PATIENT ACCESS
& AFFORDABILITY
PROJECT

Current Issues

ABANDONING QALYs:
AN UNNECESSARY DISTRACTION IN VALUE ASSESSMENT

A concern often expressed by patient advocacy groups is that the focus by health technology assessment groups on quality adjusted life years (QALYs) is misplaced. The Institute for Economic and Clinical Review (ICER) QALY measure, it is claimed, overlooks and even ignores the patient voice in claims made for competing therapies. This is an entirely valid criticism. Certainly, the QALY in the imaginary worlds created by ICER is an absurd construct. Media attention on ICER pronouncements typically, if not invariably, fail to point out that ICER is an unnecessary distraction; its recommendations lack any claim to being credible. They fail to meet the standards we expect of true science; they are pseudoscience on a par with intelligent design and not natural selection. Here's why.

ICER's Imaginary Worlds

ICER asks us to believe that it is possible to create, by assumption, a model or simulation that, over a projected future timeline of 10, 20, or 30 years, can provide useful information to health decision makers faced with decisions on the pricing and access for products and devices recently approved by the FDA. To do this, ICER sets up a mathematical model that tracks therapy choices and their clinical impact for a hypothetical patient population.[1] Everything is driven by assumption, with ICER asking us to take at face value their claim that this is the most reasonable future: an imaginary but "realistic" simulation. This is clearly nonsense as the so-called ICER reference case, instructions on how to build an imaginary future reality, can be engineered to produce any number of competing models, each coming to different value claims.

The beauty of building imaginary worlds is that the imaginary claims can never be evaluated empirically. We will never know whether ICER's value claims are right or if they are wrong – and in fact, we were never intended to know. ICER has created a situation where its claims and the recommendations it makes are immune to failure – it's a one-horse race! The house always wins!

Understanding Measurement

Making effective decisions in health care requires measures of response to therapy that are robust and precise.[2] There is no reason why we should not aim for the measurement standards that characterize the physical sciences. That is, measures that meet the axioms for fundamental measurement: invariance of comparisons and sufficiency.[3] If these conditions are met, then we have an interval measure of response in which we know the order and the exact differences between item values. Unfortunately, with only a handful of exceptions, response measures that characterize value assessment in health technology assessment are ordinal; we know the order of the values, but not the differences between the values. Treating an ordinal response as though it were an interval or cardinal response is malpractice. Ordinal scales cannot support measures such as the mean and standard deviation; they cannot support change scores or effect sizes.

If we are to meet the general axioms of measurement theory, we need to transition from Classical Test Theory (CTT) to support instrument development to Rasch Measurement Theory (RMT).[4] If our response or patient reported outcome instrument meets RMT standards, then we can assume that it is unidimensional, creating a single score to reflect the latent construct or attribute we are attempting to measure.[5]

Unfortunately, issues of ordinal vs. cardinal measurement appear not to have troubled those committed to the construction of imaginary worlds in health technology assessment. The result? ICER has published literally thousands of reference case incremental cost-per-QALY claims that fail to meet even the most basic requirement for credible, evaluable and replicable claims: the absence of cardinal or interval measurement.

The ICER QALY

The ICER QALY is constructed from (1) utilities and (2) simulated time spent by the average hypothetical patient in different stages of a disease over their hypothesized lifetime (as imagined by ICER). Add to this, for cost-per-QALY claims, the flexibility in selecting the imaginary cost components as some variant of projected direct medical costs or even social costs over the lifetime of the model.

Utilities are considered to be measures of the health status of an individual at a point in time. In the ICER imaginary construct, the utility scores are from what are called generic pre-scored multi-attribute instruments. The scores for each patient are derived from community preference weights attached to each response within each symptom. The patient responds; a prior community preference weight for that response determines its contribution to an overall raw score. This aggregate ordinal raw score is generated by a utility function that combines the response level scores, while constraining them (ideally) to a 0 = death to 1 = perfect health range. Within each multi-attribute system there is a first order function, which assumes response levels are independent of each other plus hybrid versions which may assume interdependence. Irrespective of your choice, they are still ordinal measures.

