THE GREAT I-QALY DISASTER

WORKING PAPER NO 15 JUNE 2020

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

ABSTRACT

The QALY is an impossible construct; it defies common sense. It fails completely once we consider the axioms of fundamental measurement. Utilities as manifest scales cannot be used to create QALYS. The QALY should never have been introduced to support the value assessment of pharmaceutical products and devices. The result is 30 years of QALY based assessments of pharmaceutical products and devices which are conceptually and technically wrong. They are a charade and will have contributed mistakenly to thousands of formulary decisions. In the search for a common metric to evaluate cost-effectiveness the impossibility of a QALY was overlooked. The result is a disaster, unfolding over decades.

THE QALY DISASTER

For the past 30 years the notion of quality adjusted life years, the QALY, has occupied center stage in the technology assessment literature for pharmaceutical products and devices. Unfortunately, in retrospect, the notion of a QALY was always impossible [1]. It should have been dismissed, both as an imaginary and as an operational metric, as soon as it was proposed. The reason is quite simple. If you want to transform a period of time in a disease state to its QALY equivalent, years of perfect health, then the utility or whatever you want to call the metric must have ratio measurement properties. That is, the utility must have a true zero (i.e., no utility value can take negative values) and it must have a maximum value of unity. While, by construct, utilities have a ceiling at unity (1 = perfect health) there is no true zero. There are states worse than death; negative utility scores below the arbitrary 0 = death (which is not a true zero). In the case of the most widely used utility scale, the EQ-5D-3L the scoring algorithm yield utilities in the range -0.59 to 1. If there is no true zero, the utility cannot support multiplication. End of story. The QALY notion should have been smothered at birth. A utility is nothing more than an ordinal score. The ordinal scale also does not have interval properties because no one developing the various utility scales had the thought of building that requirement into the scale. It is, frankly, a complete disaster.

FUNDAMENTAL MEASUREMENT

Briefly, to emphasize the needed role of fundamental measurement in the social sciences, an importance recognized in the physical science for almost 400 years, we must recognize four types of measurement scale; putting to one side conjoint simultaneous measurement which underpins Rasch Measurement Theory (RMT)[2]. These are: nominal, ordinal, interval and ratio.

Each satisfies one or more of the properties of: (i) identity, where each value has a unique meaning; (ii) magnitude, where each value has an ordered relationship to other values; (iii) interval, where scale units are equal to one another; and (iv) ratio, where there is a 'true zero' below which no value exists. Nominal scales are purely descriptive and have no inherent value in terms of magnitude. Ordinal scales have both identity and magnitude in an ordered relation but the unknown distances between the ranks means the scale is capable only of generating medians and modes; it is an ordinal scale. The interval scale has identity, magnitude and equal intervals. It supports the mathematical operations of addition and subtraction. A ratio scale satisfies all properties, supporting the additional mathematical operations of multiplication and division. Recognition and adherence to these fundamental axioms of measurement theory is critical if an instrument is to have any credibility. In the physical sciences accurate measurement is the key to hypothesis testing and the discovery of new facts. The same arguments apply to the social sciences. Unfortunately, they appear all too often to be absent in health technology assessment and in the development of patient reported outcomes (PROMS) instruments.

The principal case against measurement in the social sciences, and the development of generic and disease specific outcomes measures in the discipline known (oddly) as 'pharmacoeconomics' or health technology assessment, is that when the various instruments were developed, no attention was given to the axioms of fundamental measurement. A lack of attention which sets technology assessment aside from the physical sciences, where from first principles instruments have to recognize these axioms [3].

This does not mean that there have not been ongoing criticisms of the QALY. These have ranged from critiques of the assumptions supporting the utility measure to the attributes covered in multiattribute instruments and the neglect of the patient voice. Few however have addressed the question of fundamental measurement and the fact that the various utility instruments support only ordinal scales [4 5 6 7 8] . The fact that the QALY is an impossible measure as utilities are not on a ratio scale has not been recognized by model builders. Once it is recognized, these criticisms have to be focused on instruments to assess whether or not they have the required properties. This leads, for example, to abandoning multiattribute instruments as well as those that are the simple addition of scores, typically on an ordinal response scale for clinician selected items.

