MAIMON WORKING PAPERS **www.maimnonresearch.net**

THE GREAT I-QALY DISASTER (version 2)

WORKING PAPER NO. 18 August 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 ordinal 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. Our next steps must be to abandon the QALY paradigm and look ahead to a new value assessment framework.

INTRODUCTION: THE I-QALY DISASTER

For the past 30 years the notion of quality adjusted life years, the QALY, has occupied center stage in the value assessment literature for pharmaceutical products and devices. Unfortunately, in retrospect, the notion of a QALY was always impossible; hence the application here of the term impossible or I-QALY [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 I-QALY equivalent, years of perfect health, then the utility or whatever you want to call the metric must have ratio measurement properties. 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. 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 scale cannot support multiplication. End of story. The I-QALY notion should have been smothered at birth. A utility is nothing more than an ordinal score. The ordinal scale does not have interval or ratio properties because no one developing the various utility scales and creating QALYs had the thought of building those requirement into the scale. It is, frankly, a complete disaster.

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. A lack of attention which sets value assessment aside from the physical sciences, where from first principles instruments have to recognize these axioms [2]. This does not mean that there have not been ongoing criticisms. 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 [3] . Only a few 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 DISASTER UNFOLDS

How has this I-QALY disaster come about? Basically, through ignorance of fundamental measurement and a belief in the contribution of evidence constructed to support imaginary claims for cost effectiveness. 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 or preference generic utility systems, have ratio properties or can be treated as if they do; although the term ratio scale nor any of the other measurement scales were ever mentioned. Indeed it is only in the fourth and latest edition of one of the most widely used textbooks in health system economic evaluations that the question of measurement standards has been addressed [9]. Even then, there was a lack of appreciation of the standards for ratio scales and the impossibility of creating I-QALYs.

The I-QALY 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 (e.g., 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 central health planning metric. An obvious candidate was a generic multiattribute measure of quality of life (health related quality of life: HRQOL). While no thought was given to measurement axioms, it was assumed that utilities from an instrument such as the EQ-5D-3L, a multiattribute, multidimensional, ordinal HRQOL instrument could be applied. This was the fundamental error.

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-I-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 imaginary cost-effectiveness claims through the construction of simulated lifetime 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 were, a major positive plus, non-evaluable, let alone impossible to replicate between 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 [10]. For ISPOR, the absence of evidence did not preclude imaginary claims being made; in fact the opposite, it was encouraged. Indeed, by design, the modeled claimed were impossible to evaluate. An enviable position: claims were unable to be evaluated empirically; a truly mediaeval acceptance of truth; or its various incarnations.

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 [11]. 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. Evidence for cost-effectiveness is invented not discovered [12].

Considerable effort was devoted to justifying the construction of lifetime imaginary cost-per-incremental I-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 [10]. Global acceptance soon followed (and again no mention of fundamental measurement). Promoted by one and two way 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 [13]. Yet no one recognized the fundamental and fatal flaw with the I-QALY. Individuals who raised the issue of the axioms of fundamental measurement for interpreting utilities were ignored [4 5]; yet even then the implications of this for the I-QALY were overlooked.

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 [14]. 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 unidimensional measurement property. The input data have to fit the instrument; we should not try to fit the instrument to the data.

MARKETING THE I-QALY

Basing cost-effectiveness claims on simulated lifetime imaginary world, driven entirely by assumption where the claims are impossible to evaluate empirically, offers a tempting target for reverse engineering claims. In the absence of the ability to test hypotheses where claims are credible, evaluable and replicable the pseudoscience of simulated I-QALY claims takes center stage. In defense of the imaginary world, the key is the notion of 'approximate information'.
This has been 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* [15]. 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* [16].

Taking refuge in the argument that these simulated I-QALY models are 'for approximate information' is no defense. The term has no meaning. This is not 'approximate information'; it is worse – there is no information content, there is no question of reducing uncertainty, because the I-QALY and cost-per-I-QALY lifetime models are impossible fictions; the I-QALY claims for incremental I-QALY gains, irrespective of the time frame and choice of assumption, are meaningless. Attempts to make I-QALY presentations more 'believable'; by application of sensitivity analyses are equally meaningless. It is difficult to conceptualize the likelihood, through probabilistic sensitivity analysis, of the cost-effectiveness of an impossible claim. 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 accept that the resulting aggregate score can actually meet either latent or, more broadly, ratio scale properties [17].

Unfortunately, 'approximate information', opens the floodgates to marketing simulations with the risk of consultants developing model frameworks that support a client's product. Concern with the construction of 'favorable' imaginary claims has led, inevitably, to a cadre of inquisitors tasked with reporting on the purity of the imaginary modeled claims. Academic groups in countries such as the UK and Australia (reporting to NICE and the Pharmaceutical Benefits Advisory Committee respectively) were tasked with reviewing models submitted to the agency to pronounce on their methodological purity [18] [19]. Inquisitors could suggest modifications or alternative model structures and assumptions, even giving a submission a good housekeeping seal of approval. These critiques of the submitted models illustrated the flexibility of these non-evaluable claims with competing model frameworks. Journals did not have this luxury; the peer review process was not intended to support this level of unbundling a model. Even so, it is

difficult to believe that the peer review process overlooked the impossibility of the I-QALY, let alone the lack of appreciation of the standards of normal science.

