MAIMON WORKING PAPERS No. 17 22 JULY 2020

IMPOSSIBLE QALYs AND THE ICER EVIDENCE REPORT FOR NASH: AVOIDING NONSENSICAL RECOMMENDATIONS

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

ABSTRACT

Despite the fact that its reference case lifetime simulation model rests on untenable assumptions, The Institute for Clinical and Economic Review (ICER) perseveres in its commitment to final evidence reports for product value assessment. Previous reports and papers have made the case against ICER cost-per QALY models abundantly clear. The QALY (or more appropriately labeled I-QALY for impossible) is a mathematically impossible construct. It rests on ICER's ignorance of the axioms of fundamental measurement. Utilities are ordinal scales; they cannot be used to create QALYs as we cannot multiply time by an ordinal score. The latest ICER evidence report for the treatment of Nonalcoholic Steatohepatitis with Fibrosis (NASH), with obeticholic acid (OCA), released on 21 July 2020, falls squarely in the nonsense modeling category. The purpose here is to briefly make the case for rejecting the ICER model and its recommendations and advise manufacturers who have late stage products in the NASH pipeline to ignore ICER as an analytical dead end and unnecessary distraction.

INTRODCTION

Evidence for product impact in therapy areas, notably in respect of claims for cost-effectiveness, is discovered not invented [1]. Unfortunately, in the area of health technology assessment, a decision was made by leaders in the field some 30 years ago to pursue the invention of claims for cost-effectiveness [2]. The reason for this approach is obvious: at product launch evidence for costs and outcomes between competing therapies is limited. The options are to either establish a research program to generate real world evidence or to put caution to the winds and propose a model simulation that, on a series of assumptions, incents or creates claims for cost-effectiveness. Alone among the social sciences, therefore, rather than the discovery of new evidence, technology assessment rests on the creation of claims which are neither credible nor evaluable.

At the same time, not only was the notion of hypothesis testing rejected in favor of creating approximate information, but the leaders in the field hit upon the potential for a single, global metric as the centerpiece in the invented information odyssey. The metric was the quality

adjusted life year: the generic QALY as the measure of health related quality of life (HRQOL). The standard of analysis was to be a lifetime incremental invented cost-per-QALY model. The value of a product was to be judged by comparing the imaginary modeled cost-per-QALY against various threshold cost-per-QALY standards.

The case against inventing evidence through lifetime modeling simulations has been made on numerous occasions; most recently in a commentary on the National Pharmaceutical Council's embrace of imaginary information constructs [3]. The purpose of this present report is to provide an assessment of the recent (July 21, 2020) evidence report by the Institute for Clinical and Economic Review (ICER) for the treatment of nonalcoholic steatohepatitis with fibrosis (NASH) with obeticholic acid (OCA) [4]. An earlier evidence report on obeticholic acid was published by ICER in 2016 [5]; the focus here is on the latest report.

THE NASH OBETICHOLIC ACID IMAGINARY WORLD

The NASH model, developed by the by the University of Washington, School of Pharmacy, Modeling Group, Seattle, WA, is characteristic of the creating approximate non-evaluable information genre. In common with similar models the Washington group have contracted to develop for NICE, the NASH model uses a Markov model structure. This comprises two cardiovascular (CV) event history submodels with equivalent liver disease-specific state transition probabilities. Each submodel allows for transitions among no fibrosis (F0) and discrete fibrosis (F1-F3) stages, compensated cirrhosis (F4), decompensated cirrhosis, hepatocellular carcinoma (HCC), post-liver transplant, and death; the costs and health impacts of undergoing liver transplant were assessed within the transition to post-liver transplant. Patients were able to transition from any of the alive health states to death from all causes including compensated cirrhosis, decompensated cirrhosis, HCC, liver transplant, CV events, or background mortality. The transition from the first submodel (no prior CV event) to the second submodel (prior CV event) was driven by the first occurrence of a nonfatal CV event; the costs, quality of life, and survival of first CV events were assessed with the transition between submodels. NASH patients who entered the prior CV event submodel were assumed to be at increased risk for recurrent CV.

The modeled base-case analysis utilized a hypothetical cohort of patients with NASH fibrosis stages 2 and 3 in the U.S. being treated with either OCA 25mg or standard care, using demographic characteristics from the REGENERATE trial. Imaginary outcomes from the model included discounted lifetime drug costs and total treatment costs, likelihood of progression to advanced liver disease, cardiovascular events, life years equal value and I-QALYS compared to standard of care. Of course, given the lifetime model structure, discounting and range of assumptions any outcome claims would be impossible to evaluate. Indeed, they were never intended to support hypothesis testing; the focus was on approximate comparative information created from one of a possible multiverse of competing models. The analysis adds nothing to our understanding of the long term or even lifetime impact of obeticholic acid

versus standard of care. There was no attempt to provide suggestions for a research program to discover new facts.

