MAIMON WORKING PAPER No. 24 OCTOBER 2020

MORE ICER NONSENSE ON SEVERE HEMOPHILIA A: THE UNIVERSITY OF ILLINOIS MODELLED EVIDENCE REPORT FOR VALOCTOCOGENE ROXAPAVOVEC AND EMICIZUMAB

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

Abstract

The persistence of belief in and application of modelled incremental cost-per-QALYs continues. The draft and final evidence reports reports from the Institute for Clinical and Economic Review (ICER) for valoctocogene roxapavovec and emicizumab for severe adult hemophilia A continues this tradition; creating imaginary evidence to support pricing recommendations for pharmaceutical products and devices. The concern is that the ICER model is taken seriously. The case has been made for a number of years that the ICER reference case framework fails to meet the standards of normal science; claims made are neither credible, nor evaluable and replicable. This is seem, most egregiously, in the ICER belief that the EQ-5D-3L utility scores have ratio measurement properties. This is demonstrably false. The utility score has only ordinal properties. It cannot be used to create QALYs. The ICER reference case collapses. This means that cost-per-QALY lifetime estimates are meaningless. Any conclusions or recommendations must be rejected out of hand. The purpose of this commentary is to present the case against the University of Illinois model for hemophilia A. The critique focus not only on denial of the standards of normal science but, more specifically, on the limitations imposed by the axioms of fundamental measurement for both QALYS but also for the Haem-A-QoL instrument which, although used extensively in clinical trials is, once again, a measure with only ordinal properties and which cannot support claims for therapy response. Given the pipeline for hemophilia therapies this should serve as a salutary lesson for those in health technology assessment faced with possible assessments of cost-effectiveness by ICER and their academic consultants..

INTRODUCTION

The lack of appreciation for the axioms of fundamental measurement for those involved in health technology assessment is pervasive. For over 30 years professional associations such as the International Society for Pharmacoeconomic and Outcomes Research (ISPOR) have been advised that a failure to address the issue of measurement in the creation of patient reported outcomes instruments (PROMS) must call into question claims for response and cost-effectiveness. In fact the situation is more concerning if we are to focus on response to therapy. If we wish to measure response then the instrument must, at least, have interval if not ratio properties. Few PROMS meet this standard for the simple reason that none were designed to have this property. An instrument might meet psychometric classical test theory (CTT) standards or item response theory (IRT) criteria, but this does mean they have a required unidimensional interval priority.

The purpose here is twofold: first, to make clear that the University of Illinois model is a waste of time; it contributes nothing to an evaluation of the various hemophilia A therapies and, second, to make the case that if we are to apply the Haem-A-Qol instrument in hemophilia trials this is to endorse a PROM that lacks the ability to monitor response to therapy [1]. This critique applies with equal force to the earlier ICER evidence report for emicizumab for hemophilia A with inhibitors and its resulting endorsement by the ICER-nominated New England Comparative Effectiveness Council (CEPAC) a creation of ICER [2].

THE CASE FOR THE PROSECUTION

The case that the ICER reference framework for cost-per-QALY in hemophilia A and other disease areas lacks any pretense to be taken seriously rests on five observations:

- Normal Science: since the scientific revolution of the 17th century, progress and the discovery of new facts has rested on hypothesis testing; claims to be evaluated must be credible, evaluable and replicable. ICER disagrees. For ICER claims are to be constructed. Lifetime simulated cost-per-QUAY claims, linked to value assessment thresholds are the arbiter of pricing and access recommendations. Evidence for formulary decisions for ICER is invented not discovered.
- Approximate Information: in line with the dictates of ISPOR, ICER believes not in the discovery of new facts to support formulary decisions, but the construction of 'realistic' simulation models that provide a robust guide to the next 10, 20 or 30 years to create information. The information is necessarily approximate (but to what no one knows). Unfortunately, given the impossibility of a QALY, the information is impossible not approximate.
- Assumptions: ICER's futuristic simulation rests on assumptions, both from the literature, including clinical trials, and from just guesswork. ICER overlooks a logical problem. Assumptions built on prior observations may not hold in the future. The simulation model is an unacceptable construct as it is built entirely on assumptions.
- Ordinal EQ-5D-3L Scores: it has been argued convincingly that the EQ-5D-3L utility scale (in common with other generic utility instruments) has only ordinal properties; it can only support medians, modes and non-parametric statistics. It cannot support any of the four arithmetical operations [3].
- I-QALY: the impossible QALY is the undoing of the ICER reference case and other similar simulations; an ordinal score cannot be combined with time spent in a disease state to create QALY time equivalents; the QALY is an impossible (hence I-QALY) construct; it is mathematically impossible [4]

