The Impossible QALY and the Denial of Fundamental Measurement: Rejecting the University of Washington Value Assessment of Targeted Immune Modulators (TIMS) in Ulcerative Colitis for the Institute for Clinical and Economic Review (ICER)

Paul C Langley, PhD, 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, which they do not. 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. More to the point, if you have been an advocate of imaginary simulations a Damascene epiphany creates both psychological and professional challenges. 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 September 11th 2020 evidence report in ulcerative colitis, with particular reference to ICER's responses to questions raised in the public comment period on the measurement properties (or their absence) for utility scales; in this context the EQ-5D instruments. The critique pointed out that the utility scores had ordinal properties. ICER, without proof, disputed this statement asserting that health economists believed (or assumed) they were ratio scales. This is nonsensical. ICER has two options: first, to continue to believe that the EQ-5D instruments had ratio properties or second, to acknowledge that they indeed only had ordinal properties, rejecting their many modeled claims for pricing and access. Not surprisingly, the possibility of a Damascene epiphany was rejected. ICER maintained its assertion that health economists, presumably all of them, believe or possibly just assume for analytical convenience that the EQ-5D-3L and similar measures are in fact on a ratio scale. This introduces a new concept in fundamental measurement: a ratio scale without a true zero but with negative values. ICER is quite prepared to admit that negative I-QALYs are possible and their lifetime cost-per-incremental I-QALY modelling can yield negative I-QALYs.

Keywords: ulcerative colitis, quality of life, ordinal scores, rejecting QALYs, I-QALYs

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 health technology assessment, notably in the development of patient reported outcomes (PROMS) instruments, fails 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 ordinal scores. The failure to recognize the importance of fundamental measurement not only characterizes the September 11th 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 where protocols include primary outcome measures that fail to meet the required measurement standards [2].

A further 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, with recognition of 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 had been recognized 30 years ago.

Corresponding author: Paul C Langley, PhD
Adjunct Professor, College of Pharmacy
University of Minnesota, Minneapolis, MN
Director, Maimon Research LLC; Tucson, AZ
Email: langley@maimonresearch.com
Website: www.maimonresearch.net

DOI: https://doi.org/10.24926/iip.v11i3.3330

The ICER final evidence report for ulcerative colitis, as a recent example, 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. This is the purpose of this commentary with its focus on measurement and ICER's response to a series of questions raised regarding the earlier draft evidence report focusing on ICER's understanding of the axioms of fundamental measurement [4]. Specifically, the mathematical impossibility of creating QALYS by multiplying time spent in a disease stage by an ordinal score. Hence the use, in this review, of the term imaginary QALY (or I-QALY).

It might also be noted that the ICER approach puts to one side the standards of normal science. None of the ICER claims meet these standards: none are credible, evaluable and replicable; add to this ICER's use of the term 'evidence'. The term 'evidence' as it is used by ICER is not 'evidence' as it would be interpreted in the physical and social sciences. For ICER, evidence means any claim that can be made; it can include evidence for randomized clinical trials which meet the standards of normal science as well as 'imaginary' or 'constructed' evidence which is created by the modeled simulation. This needs clarification. The term 'scientific evidence, to which ICER is presumably attempting to subscribe, means evidence that serves either to support or counter a *scientific* theory or hypothesis. Such *evidence* is expected to be empirical *evidence* and interpretable in accordance with *scientific* method [5]. Or, following the OED, *the available body of facts or information indicating whether a belief or proposition is true or valid*. If this standard is applied, then the ICER use of the term evidence is, to say the least, misleading. The 'facts' of the ICER case are constructed not discovered.