THE DISASTER UNFOLDS

How has this I-QALY disaster come about? Over the last 30 years thousands of papers have been published; conferences promoted and seminars attended, professional groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) founded, formulary decisions made and patients excluded from potentially innovative therapies because in the 1980s no one gave a thought to the axioms of fundamental measurement. This is unforgivable. Professional organizations such as ISPOR, numerous gatekeepers for single payer health systems and, in the US, the Institute for Clinical and Economic Review (ICER) have, possibly unwittingly, perpetuated the idea that utility scales, in particular the various

multiattribute generic utility systems, have ratio properties; although the term ratio scale nor any of the other measurement scales were ever mentioned. We now have, and it is being added to daily in journals and conference presentations, a growing elephant's graveyard of I-QALY studies that lack any pretense at scientific credibility; a veritable bone yard.

The disaster can be traced back to the late 1980s with the focus on cost-effectiveness and how claims for competing products were to be presented. The decision makers were assumed to be single payer health systems (the National Health Service in England) which were charged with evaluating the benefits of competing products or new products versus standard of care. With limited resources and the possibility of having to ration or replace existing products, there was seen to be a need for a common metric. The obvious candidate was a multiattribute measure of quality of life (healthy related quality of life). While no thought was given to measurement axioms, it was assumed that utilities from an instrument such as the EQ-5D-3L could be applied.

At the same time, it was recognized that there was limited information to hand at the time of product approval to support cost-effectiveness (now cost-utility) claims; apart from modeling on the clinical trial data. Even so, cost-per-QALY claims would have been invalid; a protocol may have included the EQ-5D-3L (for example) as a primary end-point (most unlikely) but this would still preclude any I-QALY claims based on the timeframe of the randomized clinical trial (RCT). The answer was to create evidence for putative cost-effectiveness through the construction of imaginary decision-framework models. Hypothesis testing was put to one side in favor of the 'approximate information' modeling constructs which lacked even a semblance of empirical credibility. Their claims are non-evaluable, let alone replicable in treatment settings. It was nothing more than pseudoscience. Failing the demarcation test between science and pseudoscience, the proposed technology assessment models, enthusiastically endorsed by ISPOR in the collection of best practice monographs, joined intelligent design, not natural selection, in the Dover courtroom [9]. For ISPOR, the absence of evidence does not preclude imaginary claims being made. Indeed, by design the modeled claimed were impossible to evaluate. An enviable position: claims were unable to be evaluated empirically. As such modeled claims were all too often seen as marketing devices with consultants developing model frameworks that supported a client's product. This led, inevitably, to a cadre of inquisitors. Academic groups that were tasked with reviewing models submitted to agencies (but not models submitted to journals) to pronounce on their methodological purity.

Unlike the other social sciences where the focus is on hypothesis testing and the discovery of new facts, emulating the physical sciences, the world of ISPOR, ICER and gatekeeper technology assessment agencies is devoted to the construction and marketing of fatally flawed imaginary worlds. A belief system, a meme, that is widely entrenched to the extent that thousands of researchers believe that they can ignore fundamental measurement (if they are aware of it) and drive formulary decisions, not by the discovery of new facts and the evaluation of hypotheses as to product performance in target patient populations, but by the putting to one side of the standards of normal science and creating imaginary evidence for cost-effectiveness [10]. Health technology assessment is the only social science branch, if that is not too inappropriate a label,

that actively promotes imaginary constructs and puts its faith in non-evaluable claims. Absent the standards of normal science: truth is consensus.

Considerable effort was devoted to justifying the construction of lifetime imaginary cost-per-incremental QALY imaginary worlds. By the end of the 1990s and the adoption of the NICE reference case which mandated manufacturer constructed imaginary worlds and thresholds for pricing [11]. Global acceptance soon followed (and no mention of fundamental measurement). Promoted by sensitivity analyses, probabilistic sensitivity analyses and supplementary scenarios, the reference case model was accepted and endorsed by groups such as the Academy of Managed Care Pharmacy and the National Pharmaceutical Council. Yet no one recognized the fundamental and fatal flaw with the impossible QALY. Individuals who raised the issue of the axioms of fundamental measurement for interpreting utilities were ignored; yet even then the implications of this for the QALY were overlooked.

AN UNSHAKEABLE BELIEF

If anyone finds this a somewhat alarming picture then the belief in the role of imaginary constructs and approximate information (or disinformation) is clearly stated by ISPOR: *leaders in the field of economic evaluation have long recommended that analysts seeking to inform resource allocation decisions approximate the value of interventions in terms of incremental cost-per-QALY gained* [12]. This position is endorsed by the latest Canadian guidelines: *Economic evaluations are designed to inform decisions. As such they are distinct from conventional research activities, which are designed to test hypotheses* [13].