THE UNIVERSITY OF ILLINOIS MODEL

Set against these observations, which ICER has been aware of yet ignores, the imaginary modelling endeavor developed by the modelling group at the College of Pharmacy, University of Illinois is, quite frankly, not only misleading but a waste of time. The Illinois model assesses long tem cost-effectiveness within a modeled reference case simulation. The primary aim of the economic analysis was to compare valoctocogene roxaparvovec and emicizumab to prophylaxis with factor VIII in patients with hemophilia A without inhibitors to factor VIII who are eligible for prophylactic therapy. A Markov de novo decision analytic model was developed for two imaginary assessments: (i) the evaluation of valoctocogene roxaparvovec in adult patients with severe hemophilia A without inhibitors with dual base case analyses following ICER's imaginary ultra-rare disease frameworks and (ii) the evaluation of emicizumab in patients with hemophilia A without inhibitors eligible for factor VIII prophylaxis. The cycle length in each model was 6 months within a lifetime horizon. Given the importance of acute bleeds in hemophilia A the model is structured with tunnel states that ranged from 0-28 Petersson scores (PS). Transitions through the PS states were based on the modelled expected frequency of joint bleeds associated with treatment and consequential expected increase in PS. The model can be viewed as having four states: no arthropathy, arthropathy, joint bleeds along with related costs and impacts on patient utilities. Patients remained in the model until they died.

Needless to say, this modeling framework is best described as pseudoscience. There are no credible and evaluable claims for competing therapy interventions nor any possibility of claims assessment and replication. Claims are presented specific to a model, its scenario analyses, its assumptions regarding an unknown yet competitor free future and a naïve belief in implicit ratio measurement properties of the EQ-5D-3L utility score and the I-QALY. While it is possible to debate choice of assumption, to argue over whether some assumptions are more realistic than others (looking forward 10, 20 and 30 years) and the choice of appropriate scenarios, this is clearly a waste of time. The ICER reference case model is doomed not only by the failure to recognize the standards of normal science, but compounding this by a lack of awareness of the axioms of fundamental measurement.

Summarizing the imaginary modeled results, even for illustrative purposes is a waste of time. The manifest failures evidenced in the exercise make any discussion of lifetime I-QALYs and thresholds redundant. To this should be added the FDA’s rejection of approval for valoctocogene roxaparvovec in August 2020 with a request for a further two years of clinical data followed by the pushback by the European Medicines Agency in September 2020 asking for a further 12 months of data. This is illustrative not only of the attempts by groups such as ICER to build imaginary models on limited data but of the need for a research program to support real world data and evaluate outcomes.

ORDINAL UTILITIES AND THE I-QALY

A characteristic that is all too common in studies that have reported generating EQ-5D-3L and similar utility measures by stage of disease within target populations is the failure to recognize the ordinal nature of the utility scale. This has been noted for at least the past 30 years. Instead, the assumption is made that the utility has ratio properties to allow the range of arithmetic operations. This is held to irrespective of the floor and ceiling effects noted for the scale, the obviously ordinal nature of the symptom response levels and, most oddly, the act that the algorithm for generating the utility score from preferences can create negative values (-0.59 for the EQ-5D-3L). This allows for the creation of negative QALYs. Even if a case could be made for interval EQ-5D-3L scores (which it cannot), the requirement for a ratio scale with a ‘true zero’ is absent. Surprisingly, this is not as if in the very early days of the embrace of the EQ-5D-3L warning signs were not apparent [5].