The University of Washington Model

The University of Washington model framework follows the standards established by ISPOR for the creation of approximate [and impossible] 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, in ICER's words, 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 Washington model is to create by assumption imaginary lifetime incremental cost-per-I-QALY claims for the various products. This is achieved by simulating time spent in each of the four health states. EQ-5D-3L ordinal 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 ordinal utility score. In the draft evidence report utility scores were from a systematic literature review and meta-analysis of a mix of utility instruments to create a synthetic amalgam of ordinal scores for ulcerative colitis disease states (active disease, clinical response without remission and clinical remission) together with post-colectomy EQ-5D-3L utilities from a cross section survey of patients (Draft Evidence Report Table 5.12).

Utility scores applied to the final evidence report differ from those presented in the draft evidence report. The more recent utility scores are from an observational Australian study of UC with the EQ-5D-5L [6]. Even so, the earlier study was a reference utility point for the later study (footnote Table 5.12 final evidence report). There is no prior assessment of the fundamental measurement properties of the various PROMs used in the study. Nevertheless, the assumption is made that the EQ-5D-5L has ratio measurement properties with summaries presented as means and standard deviations. These are meaningless as the scale has ordinal properties. This applies to both the authors of this and a companion observational study using the same metric in the UK [7], as well as the authors of the Washington model.

In any event, the upshot of this fantasy creation with I-QALYS was to assume that all utilities had ratio properties with four utility values for the various modeled UC health states: active UC 0.68 (95% CI 0.63 to 0.73); clinical response without remission).78 (95% CI 0.71 to 0.85); clinical remission 0.81 (95% CI 0.77 to 0.85; and post-colectomy 0.79 (95% CI 0.77 to 0.81). Although these statistics are clearly nonsensical with an ordinal utility scale, it is worth noting that the 95% CIs overlap for the active UC and clinical response without remission, as do clinical

DOI: https://doi.org/10.24926/iip.v11i3.3330

remission and post-colectomy. It is not clear what the implications are for the modeling.

A point to note is that depending on the instrument or competing instruments, utility scores can differ significantly. This raises the obvious question: can we choose the utility scores that best meets our needs? One score could create substantially greater I-QALYs than another within the same imaginary model framework. On what criteria should we select our ordinal utilities? In UC there is no agreement on which set of ordinal EQ-5D-3L utilities should be commonly applied in modeled imaginary claims; but perhaps we are saved by sensitivity analyses to capture the possible range of ordinal scores (which is disallowed as the utilities are ordinal). The result of this fantasy construction yields a table of I-QALYs for each UC product. For the biologically naïve hypothetical population the discounted lifetime I-QALYs range from 15.596 to 15.681 for the seven selected TIMS, compared to 15.574 for conventional treatment. For the biologically experienced hypothetical population the discounted I-QALYs yield a similar picture. With only four TIMs compared to conventional treatment, TIM I-QALYs range from 15.410 to 15.449; conventional treatment yields an I-QALY estimate of 15.393. Apart from the incredible precision with which these imaginary claims are made, the differences are such that they can all be expected to yield essentially the same imaginary benefit. Cost-per-QALY claims will be driven almost exclusively by price. However, before the media and others who are unaware of the imaginary nature of ICER claims argue that this is hardly a ringing imaginary endorsement of competing benefits, the point should be emphasized that a closer understanding on the nature of assumptions driving ICER simulations and non-evaluable nature of these claims, means that they are mathematically impossible, apart from failing the standards of normal science. They are the equivalent of claims from advocates of intelligent design.

Enter the thresholds. These are expressed in imaginary terms of cost-per-I-QALY ranging, in $50,000 increments from $50,000 to $250,000. For the biologically naïve the discounted cost per I-QALY gained range from $186,000 (Infliximab–dyyb) to $1,870,000 (Adalimumab). The corresponding fantasy estimates for the biologic experienced were $495.00 for Tofacitinib and $1,885,000 for Adalimumab. These corresponded to proposed price discounts for all TIMs. Threshold analyses, where cost per I-QALY proposals are compared to lifetime discounted cost per I-QALY fantasy constructs annual TIM prices proposed ranged from $6,824 for adalimumab in the biologic–experienced population to $26,624 (ustekinumab in the biologic naïve population. The author point out that these proposals are due to the minimal I-QALYs gained generated by the model; more to the point they are nonsensical. They defy, not only the axioms of fundamental measurement in even attempting to construct and model I-QALYs but the more general critique of ICER imaginary worlds that these claims are neither credible, nor evaluable or replicable.