In the case of the EQ-5D-3L, probably the most popular ordinal measure, there are five health dimensions: mobility, self-care, usual activity, pain/discomfort, and anxiety/depression. The developers selected these, limiting them to five to allow ease of completion. Within each symptom there are three response levels: no problem, some problems, and major problems. This yields 245 health states (defined if the response levels are designated 1, 2 and 3) as, for example, 11221. If we add unconscious or dead, we have 247. The preference weights were created by applying the time trade-off technique to a sample of health states and asking a community random sample to value them.

A revised version issued in 2009 changed the ordinal response levels to five: no problem, slight problems, moderate problems, severe problems and extreme problems. This change reflected concerns over sensitivity. Unfortunately, it resulted in two different ordinal instruments with ongoing efforts to provide, mapping algorithms to translate from one to the other.

Utility measures, the assumed benefits of a health care intervention, are, in constructing the ICER imaginary world, plucked from the literature. Although the gestation period for an ICER evidence report is eight months, ICER has no interest in generating its own utility scores. Everything is second-hand. The utilities that are found are applied to the stages of the disease being modeled, including adverse events within disease stages. The stages are modeled; with the time spent in each stage created by the model and driven by assumption. Combining the utility with the time spent in each stage yields the lifetime QALY estimate. If there are no "available" utilities for that disease stage then either measures may be "interpolated," or other utility scores (e.g., HUIMk3) used in their place. After all, as ICER has stated, different generic utility systems tend to give the same results.[6] Even so, they are all ordinal scores, so the result is no different: manipulating them is logically invalid.

It is worth noting that the symptom categorization and response levels captured by the EQ-5D-3L are "operational measures"; they are clinically focused and are not intended to go beyond the measured impact of a health intervention in its impact on response level within symptom categories. The responses are valued not by the patient, but by the community. The patient voice is muzzled. We have no idea whether those responses are meaningful to the patient. The community preferences are determined from respondents valuing health states, where probably 95% or more have no personal experience (or even experience as a caregiver) with the disease state that ICER is taking upon itself to value and report on therapy options.

There are other ordinal generic measures, all of which would give different responses or raw scores for the same patient with their different health dimensions and different response levels. There is no "correct" ordinal measure; you are free to choose but are advised not to combine them in the same model. This may appear redundant as the end product is still an imaginary construct.

If ICER wishes to persevere with the construction of imaginary worlds to support imaginary recommendations, it has to demonstrate for each utility score introduced into its model that for the disease state in question, the utility can be shown to have acceptable RMT properties. This requires access to the original unit record data, if it exists, to support RMT assessment. It is doubtful that ICER would support this. By default, therefore, we have to assume that RMT criteria are not met. This results in the estimated lifetime QALYs being an absurd product of a cardinal raw score and time spent in a stage of disease. Of course, as the ICER imaginary world is built on assumptions, ICER could announce that it is assuming that the utility score extracted from the literature is, in the absence of any evidence, deemed to be cardinal. If not, The ICER reference case model collapses.

Time spent is the other component of the QALY. Depending on the choice of mathematical simulation and assumptions regarding the structure of the model and further assumptions regarding the likelihood of members of the hypothetical target population transitioning between disease stages, time spent in each stage is created. Death is the final stage: described euphemistically as the 'absorbing' state from which no hypothetical patient returns. Time spent in each stage is projected for 10, 20, or 30 years in the future.

Direct Medical Costs

The role of assumptions is not over. ICER then produces an estimate of the average projected direct medical cost of treatment (with often a variation to capture some aspect of social cost) for the average patient over the hypothetical average lifetime. This involves some imagination and choice as, unlike those with a crystal ball, ICER has no idea what these future costs are likely to be. Indeed, ICER admits that is has no unique crystal ball; it makes no effort to project actual future drug costs.

