio properties this is demonstrably false given negative utilities. But perhaps this is not as egregious as the 'false assumption' position taken by authors where it is acknowledged that the EQ-5D-3L lacks a true zero but that, in order to maintain the I-QALY illusion, we assume it has ratio properties [15]:

The situation does not change when we move from the EQ-5D-3L to the EQ-5D-5L (introduced in 2009) where there are 5 response levels. Increasing the allowed ordinal responses to five reduces the number of respondents with extreme problems. The result is a range, still including negative utilities, from -0.29 to 1.0. Even so, it is still an ordinal or manifest score. There is an ongoing debate over the switch from the EQ-5D-3L to Eq-5D-5L is defensible. They best answer is 'why bother'; if only exchanging one manifest score for another.

If the objective is to create QALYS that are not imaginary constructs then the time spent in a disease state (which has a true zero with interval properties) then the utility 'adjustment' score must also have ratio properties with a true zero and constrained to the range 0 to 1 with interval properties. Just as it is nonsensical to have a negative utility, so a utility > 1 would be nonsensical. If utilities for QALYs are to be constructed then the utility instrument must be designed to have those properties. Perhaps the Washington group and ICER might make the first steps in that direction by acknowledging RMT. Otherwise we are left with the I-QALY; indeed a manifest score that can create negative I-QALYs. It is entirely possible that over a lifetime a target population can generate only negative I-QALYs. This is an interesting perspective: can we interpret for ICER pricing and access recommendations incremental negative I-QALYs?

The fact that the EQ-5D-3l and EQ-5D-5L lack ratio properties should not be thought of as a case for completely abandoning these instruments. As long as the limitations of the measure are recognized there can be no objection to the reporting of manifest scores and the distribution of respondents across the ordinal response levels within symptom categories. Indeed there is a substantial literature that may be of interest yet has been ignored in the ICER evidence report. It is surprising that the report fails to follow up on the ISPOR recommendations for systematic utility reviews. The focus by the University of Washington model group on a single slipshod systematic review (covering the period to August 2015) that proposes a synthetic composite manifest utility score for disease states in Crohn's disease and ulcerative colitis is nonsense [5]. The authors of this review fail to recognize the limitation of manifest scores to support econometric and statistical operations. The resulting so-called utilities have no known measurement properties; indeed where the EQ-5D is included they certainly fail to have ratio properties. If ICER and the Washington group wish to provide a systematic reviews of a preferred HRQoL measure such as the EQ-5D-3L or its successor the EQ-5D-5L then that is their decision. It might be interesting to devotees of manifest scores but irrelevant if the purpose is to justify creating I-QALYs. Such a review of manifest scores and I-QALYs could be presented for the period to end-2019.

Even if ISPOR/ICER were willing to recognize the absence of fundamental measurement properties in the EQ-5D-3L (and other generic utility instruments), this does not mean that this would give succor to their belief in fabricated imaginary evidence. The ICER value assessment framework would still fail the demarcation test as pseudoscience. It is also difficult to see how ICER might underwrite a 'utility' instrument that met the standards required (a true zero yet capped at unity). After all, instruments developed by application of RMT focus, as noted above, on the response to interventions on a constructed interval scale from ordinal responses rather than attempting to go the further step of creating instruments which have ratio properties [16] [17] [18].

Developing unidimensional RMT standard instruments for ulcerative colitis, even if they only captured interval properties to report response to therapy for needs-fulfillment quality of life, would be a positive step forward compared to the ICER belief in imaginary I-QALYs. The apparent lack of awareness of the axioms of fundamental measurement is a major obstacle. It is difficult to see how this message can be conveyed either through the University of Washington model group (to acknowledge their errors) or directly to ICER.

Although overlooked by the Washington model group there is an instrument for Crohn's disease that has been developed utilizing RMT. This instrument, the Crohn's Life Impact Questionnaire (CLIQ), was developed some seven years ago [19]. Two references describe the development of the CLIQ [20] [21]. Sample questions from the CLIQ are:

- I avoid eating large meals (true/not true)
- I worry about where the nearest toilet is (true/not true)
- I need a lot of time to do things in the morning
- I find it hard to manage (true/not true)
- There is not much fun in my life (true/not true)
- I have lost interest in sex (true/not true)
- I often feel lonely (true/not true)
- It affects my image of myself (true/not true)

As noted above, the items or questions are designed to capture the probability of affirmation given the ability of a respondent to affirm a response to an item and the difficulty of the item to be affirmed. The impact of a therapy will then be judged by its ability to improve QoL for the target patient group with Crohn's disease. QoL is high when most needs are fulfilled; low when most are not. The therapy response, or comparative assessment of therapies, is to assess the response to therapy from baseline. This is captured by the sufficient or summary statistic which is the sum of the ‘true’ items reported by study subjects taking binary values. The more ‘difficult’ it is to affirm an item, the lower the probability it will be affirmed. The challenge for competing therapies is to affirm more items where the item order reflects the likelihood that the item may be confirmed given respondent ability. The more the items that are affirmed by respondents over baseline the greater their needs are fulfilled and claims for comparative response to therapy. As a true latent scale, the CLIQ responses support a range of statistical techniques to evaluate response to therapy. QALYS play no part, nor do we have to indulge in creating imaginary lifetime worlds.