Manufacturers should be advised not to take these recommendations for discounting seriously. It is just a fantasy or imaginary exercise. The analysis fails at the first hurdle: the I-QALY. While, as noted below, ICER puts the axioms of fundamental measurement aside in applying the novel concept of the "ICER I-Ratio QALY", a ratio scale without a true zero with by assumption the ability to support all arithmetic operations, the model itself is just a series of assumptions. The Washington model builders should have known better; perhaps a course on fundamental measurement?

If manufacturers are challenged by those who insist in factoring ICER fantasy claims for I-QALYs and discounting into pricing negotiations, then the answer is to point to the failure to meet the standards of normal science and the retreat to an unsustainable belief that the EQ-5D-3L/5L utility has other than ordinal properties. Fantasy should not eclipse reality in formulary decisions. All claims must be evaluated; real world evidence is more than constructing fantasy simulations.

Deconstructing the ICER Fantasy Construct

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. Deconstructing the ulcerative colitis imaginary model requires addressing four issues: (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 Washington model builders.

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. 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. Certainly, there are different

DOI: https://doi.org/10.24926/iip.v11i3.3330

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. We can only justify our preference for a theory by continued evaluation and replication of claims.

Approximate Information

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. Even so, for ICER presumably 'the truth is out there'.

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. Unfortunately, this is not even 'approximate information'. Given the mathematical impossibility of creating I-QALYs this is actually 'impossible information'. If it has meaning then this must be in some ICER alternative reality.

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. 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[8]. Given this, it has always been a puzzle why reviewers suggest options for new assumptions when an ICER-type model is considered; it seems, to bring in a tired cliché, rather like rearranging deckchairs on the Titanic.

Fundamental Measurement

There are four main types of measurement scale; putting to one side conjoint simultaneous measurement which underpins Rasch Measurement Theory (RMT)[9]. 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 an ordinal 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, including those designed to capture patient outcomes is to have any credibility [10]. 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.

It is also apparent that, in utilizing utility scales as if they had ratio properties, the authors of the evidence report have also overlooked the issue of dimensional homogeneity [11]. In the physical sciences instruments are designed to capture and report on a single attribute. This avoids confusion in attempting to unscramble aggregate scores that are the result of combing different attributes as well as being, from the perspective of measurement theory, inconsistent with fundamental axioms. If attribute scores are to be combined then they must exhibit dimensional homogeneity. Otherwise we are left with a ratbag of the sum of ordinal scales that says little if anything about response to therapy; a multidimensional composite index. Dimensional homogeneity is critical to instruments that meet the standards of fundamental measurement. Variables can only be combined if they have the same dimension. If they fail, then they lack construct validity. It is invalid to add together variables that lack a common dimension.

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, including dimensional homogeneity, or prefers to duck behind the defense that constructing I-QALYs and then imaginary worlds is the 'gold-standard' in health technology assessment (i.e., follow the leaders in the field). The implicit assumption is that utilities and other composite PROs have ratio properties.

DOI: https://doi.org/10.24926/iip.v11i3.3330

Unfortunately, they are constructed to have only ordinal properties with the added bonus of dimensional heterogeneity. The argument that the EQ-5D-3L can be defended because its value sets are based on time trade off (TTO) assessments is not a defense. These are described by Drummond et al as TTO tariffs with the scoring formula created by econometric modelling; i.e., fitting the model to the data [12]. There was apparently no intention to construct a scoring formula that yielded an interval let alone a ratio scale. TTO, as a recent paper by Lugnér and Krabbe points out, not only lacks a coherent analytical framework for valuing states worse than death but that the TTO does not take into account crucial requirements in measurement theory such as unidimensionality and the invariance principle [13]. ICER's and the Washington group's apparent ignorance of the axioms of fundamental measurement is seen, as an example, in their earlier approach to alternative utilities where they propose to average over EQ-5D baseline scores without realizing that ordinal scores cannot be added and averaged let alone combining composite scores from different instruments.