EMBRACING THE IMPOSSIBLE

The ICER report on obeticholic acid for the treatment of NASH is clearly in the pseudoscience camp. The ICER reference case framework fails, as has been noted on numerous occasions, to meet the evidence standard for normal science: all claims for products should be credible, evaluable and replicable. This has been the standard distinguishing the scientific method for 400 years. The ICER model is more properly placed alongside intelligent design rather than natural selection [6].

However, with the weight of opinion in favor of imaginary constructs, with advocates such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), ICER takes refuge in the belief that truth is consensus.

A review of the ICER NASH model makes it quite clear that the choice of assumption is paramount. Assumptions are chosen from the literature to populate the NASH model framework with the belief that looking 10, 20 or 30 years into the future, ICER can generate approximate information to support formulary decisions and its pricing recommendations for obeticholic acid.

Unfortunately, ICER misses an obvious point – the fallacy of induction. There is no reason to believe that assumptions that have held in the past will continue to hold in the future. We can't secure assumptions as 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* [7].

Attempting to escape the induction problem (although ICER has apparently not heard of it) through shrouding simulated claims with one and two way sensitivity analyses and the absurd probabilistic sensitivity outcome claims, is no escape.

Certainly assumptions have their place in modelling, but in models which generate credible, evaluable and replicable claims. If evaluable claims are falsified, then we can reassess our assumptions as part of a modeling review to discover new facts. The ICER paradigm is to create new yet imaginary 'facts'; facts which have no possibility of ever being evaluated

MEASUREMENT THEORY

ICER's failure to appreciate the standards of normal science, is made even more apparent in its ignorance of the axioms of fundamental measurement and the limitations imposed by the various measurement scales. If for no other reason, this profound ignorance renders the entire NASH modeling exercise and the recommendations for pricing and access ridiculous.

Briefly, the typical multiattribute (e.g. the EQ-5D-3L) instrument produces only ordinal scores. That is, the utility values have magnitude, an ordered relationship to each other but an unknown distance between the ranks. We have known this from the contribution by Stevens in the 1940s, the recognition of the role of Rasch measurement theory and a number of commentaries over the past 30 plus years [8] [9]. The ordinal scale can only generate medians and modes, supporting nonparametric statistics. What the utility values lack are interval properties, where there are equal intervals, supporting the mathematical operations of addition and subtraction and, at a higher order, ratio properties. A scale with ratio properties has interval properties but, with a true zero, can also support multiplication and division. If you want to create a QALY then you have to have a ratio scale. The EQ-5D-3L demonstrably fails on this criterion as the utility values can range from -0.59 to 1.0. The former utility (-0.59) indicates, apparently, a state worse than death (death is arbitrarily assigned zero) with 1.0 perfect health. Thus, we can have negative QALYs. This is a common feature of all multiattribute utility instruments.

What ICER fails (or refuses) to recognize is that the I-QALY is a mathematically impossible construct. A ordinal scale will not support multiplication, which is required if the utility is to discount time spent in a disease stage to create an I-QALY equivalent.

NONSENSENSICAL NASH QALYs

To illustrate the absurdity of the ICER reference case simulation, we only have to consider the utilities presented in Table 5.8. The utilities presented here are drawn from three sources. The first, the Global Assessment of the Impact of NASH (GAIN) study, presents EQ-5D-5L utilities for two NASH fibrosis stages 0-2 and 3 [10]. The utilities are, respectively. 0.76 and 0.73. There is, it should be noted, no problem in presenting these multiattribute scores. All it tells us is the utility of each stage but says nothing about the distance between these scores. The implied difference of 0.03 has no meaning. Putting these on an equal interval number line might give the impression that we know the distance between them; even the distance from zero. But this is arbitrary. All we know is that the more serious the condition the lower the utility, but not by how much. As this is an ordinal and not a cardinal or interval scale attempts to create confidence intervals (as both the original authors and the ICER academic group do) is mathematical nonsense. Even so, a wag might point out that if we accept the confidence interval argument, the overlap between the two utilities means they are not different from each other anyway.

The second set of utilities are from a clinical and economic burden of disease study where they were obtained directly from patients using Short Form-6D scores [11]. Although unremarked by the authors of the ICER model, the Short Form-6D multiattribute utility instrument (with responses first from the SF-36 or SF-12 questionnaire) is on different conceptual basis from the EQ-5D-5L and EQ-5D-3L. The utilities created are not from the same system, as the ICER model would imply. This has been obvious and reported on for decades[12] . Our wag would also note

that none of the utilities appear to be significantly different from each other (if we believed in ordinal confidence intervals).

