MAIMON WORKING PAPERS

Working Paper No. 7 March 2020

WHERE IGNORANCE IS BLISS: THE ICER DRAFT EVIDENCE REPORT MODELING IMAGINARY WORLDS FOR OBETICHOLIC ACID FOR THE TREATMENT OF NONALCOHOLIC STEATOHEPATITIS (NASH) WITH FIBROSIS

Paul C Langley, PhD
Adjunct Professor, College of Pharmacy, University of Minnesota

Abstract
A number of commentaries have been published over the past 4 years by the present author on the manifest flaws in the reference case value assessment framework of the Institute for Economic and Clinical Reviews. The recent release of a draft evidence report on Nonalcoholic Steatohepatitis (NASH) with fibrosis gives an opportunity, as part of the public comment process, to attempt to ascertain ICER's views on the criticisms of their commitment to constructing imaginary worlds and the application of EQ-5D utilities to construct QALYs. A series of detailed questions have been submitted to ICER. A similar approach was taken with sickle cell disease. In the case of sickle cell disease ICER's response indicated quite clearly that they were not interested in any critique of the merits of their reference case methodology. Their argument was quite simple: it's what everyone else does. They apparently had no concept of the role of the scientific method, hypothesis testing, or of the dubious role of health technology assessment in focusing on the fabrication of 'approximate information'. The purpose of this commentary is to provide an overview of the arguments against the ICER modeled recommendations for NASH pricing. The model presented by ICER fails to meet the standards of normal science. It is irrelevant to formulary decisions. We should reject the ICER approach, focusing instead on disease specific, patient centric measures that capture quality of life from both patient and caregiver perspectives.

Keywords: imaginary worlds, NASH with fibrosis, ICER, pseudoscience, nonsense claims, nonsense recommendations

---

Introduction

Insinuate: to introduce by stealthy, smooth, or artful means (Merriam Webster)

Over the past few years the Institute for Clinical and Economic Review (ICER) has attempted to insinuate itself as the principal arbiter for value assessments in the US. The ICER business model is built around the construction of lifetime imaginary simulations which claim to provide a framework relevant to health system decision makers for pricing and access with pharmaceutical products and devices. As detailed in a recent review of the ICER value assessment framework, the ICER modeling approach fails to meet the standards of normal science; the discovery of new facts [1]. It is best characterized as pseudoscience (i.e., bunk). Constructing imaginary worlds to support pricing and access recommendations has certainly characterized health technology assessment of the past 30 plus years. Indeed, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) makes clear that it is not interested in hypothesis testing or the discovery of new facts in treatment impact [2]. ISPOR sees its principal role in generating 'approximate information'; imaginary world evidence, created by its focus on lifetime incremental cost-per-quality adjusted life year (QALY) estimates and willingness to pay thresholds, in contrast to real world evidence where meaningful claims for therapy impact and quality of life in disease areas can be evaluated from patient-centric evidence platforms.

The purpose of the present commentary is to point to the manifest flaws in the latest attempt by ICER to fabricate imaginary recommendations for pricing and access for obeticholic acid for the treatment of nonalcoholic steatohepatitis with fibrosis (NASH) [3]. This is a critical issue as ICER's imaginary recommendations can ensure that the access to new therapies, in this case for NASH, is barred to those most in need. ICER has the responsibility for defending its position; not only for pricing recommendations but for denying access to new therapies. Unfortunately, irrespective of ICERs claim that it adheres to 'gold standard' techniques in its fabrication of imaginary cost-per-QALY worlds to support its revelations and recommendations; its methodology is fatally flawed. Yet ICER perseveres in a program of making recommendations for price discounting and access on a value assessment framework that defy the standards of normal science. This case against ICER rests on two arguments: first, the failure to accept the standards of normal science in modeling imaginary claims for competing therapeutic products and, second, the willingness of ICER to make

unsupported assumptions in the creation of imaginary QALYs. In the latter case the fatal flaw is to ignore the standards of fundamental measurement by fabricating measures of quality adjusted life years that rely on ordinal utility manifest scores. Consequent QALY estimates and claims are nonsense.