Central to the model is the I-QALY. The utilities that are utilized are from a recent study by O’Hara et al reporting a large scale application of the EQ-5D-3L to evaluate the impact of severe hemophilia. Accepting the O’Hara et al analysis means that the model falls at the first hurdle: the implicit assumption that the EQ-5D-3L has ratio properties [6]. This leads the authors to undertake a number of misapplied statistical analyses including calculating means and standard deviations and utilizing the EQ-5D-3L score in regression modelling as a continuous variable. The resulting Table 5.4 health state utilities presented in the evidence report by age group and Petterson score is a complete nonsense. It might also be pointed out that in the O’Hara et al paper there appears to be confusion regarding what is being reported as measured in the regression model. The dependent variable is described as non-drug related direct costs (NDDCs) rather than, presumably, the EQ-5D-3L assumed ‘ratio’ scores. This is presumably accounted for by a companion model present by O’Hara et al in another journal where the objective is to assess the determinants of NDDCs [7]. Perhaps editing could have been better coordinated.

There is no objection to reporting utility scores as long as the properties of ordinal multiattribute scales are recognized. This is, all too often, not the case. An early study by Neufeld et al, cited in the ICER evidence report reports on the application, among other measures, of a small sample reporting their EQ-5D-5L scores on bleed and non-bleed days [8]. Unfortunately, the authors report mean scores for the hemophilia sample. This is nonsense as the key characteristic of an ordinal scale is that the distances between the scores for individuals are not known: we can present median and modal values but not mean values. A similar criticism applies to a further reference, Fischer et al, to assess the relation between haemophilic arthropathy with health related quality of life utilizing SF-36 summary scores and SF-6D utilities [9]. In both cases the authors failed to recognize that they were attempting to create summary statistics and measures of response from ordinal scales. The same mistake is made in the Ballal et al reference where the impact of pain, in a hypothetical comparison of two cohorts of high-titer inhibitor patients is presented [10]. Simulated EQ-5D scores for a typical patient with and without knees surgery are presented together with, as the piece de resistance, costs per QALY. Again, there is no recognition that EQ-5D scores are ordinal and that the I-QALY is an impossible construct. The final reference to note is Naraine et al where the utility measure employs the standard gamble technique [11]. Again, this technique creates ordinal scores. Of note is its application in the University of Illinois model. No thought seems to be given to the fact that these ordinal scores are constructed on different assumptions and that, while yielding nonsensical claims, should not be captured in the same model; or will any utility suffice to create I-QALY claims? To emphasize a key point made earlier: if you want any measure to have specific properties then that has to be determined from inception.

ICERS VIEWS ON ORDINAL SCALES

Following the release of the draft evidence report offered the opportunity to respond as part of a public comment window. This is the opportunity to probe ICERs modeling with ordinal utility scores. In an important respect the conclusions are obvious: as the utility scores are ordinal then the value assessment framework collapses. Even so, there are those that subscribe to the belief, or dogma, in a 'ratio' utility; a firm commitment to imaginary simulations that ignore issues of fundamental measurement. Perhaps the more impossible a belief, the more strongly it is held!

It is instructive to consider ICER's response to questions on whether or not they could prove that the EQ-5D-3L had ratio properties in the case of ulcerative colitis. ICER's response was:

> *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 point on a scale of health related quality of life. Negative utility values on the EQ-5D scale represent states considered worse than dead.*

A detailed rebuttal of this rather weird and inconsistent response has been published [12]. Rather than repeat this rebuttal (although it might be noted that the TTO does not have interval properties [13]), ICER was asked once again to provide a proof that the EQ-5D, which features in the bladder cancer report, has a ratio scale. It is somewhat self-defeating to maintain that the EQ-5D-3L has a natural zero and in the next sentence point out that EQ-5D can create negative utility values. ICER cannot have it both ways: a pseudo-ratio scale with negative utilities and a natural zero point? Should this be seen as a major advance in fundamental measurement theory? Unfortunately, it is not clear what a natural zero point means. In the case of the EQ-5D-3L the zero is simply an artifact of the equation or algorithm that creates the utilities. Unlike, for example, a true zero in measuring weight (i.e., you can't have negative weights). If ICER or the academic group at the University of Illinois are not sure of this, they might refer to the standard textbook on health technology assessment [14].