ICER's belief system, its adherence to the I-QALY, has been highlighted in a recent publication that has considered the origin of the commitment to constructing imaginary worlds to create cost-effectiveness claims [14]. The argument presented is that in the early 1990s the leaders in health technology assessment had two options: First, a commitment to a research program to meet evidence gaps and to create a platform for monitoring and feedback of claims in target populations; or, second, a commitment to creating imaginary simulation models. The latter, the easy way out, was taken. It gave immediate results at low cost. The result is 30 years of wasted effort based on an I-QALY framework which leads to impossible conclusions.

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 [15] [16] [17].

ICER Question and Response

Given that ICER's claim to fame as the health technology assessment arbiter in the US rests in large part on its reference case model and the assumption that the EQ-5D-3L/5L utility score has ratio measurement properties, it is reasonable to ask ICER to justify this assumption. It is not sufficient to argue that they are merely following 'accepted practice' in health technology assessment in the creation of imaginary evidence. There must be an ability or a belief that they can legitimately use ordinal utility scales to create I-QALYs. To this end, a series of questions on measurement theory were submitted to ICER as part of the public comment window for the draft evidence report. These questions were intended to be as specific as possible in evaluating ICER's continued use of ordinal utility scales and whether they had 'proof' that the utility scales had not just latent but ratio measurement properties. A previous attempt to elicit a response from ICER in respect of their modeling of cystic fibrosis therapies was inconclusive with ICER working around the questions [18].

The supporting letter to ICER in the public comment period was designed to give the reasons for the questions and to give a series of references that might enlighten ICER and the Washington group on the existence of these axioms. The questions, the ICER response and comments on the ICER response are presented in Table 1. The responses are noted together with comments on their relevance. The responses, as detailed, are hardly reassuring. ICER bases its 'case for the defense' on the apparent fact that everyone does it by assumption. In this relativistic world view, for ICER, truth (or unquestioning use) is consensus.

The widespread use of the EQ-5D measure, where the value sets are defined by the TTO overlooks, in its acceptance, the manifest difficulties that attach to the TTO as a preference measure, most notably in what appears to be intractable problems with (i) administration, (ii) the concept of death and the assignment of zero to death – which is not actually a health state but the absence of health; and (iii) the accommodation of states worse than death which in the EQ-5D algorithms for creating scale values yield negative values. In addition, to reiterate the point made earlier, the TTO does not take into account the crucial requirements in measurement theory such as unidimensionality and the invariance principle. If, as ICER appears to argue, the unproven yet widely accepted ratio measurement properties of the EQ-5D instruments rest on the ratio measurement properties of the TTO, then there is a problem as the TTO cannot support, not only interval claims but certainly not ratio claims as there is no true zero. Indeed, to clarify a point, if the EQ-5D does not have interval properties then in cannot have ratio properties; unless we assume it can. Where we are dealing with abstract concepts such as quality of life, creating an instrument to capture this attribute is not easy. At best, as demonstrated by RMT, we can create instruments with interval properties for response to therapy.

We may however be missing the point entirely; it is a closely held belief. The ICER belief in the ratio properties of the EQ-5D is by assumption; the absence of proof is just put to one side as irrelevant. Indeed ICER may believe in an entirely new concept in measurement theory; a ratio scale without a true zero that can support arithmetic operations, including multiplication, to create QALYs and incremental lifetime cost-per-QALY models

DOI: https://doi.org/10.24926/iip.v11i3.3330

for value assessment and long term imaginary benefit. This allows negative QALYs to be created and, indeed, a possible lifetime incremental negative cost-per-QALY outcome. Indeed, to clarify a point, if the EQ-5D does not have interval properties then in cannot have ratio properties.