The Imaginary World of the NASH Washington Model[1]

Imaginary: existing only in the imagination (Oxford Dictionaries)

Imaginary worlds can be compelling; from Peter Pan to Harry Potter millions of children (and adults) have been enthralled with their creativity and their identification with the leading characters. Health technology assessment, as understood and proselytized by groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has been in the forefront in advocating practice standards for fantasy creations. ISPR appears committed, and this is echoed in guidelines such as the Canadian, that the focus of health technology assessment is on generating approximate information, or more accurately, imaginary approximate information or disinformation, rather than testing hypotheses: *Economic evaluations are designed to inform decisions. As such they are distinct from conventional research activities, which are designed to test hypotheses* [4]. The ICER Washington model has followed this in presenting its 'approximate information' claims, rejecting any notion of hypothesis testing for the modeled claims.

The imaginary simulated world fabricated for the Washington NASH value assessment framework has as its primary objective to estimate the lifetime cost-effectiveness of obeticholic acid compared to standard of care for adults with NASH with fibrosis. The proposed model uses a Markov structure composed of two cardiovascular event history submodels with equivalent liver disease-specific state transition probabilities. Each submodel allowed transitions among a number of health states or events: no fibrosis, discrete fibrosis (F1-F3) stages, compensated cirrhosis (F4), decompensated cirrhosis, hepatocellular carcinoma, post level transplant and, the ultimate absorbing state, death. Outcomes from the model were life years, equal value life years, quality adjusted life years (QALYs), cardiovascular events, hepatic complications and total lifetime costs. None of these outcomes were presented in a form that allows empirical assessment; they were not intended to meet standards for credibility and evaluation. The target hypothetical population for the model were patients with NASH fibrosis stages F2 and F3 being treated with either obeticholic acid 25mg or standard of care. The demographic characteristics of the target population were matched to the REGENERATE trial [5].

The intricacies of the Washington NASH SCD model are not of concern here. After all, although data to support the model are drawn from REGENERATE trial together with data drawn from a variety of other sources, the purpose is to provide support for the assumptions; the model is driven entirely by assumption. Obviously, this model is one of many, if not a multiverse of different models, to support competing claims for therapy impacts in NASH. The key point, from the perspective of an imaginary construct with lifetime non-credible (and obviously non-evaluable) claims to support pricing and access recommendations is the construction of QALYS, lifetime QALY estimates for the various assumed and modeled treatment pathways, lifetime costs and, *the piece de resistance*, incremental cost-per-QALY estimates to support threshold analysis and the much awaited ICER recommendations for price discounting from WAC. Our focus, therefore, is on the QALY and whether the model assumptions regarding how utilities are 'discovered' and assigned and QALYs created make sense from the perspective of normal science.

Lifetime cost-per-QALY claims for any imaginary value framework constructs rests on (i) claims for increment benefits expressed as QALYs and (ii) the lifetime direct medical costs. The base case results for the NASH model yielded a lifetime incremental QALY of 0.50 years (obeticholic acid 10.13 QALYS vs, 9.63 QALYs for standard of care). Total costs were respectively $1,291, 000 (discounted at 3%) and $419,000. This yields in turn a lifetime cost per QALY gained of $1,756,000. As will be demonstrated, this claim is complete nonsense. The Washington NASH model builders cannot make any claim for incremental QALYs or costs per QALY gained against standard of care because the QALY construct is a mathematical absurdity as it neglects the requirements of fundamental measurement.

The Standards of Normal Science

Pseudoscience: a collection of beliefs or practices mistakenly regarded as being based on scientific method (Oxford Dictionaries)

[1] The model was developed by the Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, University of Washington, Seattle WA

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. This applies in NASH just as it does in other therapies. 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, 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., bunk).

Approximate Information (or Disinformation)

Approximate: close to the actual, but not completely accurate or exact (Oxford Dictionaries)

To add to this litany of disbelief, it is worth emphasizing that ISPOR, as ICER's methodological mentor, explicitly disavows hypothesis testing as a core activity in health technology assessment. The primary role of health technology assessment 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 ICER modeling, any known reference point for 'true information' to judge approximation. How close are we? Is the truth out there? 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.