The third reference point for utilities is a somewhat dated assessment of disutilities for chronic conditions in the US. The point of interest here is that these utilities are based on the EQ-5D-3L instrument. Given that the first reference for utilities by NASH stage are from the EQ-5D-5L, there is the somewhat odd situation where the utilities to populate the cost-per-QALY claims in the NASH model are from three different multiattribute instruments [13]. The EQ-5D-3L and EQ-5D-5L produce quite different utility scores for the same symptoms; the instruments differ only in the number of response levels. This may not, from the Washington groups perspective, be a barrier to creating evidence for obeticholic acid; presumably any utility estimate will suffice as long as no one digs too deeply into the source references. Any port in a utility storm! After all, the model is only an imaginary construct.

Aggregating the Washington I-QALYS over the lifetime of the hypothetical NASH patient cohort yielded a discounted aggregate I-QALY count of 9.92; the corresponding standard of care count was 9.43 I-QALYs. Bearing in mind that creating, aggregating discounting and subtracting I-QALYs violates the axioms of fundamental measurement, means that the difference of 0.49 I-QALYs will yield, given the projected discounted lifetime drug cost, a substantial cost-per-I-QALY; in this case $1.48 million.

While this is clearly complete nonsense, the concern is that the media and health system decision makers will take this at face value, claiming that ICER has brought down from Mt Sinai (or mount I-QALY) golden tablets to make the case for substantial price discounting with obeticholic acid. To this should be added the budget impact assessment where ICER announces that, at the product rice, only a limited number of NASH patients would be able, given ICER nominated budget constraints, to be treated.

Looking to the NASH pipeline, manufacturers will be concerned that the further application of the ICER NASH model and the creation of I-QALYs will have a potentially, and unwarranted, adverse impact on the willingness of investors to support the NASH pipeline product and the willingness of health systems to enter into price negotiations. It is up to manufacturers to point out that this is an absurd situation, resting on a denial of normal science and a determined effort to ignore the axioms of fundamental measurement.

CONCLUSIONS

It is difficult to understand why ICER continues to create cost-per-QALY claims that are clearly nonsensical. The NASH model is merely one example where the model builders are clueless when it comes to the standards of normal science, the role of assumptions and the axioms of fundamental measurement. Presumably, having been trained in the construction of imaginary reference case worlds, it is difficult, as noted in previous studies, to overturn such a strongly held belief system.

It is not just the question of the construct, but in the NASH case the obvious failure to understand that different utility systems will yield different scales. Applying these incompatible ordinal scales to generate I-QALYs, adding them together to create lifetime discounted QALYs leads us further into fairyland. It is not just impossible QALYs (I-QUALs) that are created but the incompatibility of the I-QALY constructs that defies any attempt at imaginary aggregation.

The ICER reference case model for NASH is a disaster. The concern must be that if ICER continues to employ the 'any utility will do' model for pipeline NASH products that manufacturers and health system decision makers must be advised that the ICER NASH model, and all other ICER modeled evidence reports are an analytical dead end and an avoidable distraction.

REFERENCES

[1] Langley P. Langley P. Nonsense on Stilts Part I: The ICER Value Assessment Framework for Constructing Imaginary Worlds. *Inov Pharm*.2020;11(1): No. 15

[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;21:119-123

[3] Langley P. The National Pharmacy Council: Endorsing the construction of imaginary worlds in health technology assessment. *Pharmacy*. 2020; doi:10.3390/pharmacy 8030119

[4] Rind DM, Hansen R, Guzauskas G, Beinfeld M, Chapman R, Bradt P, Pearson SD. Obeticholic Acid for the Treatment of Nonalcoholic Steatohepatitis with Fibrosis: Effectiveness and Value. Institute for Clinical and Economic Review, July 21,2020

[5] Institute for Clinical and Economic Review. Obeticholic Acid for the Treatment of Nonalcoholic Steatohepatitis: Comparative Clinical Effectiveness and Value. ICER. July 26, 2016

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

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

[8] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-80

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

[10] O'Hara J, Finnegan A, Dhillon H, et al. The Cost of Non-Alcoholic Steatohepatitis (NASH) in Europe and USA: The GAIN Study. 2019; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506175. Accessed 1/22/2019, 2019 Preprints with *The Lancet*

[11] Younossi ZM, Blissett D, Blissett R, et al. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. *Hepatology* (Baltimore, Md). 2016;64(5):1577- 1586

[12] Sullivan PW, Ghushchyan V. Preference-Based EQ-5D index scores for chronic conditions in the United States. *Med Decis Making*. 2006;26(4):410-420

[13] 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