ICER's responses show a complete lack of understanding of the constraints imposed by the axioms of fundamental measurement. This is not a unique situation; over 16,000 cost and QALY papers have been published over the past 30 years. Acceptance of the QALY as a legitimate, if a mathematically impossible construct, seems pervasive. ICER does not want to know the absurdity of its position in promoting impossible information for pricing and formulary decision making. ICER insists, by assumption not proof, that an ordinal utility scale mysteriously transforms to a ratio scale. If so, this opens up an entirely new concept in fundamental measurement; the "ICER-ratio or I-ratio scale", the scale that, despite all arguments to the contrary, has ratio properties. The ratio scale you need when you don't have a ratio scale.

Conclusions

There are two conclusions to draw from this evaluation of the ICER evidence report. First that the report, specifically the incremental cost per I-QALY model developed by the University of Washington model group, should be rejected. Second, that if claims are to be made about the impact of alternative TIMs then they should be restricted to clinical outcome measures that meet the axioms of fundamental measurement. There is no evidence to support such an evaluation. 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.

Much as ICER (and its supporters in ISPOR) might attempt to argue that their imaginary reference case framework is a standard in health technology assessment, it is an 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 if it had to admit to years of impossible recommendations for pricing. 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 a pole dancer, at center stage.

A final word to manufacturers: whether your pricing and rebate policy is considered 'out of line' is a contractual issue with health decision makers. What should be rejected from negotiations is a naive belief in the relevance of ICER simulations and I-QALY constructs to any discussion. It must be made clear that ICER is an analytical dead end. The criticisms presented here should be sufficient for that case to be made.

Conflict of Interest: PCL is an Advisory Board member and consultant to the Patient Access and Affordability Project, a program of Patients Rising

DOI: https://doi.org/10.24926/iip.v11i3.3330

References

[1] Ollendorf DA, Bloudek L, Carlson JJ, Pandey R, Fazioli K, Chapman R, Bradt P, Pearson SD. Targeted Immune Modulators for Ulcerative Colitis: Effectiveness and Value; Evidence Report. Institute for Clinical and Economic Review, September 11, 2020. https://icerreview.org/topic/ulcerative-colitis/.

[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 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348

[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] Ollendorf DA, Bloudek L, Carlson J, Pandey R, Fazioli K, Chapman R, Bradt P, Pearson SD. Targeted Immune Modulators for Ulcerative Colitis: Effectiveness and Value; Draft Evidence Report. Institute for Clinical and Economic Review, May 26, 2020. https://icerreview.org/topic/ulcerative-colitis/

[5] Wikipedia 'Scientific Evidence'.

[6] Gibson PR, Vaizey C, Black CM, et al. Relationship between disease severity and quality of life and assessment of health care utilization and cost for ulcerative colitis in Australia: a cross sectional, observational study. *J Crohns Colitis*. 2014;8(7):598-606

[7] Vaizey C, Gibson P, Black C et al. Disease status, patient quality of life and healthcare resource use for ulcerative colitis in the UK: an observational study. *Frontline Gastro*. 2014;5:183-9

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

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

[10] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: awaiting peer review] F1000Research 2020, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[11] McKenna S, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality. *J Med Econ*. 2020 DOI: 10.1080/13696998.2020.1797755

[12] 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

[13] Lugnér AK, Krabbe P. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. *Exp Rev Pharmacoeconomics Outcomes Res*. 2020; 29(4):331-342

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

[15] 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

[16]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

[17] McKenna S, Heaney A, Wilburn J. Measurement of patient reported outcomes 2: Are current measures failing us? *J Med Econ*. 2019;22(6):S23-30