Choice of Assumptions

Assumptions: a thing that is accepted as true or as certain to happen, without proof (Oxford Dictionaries)

The ICER claim to fame is the ability of its consultants to construct or fabricate an imaginary world that sets the stage for value impact over 10, 20 or 30 years in the future. In the Washington NASH model the number of assumptions made to support the various simulations and their scenario progeny across the three therapies is truly awesome; some come from the literature, others are pure guesswork. 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]. No, Virginia, all swans are not white. You may have seen only English swans, but on my last QANTAS vacation in Western Australia, I saw black swans. In similar vein, we cannot assume that if the REGENERATE trial were to be repeated, that after any number of successful repeats, we would again replicate the results.

Achilles, Utilities and QALYs

Achilles Heel: a weakness or vulnerable point (Oxford Dictionaries)

QALYS are the Achilles heel of the ICER construction and belief in imaginary reference case lifetime worlds; exeunt QALYs and the fantasy edifice collapses. Apart from their use in the ICER contribution to the science fiction literature, QALYs can only survive if the measure is credible, evaluable and replicable. The QALY constructed by ICER in the Washington NASH model meets none of these criteria. In fact, the construct itself is nonsensical. The issue is one of failing to recognize the importance of fundamental measurement standards. In the Washington NASH model the utilities selected, and there is only one reference cited in NASH in the ICER report for utilities, in this case the EQ-5D-3L, self-assessment in a hospital environment, the balance of utilities are for other, possibly similar, disease states [11]. 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 interval scale of 0 to 1 (death to perfect health). Combining the two, multiplying time by utility is assumed to produce a QALY. In the ICER imaginary NASH world these are combined to produce QALYs for the modeled life. However, before considering the EQ-5D-3L utility that is central to the imaginary NASH simulation, a brief digression in measurement theory and its application to instrument development in the social sciences is in order. This is important because the ICER academic groups building imaginary worlds seem oblivious to these requirements.

Briefly, there are four measurement scales (putting to one side conjoint simultaneous measurement which underpins Rasch measurement theory [12]). These scales are nominal, ordinal, interval and ratio. The argument presented here is that the EQ-5D-3L generates ordinal 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. The result is that to construct QALYs by assuming the EQ-5D-3L has ratio properties is a mathematical nonsense. From an imaginary modeling perspective is should emphasized that: (i) ordinal scales (the EQ-5D-3L) only allow median and modal values to be presented; (ii) interval scales allow addition and subtraction (no true zero) and (iii) ratio scales with a true zero (i.e., no negative values) allow multiplication and division (i.e., distance from a true zero). Unfortunately, the EQ-5D-3L tariff algorithm has no demonstrable interval measurement properties (with odd ceiling and floor effects) as well as allowing negative utilities (below a true zero). 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.

As there is a firm belief in the memetic ISPOR mystery of the EQ-5D-3L having a 'true zero' we also need to combine this with the 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 reduce confidence in predictions and the score is a less useful summary. In Rasch modeling, estimates of item difficulty and person ability are meaningful if every question 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 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. But we just lump them together and attach community preference weights to the ordinal responses and assume the result is a ratio scale with a true zero even though we can have negative scores. While ICER maintains the EQ-5D has interval properties, which they cannot demonstrate, they need to assume ratio properties to create QALYs (multiplying imaginary time spent in the modeled disease by the assumed ratio utility score).

Apart from the lack of a single attribute (e.g., needs-based quality of life utilizing Rasch modeling) ICER does not appear to recognize that the responses to the five symptom levels will vary by disease state (e.g., no problem response for mobility vs. major problem in another disease state). If the EQ-5D-3L is used to create imaginary QALYs in that disease state then ICER has to demonstrate that the ratio property holds for that application. This has been ignored in all ICER value assessment models. Unfortunately, an EQ-5D-3L health state defined by 5 symptoms and 3 response levels (e.g., 13333) can easily be shown to have a negative 'utility' or a state worse than death (death scores a zero; without any questionnaire response); in this case 13333 translates to a utility of -0.28. There is no true zero.

The problem for ICER and others using the EQ-5D-3L is that it was not developed to meet the standards for fundamental measurement, in this case for constructing QALYs, a true zero for a ratio scale; let alone a ratio scale with both a true zero and an upper limit of unity (QALYs require time to be multiplied within a range of 0 to 1). It is not clear, if death is zero, but we measure states worse than death, how we would interpret a QALY if the EQ-5D-3L score is negative (the lowest possible EQ-5D tariff is -0.59)?