In the public comments on the draft hemophilia report, ICER was asked to respond to two questions:

(i) Do you have a proof that the EQ-5D-3L/5L have ratio measurement properties; and
(ii) Do you have a proof that the TTO has interval measurement properties

ICER was also asked to avoid using the phrase 'have the understanding'.

ICER did not respond directly, assuming apparently that referencing a recent paper would be the definitive response [15]. In respect of (ii) we know the time trade off (TTO) measure does not have interval properties [16]. In respect of (i) I asked colleagues in fundamental measurement to join me in reviewing the paper. While the authors assert the EQ-5D-3L has ratio properties our unanimous view was that the arguments were nonsensical; no proof was offered that utility scales actually had a 'true zero'. The authors of this study did not find a 'true zero' in the EQ-5D-3L/5L utility scale. In all cases their 'proof' of the value of zero was assigned. Nor could the authors even demonstrate that the utility scale had interval properties. Even if this had been 'demonstrated' it would not have saved the QALY because to create a QALY the utility scale has to have ratio properties [17]. Hence, the ICER evidence model for incremental cost per QALY claims is an impossible mathematical construct [18]. The EQ-5D-3L/5L utility scale is nothing more than an ordinal scale of raw scores placed for convenience on an interval scored number line. This error has been pointed out by a number of authors over the past 30 years [19] [20].

If ICER and the modelers understood the axioms of fundamental measurement, it would be obvious that to ask for a 'proof' that utility scales such as the EQ-5D-3L have a ratio property is impossible; it was proposed as a form of rhetorical question (the classical form *subjectio*) because we know the answer. You either have a true zero or you do not. But perhaps ICER and supporters have an insight that has escaped the attention of those formalizing measurement theory over the past century: there is a mystical ratio scale that admits of negative values. i.e., there is no true zero [21] . A ratio scale in disguise; the transformation of the ordinal raw scores created by utility algorithms into a mystical ratio scale (with assumed interval scoring properties as part of the mystical transformation). But perhaps a ratio scale without a true zero is just one more assumption to support ICER's modeling of cost-per-QALY claims? After all, what is one more assumption? Indeed, an assumption that is required to support the QALY measures; if utilities are actually raw or ordinal scores then the QALY is an impossible mathematical construct. Of course, if the mystical ratio scale fails to demonstrate interval properties then the scale cannot create impossible QALYs. To support multiplication and division you need invariance of comparisons.

For readers who may be unsure of the nature of utility scores, consider this an illustration from each of the EQ-5D stable and the HUI stable of measures [22]. Let's consider first the EQ-5D-3L, which is the most popular of the mystical ratio scales without a true zero.

The algorithm or equation that translates community valuations of respondent health status comprises 5 symptoms and three ordinal response levels within each symptom. Starting from an assigned value of perfect health of unity (where all responses have an assigned value of 0 indicating ‘no problem’)the utility value is created for each of 243 health states:

- subtract a constant term (for any dysfunctional state) [- 0.081]
- subtract five dimension scores for mobility level, self-care level, usual activities level, pain or discomfort level, anxiety or depression level (3 levels: no problems, some problems, extreme problems) [no problems = 0; some problems range 0.069 to 0.123; extreme problems range 0.049 to 0.386];
- subtract N3 level (where level 3 occurs within at least one dimension) [– 0.123}

As an example of the raw score calculation consider a situation where the respondent reports extreme problems on each of the five symptoms (health state 33333). The equation is:

$$U = 1 – (0.081 + 0.314 + 0.214 + 0.094 + 0.386 + 0.236 + 0.269) = 1 – 1.594 = -0.594$$

For the more dysfunctional states the algorithm will yield negative scores within the possible range of 1 to -0.594. The negative values (below death) are simply health states worse than death. Presumably, if these individuals died then the overall quality of life of the community would improve, but would be impossible to quantify as the worst state raw scores are ordinal measures. The health state [11111] yields a score of 1 = perfect health as all perfect health responses are weighted as zero. In the case of the EQ-5D-5L where there is the opportunity for five responses within each symptom dimension, the algorithm still yields negative scores. In both variants there is no demonstration that the scale has interval properties (and clearly not ratio properties). A confounding factor is that these score are typically presented on a scale with interval properties which gives a false impression that the utilities actually have interval properties (i.e., distances are known rather than unknown).