The problem, of course, is that this is not 'approximate information'; it is the antithesis – there is no information content because the QALY is an impossible construct. Claims for incremental I-QALY gains, irrespective of the time frame and choice of assumption, are meaningless. There is a fundamental disconnect between the standards of normal science and what passes for 'standards' in health technology assessment. We are dealing with constructs which fail accepted standards for instrument development: the emphasis on multiattribute 'measures' and the belief that we can add up any number of scores and believe the resulting aggregate score can actually meet either latent or, more broadly, ratio scale properties. The point that is overlooked is that if you wish to measure response, then the instrument has to have latent and, if possible, ratio properties. Developing a measure through the application of classical test theory ensures that the resulting scale or score is nothing more than an ordinal scale. The endeavor falls at the first hurdle. If you want to measure any particular attribute then your instrument must capture those items that ensure it has the required measurement property. The data have to fit the instrument; we should not try to fit the instrument to the data.

EXEUNT ICER

While ISPOR may be able to survive the QALY disaster, ICER will have difficulty. For the past 5 years ICER has built its credibility, influence and its business case on a fatal misunderstanding,

by ICER staff and reviewers, of the role of normal science. The value assessment reference case, the construction of lifetime incremental cost-per-QALY models to support non-evaluable threshold based recommendations for pricing, product access and budgetary scenarios, is a charade. The fatal flaw is the I-QALY; an impossible mathematical construct. While some might be prepared to live with approximate information based on lifetime value assessments, the edifice collapses once I-QALYs are rejected. Not only are incremental cost-per-I-QALY claims impossible, they should not even be considered as information, approximate or otherwise. The application of cost-per-I-QALY thresholds is a complete nonsense. ICER has nowhere else to go. There is no reason for organizations and manufacturers to support ICER. They would be better engaged with subscriptions to *Psychic News*.

PUSH BACK

Without doubt, after 30 years there will be sustained efforts to defend the I-QALY. The belief system, or technology assessment meme, is well entrenched. Unscrambling and rejecting such a belief system, a belief in the fog of approximate information, will not be easy. After 30 years the retreat to 'truth is consensus' will lead to the wagons being circled. No one, let alone those experiencing 30 years of indoctrination and advocacy, wants to be told that the leaders in the field are wrong. But perhaps the death will be over quickly. After all, the ISPOR/ICER paradigm only survives (and the two organizations) if the 'customers' believe that 'approximate information' built on an imaginary lifetime construct has merit. Once doubts set in, with formulary committees and patient advocacy groups pointing out the manifest deficiencies of the I-QALY paradigm, a tipping point will have been reached.

ISPOR, which has a lot to lose, may mount a save the I-QALY crusade. Critics will be pilloried; journal articles and letters to the editor will be summarily rejected, leaders in the field of technology assessment will redouble and double again efforts to assure their followers and clients in industry and technology assessment that the mystery of the QALY is safe. Advocates could argue that, yes, It might need modifying and embedding within other value assessment frameworks so that the efforts of the last 30 years are seen as a sure stepping stone to a more inclusive, even patient centric, use of modeled claims in imaginary formulary decisions. After all, an approximate assumption driven information framework, blessed by professional groups and leaders in the field, is a sure and certain hope for an I-QALY future. The issue of fundamental measurement will be just a minor issue which, by assumption, can be put to one side. Welcome to a new branch of science fiction. It is on this rejection of normal science that health technology assessment is built.

Paradigm shifts are not new in the history of science, or in the social sciences. Indeed, the disputes can become extended with academic groups and journal editors taking sides. The present case does not bode well. It is not the question of special and general relativity supplanting Newtonian mechanics. As Brian Greene puts it: *When quantum mechanics came along, Newton's edifice was not dismantled. It was renovated. Quantum mechanics provided a new foundation that deepened the reach of science and gave the Newtonian structure a fresh interpretation* [14]. In the I-QALY case it is not a question of retention; the I-QALY edifice has to