The QALY Industry

For those who live for imaginary worlds, the number of worlds that can be created is limited only by your imagination. If your business model depends on your ability to create imaginary worlds – the realm of a successful science fiction author perhaps – then any criticisms will be defended. ICER defends its models as "state of the art" in health technology assessment. Respected professional associations, such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), are quite clear on this: health technology assessment is not to test hypotheses regarding therapy response but to develop imaginary worlds as "information" for decision makers.[7] [8]

This position is supported by agencies such as NICE in the UK who mandate the construction of reference case imaginary worlds to support formulary submissions.[9] Unsurprisingly, we now have a tribe of imaginary world referees (typically in academic institutions) who are tasked to peer into their own crystal balls to assess the "merits" of a manufacturer's imaginary world submission.

Inevitably, once a government agency mandates a procedure, the response will be an industry devoted to the construction of QALYs. There is no single QALY; hundreds of different QALYS could be constructed depending upon the whims and the commercial objectives of the developer. Yet ICER gives the impression that theirs is the single "golden" QALY. This may be unintentional, but for those whose treatment options are affected by recommendations based on the ICER QALY, few would appreciate the multiverse of imaginary cost-per-QALY worlds. ICER recommendations are merely one of this (undefined) multiverse of recommendations. Apart from the fatal flaw of trying to represent that ordinal measures can be treated as cardinal measures (after all, it is really an imaginary fiction), the flexibility in QALY claims knows no bounds. Reviews of the health technology assessment journals make this quite clear[10] - and unsurprisingly, the cost-per-QALY claims sponsored by manufacturers tend to support the manufacturer's product and its recommended market entry price.

The US Response

There is no reason for the US to emulate NICE in the UK and follow ICER on its imaginary journey. First, the ICER model, in its attempt to be NICE-lite, fails to meet the standards of normal science. It fails the most basic test of measurement theory, espousing claims and recommendations that fail to be credible, evaluable and replicable. It is pseudoscience, joining intelligent design in the Dover courtroom dock.[11] Second, and more egregiously, it ignores the patient. Indeed, the imaginary world with its community preference score, must ignore the patient voice. The ICER model only "succeeds" because the needs of patients are ignored in the imperative to include multi-attribute community preference ordinal scores at center stage for imaginary QALYs that fail to meet required interval measurement standards.

We require a new framework for evaluating patient response to therapy – a framework focused on the patient's needs and the extent to which these needs are met with new therapy choices. At the same time, we need a cardinal measure of patient benefit. Achieving these objectives will be the subject of the next IPAA Current Issues report.

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

[1] Institute for Clinical and Economic Review (ICER). 2020 Value Assessment Framework. https://icer-review.org/topic/2020-value-assessment-framework/

[2] McKenna S, Heaney A, Wilburn J et al. Measurement of patient-reported outcomes. 1: The search for the Holy Grail. *J Med Econ*. 2019;22:6(516-22)

[3] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice (Editorial). *J Rehabil Med*. 2012;44:97098

[4] McKenna S, Heaney A, Wilburn J. Measurement of patient reported outcomes. 1: Are current measures failing us? *J Med Econ*. 2019;22:(523-30)

[5] Bond T, Fox C. Applying the Rasch Model (3rd Ed). New York: Routledge, 2015

[6] Institute for Clinical and Economic Review (ICER). Oral Semaglutide for Type 2 Diabetes: Effectiveness and Value. Draft Evidence Report (updated). 12 September 2019. https://icer-review.org/wp-content/uploads/2019/04/ICER_Diabetes_Draft-Evidence-Report_091219-2.pdf

[7] Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the economic evaluation of health technologies: Canada. Ottawa: CADTH, 2017

[8] 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

[9] Langley PC. Sunlit Uplands: The genius of the NICE reference case. InovPharm.2016;7(2) https://pubs.lib.umn.edu/index.php/innovations/article/view/435

[10] Langley PC. The Imaginary Worlds of ISPOR: Modeled Cost-Effectiveness Claims Published in Value in Health from January 2016 to December 2016. InovPharm. 2017;8(2) https://pubs.lib.umn.edu/index.php/innovations/article/view/519

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