The HUII Mk3 is different, classifying health state symptoms and responses in a multiplicative algorithm. The HUI Mk3 is the next most often quoted utility measure. This scoring formulation is based on standard gamble (SG) utilities from a community survey which are ordinal raw scores. The instrument questionnaire captures 8 health dimensions with 5 or 6 responses defined for each. Again, the utility score is capped at unity (1 = perfect health) with decrements defined by the multiplicative scoring equation. If we consider the most disadvantageous health state described by the worst score for the eight health dimensions (vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain) the equation is (where u = utility)and a base score of 1.37 which is then reduced by multiplying it by the product of eight SG based tariffs on a presumptive ratio scale of 0 to 1 (vision = 0.61; hearing = 0.61; speech = 0.68; ambulation = 0.58; dexterity = 0.56; emotion = 0.46; cognition = 0.45 and pain = 0.55) but with a true range of -0.362 to 1.

$$U = 1.37(0.61x0.61x0.68x0.58x0.56x0.46x0.43x0.55)-0.37 = -0.362$$

Again, the most disadvantaged health states yield a negative utility. Neither instrument restricts utilities to a 0 = dead (or unconscious) and unity range with ratio properties (i.e., death or unconsciousness represents a true zero – ‘An undiscovered country whose bourne no travelers return’ …. unless you are unconscious and are resuscitated. Both instruments allow states worse than death. Note also this is a multiattribute scale which means we have no idea, in the aggregate ordinal scores reported, what the component changes in that score is contributing). This defies the standards for fundamental measurement in failing to be dimensionally homogeneous. That is, following the physical science we should be measuring one attribute at a time.

HAEM-A-QOL: AN ORDINAL INSTRUMENT

While it may come as a surprise to many, the Haem-A-QoL instrument has only ordinal properties; it is not designed (by default) to measure response to therapy [23] [24]. Announced in 2005, the Haem-A-QoL was intended to be a multiattribute PROM to report on the health related quality of life of adults with hemophilia A or B. The instrument captures 47 items in 10 domains with each domain comprising 3 to 8 items. All responses are based on a 5 point Likert scale (range 1 = never to 5 = all the time). There is an option in certain domains for a 'not applicable' response. Adding up, domain score and total score are transformed to a 0 – 100 scale, with higher scores indicating greater impairment.

Unfortunately, this scoring system fails the standards required for fundamental measurement. Apart from the fact that, even though Rasch Measurement Theory (RMT) has 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, with 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 an interval, rather than an ordinal scale.

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 47-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.

It is also apparent that, in utilizing Likert scales as if they had ratio properties, the authors (and subsequent users) of the Haem-A-QoL evidence report also overlooked the issue of dimensional homogeneity [25]. 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 with ordinal properties.

The use of composite indices or multidimensional scores are common in clinical trials (e.g., as composite endpoints) and in patient reported outcome measures. Ordinal composite measures predominate in assessing therapy response and, at a more global level, health system performance. They typically lack any rationale for the various attributes that are 'mashed' together, a coherent construct theory, or the weights that are attached to generate a composite score. The resulting score or index is all too often meaningless as there is no discussion as to what is driving the aggregate score. Nor is there any account taken of the need for the various composite items each to have unidimensional properties which should be independent of each other.

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* [26]*.* 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, as in the case of the Haem-A-QoL, remain ordinal scales. Attempts to create responder definitions for specific domains for changes in scores ('notable improvements') is simply a waste of time. This does not suggest that the Haem-A-QoL should be abandoned. While it lacks the required, as a minimum, interval properties there is no reason clinicians might not abandon the axioms of fundamental measurement and use Haem-A-QoL scores as a crude measure of something.

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. Hence the Haem-A-QoL (in common with the EQ-5D-3L) lacks dimensional homogeneity. In mathematics all components of an equation must have the same degree of value or quantities of the same base units on both sides; only quantities having the same dimension may be compared, equated, added or subtracted. The Haem-A-QoL and the majority of PROMs fall at the first hurdle. They are a 'mashup' of ordinal scales.