Can we have negative QALYs when we calculate aggregate QALYs over a hypothetical lifetime? Apparently, users of the EQ-5D-3L can take refuge in the observation that there are only a handful of respondents who would score a negative EQ-5D so they can be assumed to be 'at deaths door' (but not dead as they are responding to the questionnaire); perhaps a utility of 0.000001? We cannot, of course, multiply a time spent in a disease state by zero if someone is alive. One view is that at a population level we can ignore the annoying negative utility issue because it is infrequent; the same may not apply by disease states; notably for rare diseases and those with major disabilities. Can ICER demonstrate for the application of the EQ-5D-3L in SCD that we have a minimal and assumed irrelevant proportion of patients with negative utilities in these disease states? And what about the utilities of caregivers?

The situation becomes even more bizarre when we move from the EQ-5D-3L to the EQ-5D-5L (introduced in 2009) where there are 5 response levels. Individuals in the same health state will have assigned preference scores to generate a score for five response levels quite different from three response levels (including states worse than death). If the ICER sickle cell model applied 5L weights then the count of QALYs (including presumably negative ones) would be different.

Even if 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 the belief in fabricated imaginary evidence. The ICER value assessment framework would still fail the demarcation test as pseudoscience (i.e., bunk). 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 Rasch Measurement Theory (RMT) focus on the response to interventions on a constructed interval scale rather than attempting to go the further step of creating instruments which have ratio properties (i.e., a true zero) [14] [15] [16]. The EQ-5D-3L horse has well and truly bolted.

Conclusion: Next Steps

The fact that the application of utility values, from a variety of sources to create QALYS, fails the standards of fundamental measurement should be sufficient to show that the ICER reference case model for NASH (and all previous evidence based disease claims) should be rejected; unfortunately, this will not deter ICER. The company has too much invested in its claim as the US technology assessment arbiter of emerging products and technologies. After all, it would be embarrassing to admit that its recommendations for pricing and access are, to say the least, nonsensical, and that the ICER value assessment framework is more appropriately classified with intelligent design than natural selection.

In NASH, it will be up to the manufacturers to make the case for ignoring ICER to health system decision makers. They will have to offer an alternative approach to evaluating the 'value' of their products. Previous commentaries have proposed that rather than focusing on QALYs, manufacturers should direct their activities to claims based on disease specific QoL instruments. Since the mid-1990s disease specific (both patient and caregiver) instruments have been developed with needs fulfillment as the latent unidimensional construct. The instruments meet the required standards of Rasch measurement theory to create an instrument that meets interval measurement standards to assess response to therapy: does a new therapy contribute to patients needs being more effectively met in a disease state? An instrument that meets these standards should be considered in NASH.

Adopting a disease specific, patient centric instrument (together with a caregiver instrument for pediatric NASH patients) provides claims that are credible, evaluable and replicable. It is a simple index of response to therapy and can be an integral part of evidence platforms such as registries in NASH. The fact is we don't need ICER (or any other group) to spend eight months from conception through gestation of an imaginary construct that fails the standards of normal science. A commitment to fantasy creations that is, surprisingly, supported financially by manufacturers; they should know better. A return to the standards of normal science, to the discovery of new facts in the treatment and response to therapies in diseases such as NASH would be a welcome respite from, and antidote to ICER.

References

[1] Langley PC. Nonsense on Stilts – Part 1: The ICER 2020-2023 Value Assessment Framework for Constructing Imaginary Worlds. *InovPharm*. 2020;11(1):No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444

[2] 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;21:119-25

[3] ICER. Obeticholic Acid for the Treatment of Nonalcoholic Steatohepatitis with Fibrosis. Draft Evidence Report. 19 March 2020 https://icer-review.org/wp-content/uploads/2019/10/ICER_NASH_Draft_Evidence_Report_03192020.pdf

[4]Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the economic evaluation of health technoloigies. Canada: Ottawa, 2016

[5]Younossi ZM, Ratziu V, Loomba R, Obeticholic acid for the treatment of non-alcoholic steatohepatitis: interim analysis from a multicentre, randomised, placebo-controlled phase 3 trial. *Lancet*. 2019 Dec 14;394(10215):2184-2196

[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] Anie K, Grocott H, White L et al. Patient self-assessment of hospital pain, mood and health related quality of life in adults with sickle cell disease. *BMJ Open*. 2012;2:e001274

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

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

[15]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

[16] 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