Conclusions

There are two conclusions to draw from this evaluation of the ICER draft evidence report. First that the report, specifically the incremental cost per I-QALY model developed by the University of Washington model group, should be withdrawn. Second, that if claims are to be made about the impact of alternative TIMs then they should be restricted to the Mayo scores as recognized markers for relapse and remission. There should be no attempt to go beyond these clinical markers to incremental lifetime cost per I-QALY fantasy constructs. If evidence for comparative effectiveness is absent, then ICER would be better placed as the arbiter of claims to suggest how evidence gaps might be remedied, together with proposals for evaluating quality of life with instruments that meet the axioms of fundamental measurement.

There is no half-way-house. As this is a draft report ICER might attempt to ‘compromise’ by asking the University of Washington model group to come up with a ‘revised’ imaginary lifetime model with EQ-5D-3L or EQ-5D-5L utilities. This would be a waste of time. Whichever generic utility measure is chosen, it will fail the axioms of fundamental measurement. It will still be a manifest score. A score than cannot be used to create modelled QALYs because it lacks ratio properties. Attempting to utilize the IBDQ is similarly a waste of time as this measure also fails the required axioms.

Much as ICER (and its supports in ISPOR) might attempt to argue that their imaginary reference case framework is a standard in health technology assessment, it is analytical dead end [2]. Its demise is long overdue. ISPOR and ICER should acknowledge this both to those groups, manufacturer’s and health systems, who fund ICER’s imaginary creations and to those health system decision makers (and media representatives) who are naïve enough to believe, take at face value, ICER’s recommendations for pricing, product access and budget impact. ICER should affirm, as it has not done so far, a commitment to the standards of normal science and the primacy of real world evidence. If this commitment is made then the imaginary value assessment, creating approximate ‘pseudo realistic’ information can be abandoned along with the absurd belief in the existence of a true zero for generic multiattribute utility scales. The likelihood of this happening is zero; ICER has too much vested in its I-QALY business model to welcome the ridicule to which it might be exposed. It would no doubt be supported by ISPOR and its contracted academic modelling groups.

A commitment to disease specific, patient centric interval response instruments provides a firm foundation for evidence based medicine. We can abandon imaginary lifetime value assessments and focus instead on claims for quality of life that are credible, evaluable and replicable. We have a way forward: the application of RMT to disease specific instrument development to capture response to therapy on interval scales. We can focus on discovering new facts rather than recycling assumptions. It is unlikely, however that a positive outcome as outlined above will have any chance of mainstream success. Health technology assessment has far too much to lose. Leaders in organizations such as ISPOR have invested 30 years of academic and pseudo-academic endeavor into constructing imaginary worlds and proposing the dominant role of approximate information or disinformation in decision making with the ubiquitous I-QALY, like Widow Twankey, at center stage.

REFERENCES

[1] Institute for Clinical and Economic Review. Targeted Immune Modulators for Ulcerative Colitis: Effectiveness and value. May 2020 . https://icer-review.org/wp-content/uploads/2019/09/ICER_UC_Draft_Evidence_Report_052620.pdf

[2] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-20234 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12

[3] Neumann P, Willke R, Garrison L. A health economics approach to US value assessment frameworks – Introduction: An ISPOR Special Task Force Report (1). *Value Health*. 2018;21:119-123

[4] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes (4th Ed.) New York: Oxford University Press, 2015

[5] Malinowski KP, Kawalec P. Health utility of patients with Crohn's disease and ulcerative colitis: a systematic review and meta-analysis. *Expert Rev Pharmacoecon Outcomes Research.* 2016;16(4):441-453

[6] Wootton D. The Invention of Science: A new history of the scientific revolution. New York: Harper Collins, 2015.

[7] Popper KR., The logic of scientific discovery .New York: Harper, 1959.

[8] Lakatos I, Musgrave A (eds.). Criticism and the growth of knowledge. Cambridge: University Press, 1970.

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

[10] Magee B. Popper. London; Fontana, 1973

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

[12] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-680

[13] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice. *J Rehabil Med*. 2012.44:97-98

[14] Richardson J, Hawthorne G. Negative utility scores and evaluating the AQoL worst health state. Working Paper 113. Centre for Health Program Evaluation. Monash University. June 2001

[15] Yang Z, Luyo N, Bonsel G et al. Selecting health states for EQ-5D-3L valuation studies: Statistical considerations matter. *Value Health.* 2018;21:456-61

[16] Tennant A, McKenna S, Hagel P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health*. 2004;7(Suppl 1):S22-S26

[17]McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes 1: The search for the Holy Grail. *J Med Econ*. 1019;22(6):516-522

[18] McKenna S, Heaney A, Wilburn J. Measurement of patient reported outcomes 2: Are current measures failing us? *J Med Econ*. 2019;22(6):S23-30993

[19] Galen Research, Manchester UK www.galen-research.com

[20] Wilburn J, McKenna SP, Twiss J et al. Assessing Quality of Life in Crohn's Disease: Development and validation of the Crohn's Life Impact Questionnaire (CLIQ). *Qual Life Res*. 24(9):2270-88

[21] Wilburn J, Twiss J, Kemp K et al. A qualitative study of the impact of Crohn's disease from a patient's perspective. *Frontline Gastroenterol*. 2017;8(1):68-73