[18] Langley PC. Value Assessment in Cystic Fibrosis: ICER's rejection of the axioms of fundamental measurement. *Inov Pharm*. 2020;11(2): No. 8 https://pubs.lib.umn.edu/index.php/innovations/article/view/3248/2395

DOI: https://doi.org/10.24926/iip.v11i3.3330

Table 1
ICER's Responses to Measurement Questions

Question to ICER	ICER Response	Comment
1 When a QALY is constructed time spent in a disease state is multiplied by a utility score on a range 0 = death to 1 = perfect health. Would ICER agree that this requires the utility score to have ratio properties?	Questions 1 -3 have been responded to as a single question. ICER states: *We (and most health economists)* ***have the understanding*** *(emphasis added) that the EQ-5D (and other multiattribute instruments) do have ratio properties. The EQ-5D value sets are based on time trade-off assessments (which are interval level) with preference weights assigned to different attributes. We fail to see why this should be considered as an ordinal (ranked) scale. ICER believes that the dead state represents a natural zero pont on a scale of health related quality of life. Negative utility values on the EQ-5D scale represent states considered worse than dead.*	This is a truly amazing response; and one that is demonstrably false. For ICER everything in constructed simulations is by assumption. ICER and others may assume anything; in this case to assume the TTO tariffs of the EQ-5D algorithms have ratio properties is complete nonsense. Unfortunately, ICER does not provide a proof of this bizarre assertion. Similarly the TTO technique does not yield interval let alone ratio properties (see the Lugnér and Krabbe reference). The TTO tariffs created by the EQ-5D scoring formulas from the econometric modelling have only ordinal properties. If ICER continues to insist that they can defy the axioms of fundamental measurement they are entitled to do so; hoping presumably that they will be believed. If, as discussed in the text, ICER insists on this ratio property then the EQ-5D-3L with a range from -0.59 to 1.0 must have (somewhere) a true zero. However, ICER, in their reformulation of measurement theory must prove that in the absence of a true zero multiplication (to create QALYs) is possible. Can we see this proof? This proof must support all arithmetic operations (but not be assumption). However, we do have the intriguing but weird possibility of negative QALYs! I suppose there is an upside. Consider the phrase 'have the understanding'. Can health economists demonstrate that the EQ-5D-3L, even with negative values, has ratio properties which requires a true zero? Can ICER show that time-trade off has unidimensionality and interval properties? The answer is that it does not: to claim that the EQ-5D-3L scale has ratio properties because the TTO has interval properties is just nonsense. ICER might demonstrate how an interval scale can be (and apparently has been) transformed to a ratio scale. We might have the understanding that the moon is made of green cheese; this does mean it has. At least this claim can be empirically assessed unlike ICER claims. Indeed, ICER admits that there can be states worse than dead (i.e., negative utilities) which means that the scale does not have ratio properties. Perhaps ICER should make its mind up.
2 Ratio property means that the measurement scale must have a true zero. This means that the EQ-5D-3L should have a true zero. In fact, EQ-5D-3L utilities can take negative values (with a range -0.59 to 1). Would ICER agree that this means the EQ-5D-3L is not a ratio scale?	*As above*	ICER does not give a coherent answer. The TTO does not have unidimensional and invariance properties. We cannot just assume that it has. The TTO preferences are ordinal. The responses to the Eq-5D-3L questions on symptoms are for ordinal scales. The result is an ordinal scale. It might be instructive for ICER to review the EQ-5D algorithms that create ordinal scales.