be dismantled. A situation where the post-I-QALY paradigm for assessing the merits of competing pharmaceutical products and devices actually replaces the I-QALY paradigm. This seems unavoidable: there is not much room to breathe if the preceding paradigm is built on a denial of the standards of normal science including a belief that it is possible to assume that an ordinal scale can have ratio properties. Creating evidence by assumption, fudging the analytical framework by describing it as creating approximate information, is not a position consistent with a post-I-QALY paradigm that attempts to meet defensible standards for hypothesis testing and the discovery of new facts. Unless, to follow Sir Thomas Browne's (1605-82) transliteration of Tertullian's (155-240 AD) maxim, advocates of the I-QALY believe in the I-QALY and the modeling of imaginary worlds because the construct is impossible (*et sepultus resurrexit: certum est, quia impossibile*). In other words, that there is a refuge and defense of the mystery of the I-QALY paradigm; the more impossible the I-QALY mystery the stronger the belief held by its supporters. There is no possibility of a synthesis of the post-I-QALY paradigm building on and encompassing the I-QALY paradigm, let alone trying to insert it in a 'wider' value assessment framework. There are just too many fundamental errors. We cannot build on an impossible construct despite how many believe in it.

ACCEPTING I-QALYs

The great I-QALY disaster is not just the result of a minor academic oversight. The implications go far deeper. For 30 years formulary decisions have been influenced, if not dictated, by a cost-per QALY construct which is mathematically nonsensical. Thresholds have been touted as the last word in value assessment. Pricing negotiations have been driven by an imaginary construct; access to care has been similarly blighted. Why? Because 'leaders' in the field determined that creating evidence for cost-effectiveness claims trumped (sorry!) real world evidence. The last 30 years have given the opportunity to develop evidence platforms to support cost-outcomes claims; that has been squandered by the obsession with reference case imaginary worlds.

Although the data are limited, there is evidence pointing to an unfortunate increase in the acceptance of ICER recommendations in the formulary decision process. The concern is that in accepting ICER recommendations to support pricing and access protocols, the decision makers have no idea of the basis on which ICER arrived at these conclusions. Few formulary or more broadly health decision makers, have the skills necessary to evaluate ICER modeled claims. The reference case framework is a black box. Questions regarding the reference scientific status and the impossibility of creating QALYs in an imaginary pseudoscientific simulation are ignored; none have the skills to argue the case for its rejection. This lack of awareness is further compounded by the failure to recognize, or even acknowledge, the limitations imposed by the axioms of fundamental measurement.

In this situation manufacturers will be rightly concerned that they are being shortchanged. Introducing recommendations for pricing and product access into formulary negotiations that have no basis in reality is hardly a comfortable situation. To what extent do formulary committees give weight to the I-QALY construct? What recourse do they have if they fail to convince formulary committees that the I-QALY modeling is an analytical dead end? More

troubling are the concerns of patients and providers. Are they being denied access to new and innovative therapies on the basis of an imaginary and impossible simulation model?

BACK TO BASICS

The fact that utilities are ordinal scales is incontrovertible. They were not designed to be anything other than ordinal scores; by design they cannot support I-QALYs. If there had been an intention to generate utilities with ratio properties this should have been central to the development of the various systems. The term 'QALY' should be put to one side; excised from the technology assessment lexicon. Certainly, focus on the notion of a latent construct that we can call quality of life. As a measure of response there are now a number of examples of needs-based quality of life latent measures that meet the standards for fundamental measurement. Abandoning the existing technology assessment paradigm clears the field for a new paradigm, a patient centric paradigm that recognizes the need for meeting the fundamental axioms of measurement theory. We can abandon the absurdity of lifetime imaginary constructs and the creation of approximate information.

REFERENCES

[1] 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

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

[3] Chang H. Inventing Temperature: Measurement and scientific progress. New York: Oxford University Press, 2007

[4] Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*. 1989;70:308-12

[5] Tennant A, McKenna S, Hagell P. Application of Rasch Analysis in the development and application of quality of life instruments. *Value Health*. 2004;7(Supp 1):S22-26

[6] 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

[7] 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

[8] McKenna S, Heaney A, Wilburn J et al. Measurement of Patient Reported Outcomes 2: Are current measures failing us? *J Med Econ*. 2019;22(6):523-30

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

[10] Dawkins R. A Devil's Chaplain. New York: Houghton Mifflin, 2004

[11] Langley P. Sunlit Uplands: The genius of the NICE reference case. Inov Pharm. 2016;7(2) https://doi.org/10.24926/iip.v7i2.435

[12] Neumann P, Willke R, Garrison J. A Health Economics Approach to US Value Assessment Frameworks – An ISPOR Task Force Report. *Value Health*. 2018;21:119-123

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

[14] Greene B. Until the End of Time. New York: Knopf, 2020