QUALITY OF LIFE IN HEMOPHILIA A

If quality of life is a key endpoint in hemophilia A studies then we have a way forward: to focus on needs fulfillment and the development of a quality of life instrument that has the required interval response properties. This, however, still means abandoning the cost-per-incremental lifetime QALY framework. ICER type lifetime modeling must be abandoned; it is just pseudoscience. At best the needs fulfillment instrument will have interval properties; a true measure of therapy response. It will also provide a basis for credible quality of life claims where the instrument is developed for the needs of target patient populations in disease areas. These claims can be empirically assessed and replicated.

Needs fulfillment as the focus in quality of life was first proposed some 30 years ago. It was pointed out that measures of health related quality of life, such as the EQ-5D-3L, determine the presence of symptoms and functional ability [27]. The instruments were not designed, the axioms of fundamental measurement aside, to determine the value to patients of alternative health states. Symptom change and functional mobility change are not ends; as judged by patients and caregivers, they are a means to fulfill human needs. The needs model focuses on the extent to which human needs are fulfilled through disease interventions. It is a patient-centric approach.

Since the mid-1990s a number of disease specific needs models have been developed through the application of RMT. This creates an instrument that is unidimensional or dimensionally homogeneous with interval or invariance of comparison properties. These properties are required and were recognized from the start. The instruments yield a single score and provide a meaningful measure of response to therapy by both patients and caregivers; a measure of value that can be applied in real world evidence studies. Clearly, this represents a new paradigm in value assessment and one that rejects completely the discredited I-QALY reference case imaginary world paradigm. There is no needs fulfillment instrument in hemophilia A.

CONCLUSION: DOES ICER HAVE A FUTURE?

Technology assessment took the wrong road 30 years ago with the decision by leaders in the field to create evidence for cost-effectiveness by simulation and assumption rather than focusing on evidence platforms to support claims assessment. The choice of the I-QALY, driven by ordinal generic utilities, not only compounded this unfortunate decision but encouraged the publication of thousands of cost-per-QALY technology assessments that were, in reliance on the I-QALY, nonsense. The draft report on hemophilia A by the modelling group at the University of Illinois is just the latest addition to this sad collection. This joins modelling groups, not only in university centers but also in manufacturers, who have focused on the modelling with recognizing the standards of normal science, compounded by an ignorance of measurement theory.

We have to put this pseudoscience paradigm behind us. Fundamentally, if evidence is limited at product launch then manufacturers proposing claims for their product should provide protocols to detail how these are to be evaluated. We require a coherent new paradigm to support claims creation and assessment. One possible framework is provided by the recently released Version 3.0 of the Minnesota Guidelines for Formulary Assessment [28] Evidence must be created through a process of trial and error, not constructed by modelling groups who not only fail to recognize the limitations imposed by the axioms of fundamental measurement but who are wedded to a deep belief in constructing imaginary claims. It is not enough to claim, as many do, that they have a 'heavenly' dispensation from the discovery of new evidence, putting aside the standards of normal science and the axioms of fundamental measurement. Truth for these model builders is created not discovered. The fact is that they have been given rein to develop reference type models, not because senior executives have any idea of the merits of this approach, but because everyone else does it.

Rejecting the I-QALY paradigm will claim many scalps. The incontrovertible fact that the EQ-5D-3L measure (as an exemplar) yields only ordinal scores will probably cause some disquiet. For those who are aware of the required measurement properties this is the obstacle. The fact that the I-QALY utility scale disallows any attempt to create and present the QALY is the more disconcerting outcome. After all, those pharmacists and other involved in model building for selected manufacturers face a questionable future for their product. Can they claim immunity from the standards of normal science? To do so would be laughable: the claim that the distinguishing feature of health technology assessment is the production of non-evaluable, let alone credible claims for competing therapies resting on a denial of the axioms of fundamental measurement. The answer is obvious: to reject the creation of I-QALY models to support formulary evaluations and close down groups committed to their creation. ICER does not have a future.

It is recognized that, from a business and reputational perspective, ICER is determined to maintain the illusion of lifetime cost-per-QALY creations. On the other hand there is always credibility to admitting that you may have misled decision makers, manufacturers and others in believing that the QALY and your claims for pricing and resource allocation are intellectually robust.