DOI: https://doi.org/10.24926/iip.v11i3.3330

3 In respect of Q2, if ICER believes that the EQ-5D-3L instrument has, despite negative values, a true zero, could ICER provide a proof?	*As above*	No proof (or even a reference) was forthcoming. The key issue that ICER and others overlook is that if you want an instrument to have fundamental measurement properties then it has to be designed to have them.
4 It appears to be commonly assumed that the EQ-5D-3L (in common with other generic instruments) meets the axioms for invariance of comparisons. That is, it has interval scoring properties. Would ICER agree?	*The EQ-5D multiattribute utility function is designed so that a utility difference of 0.05 is considered equivalent regardless of the starting point.*	If so, no references were provided nor any proof. The EQ-5D lacks unidimensionality and does not have interval properties. It has negative values which mean negative QALYs. It was not designed with invariance as a characteristic. ICER's response is incorrect.
5 In respect of Q4, if ICER believes that the EQ-5D-3L has interval properties, could ICER provide a proof?	No answer	No proof; presumably because it is impossible to provide one
6 If ICER cannot provide a proof that the EQ-5D-3L has interval properties, how does ICER justify the creation of QALYs as responses to therapy?	*Please see above responses*	The above responses do not answer the question: issue is you cannot use an ordinal score to multiply time spent in a disease state. The QALY is mathematically impossible
7 Over the past 20 years commentaries from measurement theory specialists have made the case that instruments such as the EQ-5D-3L, in fact the majority of patient reported outcomes instruments, have lacked ratio (and interval) properties. Is ICER aware of this literature and would ICER care to comment? [3] [4]	*Please see above responses*	The above responses do not answer the question. It would be useful if ICER and the Washington group could have reviewed the references provided. We can only assume they are unaware of this literature which goes back 60 years (or a century if the early formalization by Thurstone and be Stevens is included)
8 In respect of the draft evidence report for TIMs in ulcerative colitis, the utility scores appear to be an amalgam over different generic instruments (the Malinowski & Kawalec paper). How does ICER justify this given the differences that exist between the various instruments?	No answer	Apparently, where required, the University of Washington modelling group will use any available utility score. These scores were abandoned in the final evidence report (no reason given)
9 Table 5.12 in the draft evidence report provides 95% confidence intervals for four disease states. As the various scales, on which these are based, are ordinal how are	No answer	Which means they believe it has the required ratio properties; how is not discussed. Once again 'proof by assumption'.

these justified? If ICER believes these are justified could ICER demonstrate that the consequent utility scale has the required measurement properties to support confidence intervals? Given the 95% confidence intervals overlap, can you claim that the utility scores are significantly different? At what level? Does this mean there are only two disease stages that yield significantly different utilities (clinical remission vs. other three)? Should the model be reworked for two utility measures?		ICER in the evidence report has moved on by rejecting the utility scores in the first model by those generated directly (in this case from an Australian observational study). Irrespective, the EQ-5D-5L scales are still ordinal; the authors of the Australian study mistakenly believe otherwise and assume (with references or proof) that they are ratio scales.
10 Again, is respect of Table 5.12, could you detail: (i) the attributes captured by the utility scale (i.e., symptoms); (ii) the ordinal response levels for each attribute; and (iii) the preference weights or values for each response level by symptom attribute? Or is this impossible given the meta-analytical basis for aggregating over quite different utility systems? Do these utility scores have ratio properties? Can ICER demonstrate that this is the case?	No answer	Apparently any utility score will do; they are still ordinal scores even if derived from an observational study. They are, of course, all ratio scales by assumption.
11 If the EQ-5D-3L (or other generic utility scale) cannot be shown to have ratio (and interval) properties why does ICER persist in creating lifetime cost-per-QALY claims? As the utility scale is ordinal then the QALY is an impossible construct? Would ICER agree?	No answer	Because the ICER business case rests on the assumption that QALYs can be created by assuming the utility scale has ratio properties.
12 If, given that the QALY is mathematically impossible, would ICER inform its audience that it recognizes this but insists that there is still merit in constructing imaginary value assessments on imaginary QALYS to create imaginary claims information?	No answer	This won't happen; the business case is inviolate

DOI: https://doi.org/10.24926/iip.v11i3.3330