If ICER has a deeply held belief in the mystical ratio scale without a 'true zero' or a scale without negative values than it should make this clear to its model builders in the various university groups who apparently have no idea. Consumers of ICER reports should also be apprised of the existence or belief in this new axiom or dimension in measurement theory. A proof would be of interest, but it seems a waste of time to ask for it.

REFERENCES

[1] Rind DM, Walton SM, Agboola F, Herron-Smith S, Quach D, Chapman R, Pearson SD, Bradt P. Valoctocogene Roxaparvovec and Emicizumab for Hemophilia A: Effectiveness and Value: Draft Evidence Report. Institute for Clinical and Economic Review, August 26, 2020. https://icer-review.org/material/hemophilia-a-update-draft-evidence-report/

[2] ICER. Emicizumab for Hemophilia A with Inhibitors: Effectiveness and Value. Final Evidence Report. April 16, 2018. https://icer-review.org/wp-content/uploads/2017/08/ICER_Hemophilia_Final_Evidence_Report_041618.pdf

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

[4] 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

[5] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *Inov Pharm*. 2020;11(1):No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348

[6] O'Hara J, Walsh S, Camp C, et al. The impact of severe haemophilia and the presence of target joints on health-related quality-of-life. *Health Qual Life Outcomes.* 2018;16(1):84

[7] O'Hara J, Walsh S, Camp C et al. The relationship between target joints and direct resource use in severe hemophilia. *Health Econ Rev*. 2018;8:1

[8] Neufeld EJ, Recht M, Sabio H. Effect of acute bleeding on daily quality of life assessments in patients with congenital hemophilia with inhibitors and their families: Observations from the Dosing Study in Hemophilia. *Value Health*. 2012; 18:916-925

[9] Fischer K, de Kleijn P, Negrier C, et al. The association of haemophilic arthropathy with health related quality of life: a post hoc analysis. *Haemophilia*. 2016;22(6):833-840.

[10] Ballal RD, Botteman MF, Foley I, Stephens JM, Wilke CT, Joshi AV. Economic evaluation of major knee surgery with recombinant activated factor VII in hemophilia patients with high titer inhibitors and advanced knee arthropathy: exploratory results via literature-based modeling. *Curr Med Res Opin*. 2008;24(3):753-768

[11] Naraine V, Risebrough N, Oh P, et al. Health-related quality-of-life treatments for severe haemophilia: utility measurements using the Standard Gamble technique. *Haemophilia*. 2002;8(2):112-120.

[12] Langley P. 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). *InovPharm*. 2020;11(2): No 17
https://pubs.lib.umn.edu/index.php/innovations/article/view/3330/2533
https://doi.org/10.24926/iip.v11i3.3330

[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] 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

[15]Rind DM, Walton SM, Agboola F, Herron-Smith S, Quach D, Chapman R, Pearson SD, Bradt P. Valoctocogene Roxaparvovec and Emicizumab for Hemophilia A: Effectiveness and Value; Evidence Report. Institute for Clinical and Economic Review, October 16, 2020. https://icerreview.org/material/hemophilia-a-update-evidence-report/

[16] 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

[17] 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

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

[19] Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. *Arch Pjys Med Rehabil*. 1989. 70:308-312

[20] Grimby G, Tennant A,m Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? *J Rehabikl Med*. 2012; 44:97-98

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

[23] von Mackensen S, Gringeri A & the Haem-A-QoL study Group. Health-related Quality of Life in Adult Patients with Haemophilia – Assessment with a New Disease-specific Questionnaire (Haem-A-QoL). *J Thrombosis Haemostasis*. 2005;3(Sup1):P0813.

[24] von Mackensen S, Gringeri A. Quality of Life in Hemophilia. In: Handbook of Disease Burdens and Quality of Life Measures. Heidelberg: Springer; 2009;1910-1.

[25] 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

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

[27] McKenna SP, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21)5):474-80

[28] Langley P. Guidelines for Formulary Evaluations [Proposed]. Program in Social and Administrative Pharmacy, College of Pharmacy, University of Minnesota. Version 3.0. October 2020. https://www.maimonresearch.net/minnesota-guidelines/