FIDELITY AND ISPOR CHEERS

Beliefs must be sustained; the technology assessment meme must be codified and reinforced to ensure copying fidelity (e.g., Council of Nicaea and Nicene Creed). Practitioners building imaginary worlds require assurance that they are meeting required standards for modelling. The ISPOR CHEERS reporting guidance checklist provides such an assurance [20]. Published in 2013, CHEERS (Consolidated Health Economic Evaluation Reporting Standards) builds on previous ISPOR standards, providing imaginary and I-QALY driven model builders with a framework for justifying their construction of simulated claims. Certainly a number of the questions would be relevant to modeled claims consistent with the standards of normal science, but it is clear from the context and the questions addressed that the primary audience is those creating imaginary I-QALY driven cost-outcomes claims. These include analysts and reviewers; indeed it is noteworthy that to support fidelity in imaginary model construction and assessment that the CHEERS checklist was distributed to leading journals that had (and continue) to accept cost-per-I-QALY imaginary constructs.

Given the present I-QALY critique it is worth noting that there is no requirement for model builders to determine whether or not their claims meet the standards for credibility and empirical evaluation. There is no discussion of instrument standards and the need to meet the axioms of fundamental measurement. The EQ-5D-3L is given as an exemplar outcome measure. In practical terms CHEERS is a checklist for reporting and reviewing imaginary economic evaluations. A tool designed to reinforce existing imaginary modeling standards proposed by ISPOR. As would be expected there is no discussion of claims evaluation in target treating populations. Nor should we expect this as it is simply a checklist for the imagination. The reporting of results section is focused on the model as a self-referential closed system: study parameters (reporting vales, ranges, references and parameter probability distributions), incremental costs and outcomes, the characterization of uncertainty and heterogeneity. All that would be needed to clarify the intent of the CHEERS checklist would be to ask whether the economic evaluation was designed to propose credible and empirically evaluable claims. This would, perhaps unfortunately, put the fox in the imaginary henhouse.

THE WORLD TURNED UPSIDE DOWN

But the hens are safe. There is no doubt as to the acceptance of the I-QALY. As an indication, a Pub Med count on the keywords “Cost AND QALY” [accessed 29 July 2020] yielded a total of 16,378 citations with hundreds of citations for a sample of individual journals (Table 1). The majority of these published in the last 15 years with an almost exponential growth, particularly for the health technology journals such as *Pharmacoeconomics*.

TABLE 1

PUBMED CITATIONS FOR "COST AND QALY"

Total count	**16,378**
PharmacoEconomics	**552**
Value in Health	**519**
Journal of Medical Economics	**369**
British Medical Journal	**374**
Lancet	**193**
Journal of the American Medical Association	**164**
New England Journal of Medicine	**124**

To supporters of the I-QALY and approximate information this is a sure and certain sign of the acceptance of their paradigm; to others who are aware of the limitations imposed by the axioms of fundamental measurement and the pre-eminent role of the scientific method, this is an indictment. How so many could engage with a pseudoscientific belief system is testament to the willingness to engage with a belief system founded on ignorance.

Unrecognized after some 30 years of I-QALY application, the incontrovertible fact is that generic multiattribute instruments, the standard gamble and time trade off, together with virtually all PROMS yield ordinal scores. This is dictated by the axioms of fundamental measurement. The even more unfortunate corollary is that the implicit belief that the various instruments have ratio properties can be simply dismissed. The logical conclusion is that the I-QALY is an impossible mathematical construct. This is, to say the least, an embarrassing situation. How are the erstwhile supporters of the I-QALY and the cost-per-incremental I-QALY value assessment framework to respond?

This 'embarrassment' is made the more concerning when the sheer number of studies is considered. Although the PubMed count of 16,378 could be refined by a more rigorous systematic review (which seems hardly worth the effort) it is difficult to think of another example of misapplied science that reaches this publication magnitude. The willingness, without question, of analysts to accept the I-QALY is astounding; but who is to give the bad news? Journal editors are in an invidious position with the 'leading' journals in health technology assessment such as *Pharmacoeconomics* with 552 'hits' having to provide some explanation. The same applies to the leading medical journals.

DEFENDING THE I-QALY

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 merits of impossible (not approximate) information, will not be easy. After 30 years the appeal and 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. 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 I-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, it might be argued, an approximate assumption driven information framework supporting non-evaluable claims, blessed by professional groups and leaders in the field, provides 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.

But perhaps the death will be over quickly. After all, the technology assessment paradigm only survives if the customer, such as a formulary committee, believes it 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. If so, this raises the questions of explaining how this unfortunate state of affairs was maintained and how the process of re-education might be initiated.

Paradigm shifts are not new in the history of science, or in the social sciences. Indeed, the disputes can become extended (and noisy) 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 points out: *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* [21]. In the I-QALY case it is not a question of a fresh interpretation within a successor paradigm; the I-QALY edifice has to be rejected. A situation where the post-I-QALY paradigm for assessing the value of competing pharmaceutical products and devices actually replaces the I-QALY paradigm. This seems unavoidable: there is not much room, if any, for maneuver. As the I-QALY 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, then this will be totally at odds with a new value assessment paradigm that recognizes these axioms. There are just too many fundamental errors. We cannot build on an impossible construct despite how many believe in it. Real world evidence must replace imaginary world created evidence.

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 I-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 trumps real world evidence. The last 30 years could have provided 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 I-QALY based 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 I-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 on instrument development.

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 manufacturers have if they fail to convince formulary committees that the I-QALY modeling is useful and not 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? Are health care decision makers being misled?

CONCLUSIONS

The fact that utilities are ordinal scales is incontrovertible. They is no way they could be anything other than ordinal scores; by design they cannot support cost-per-QALY claims and threshold criteria for pricing and access. If there had been an intention to generate utilities with interval, if not ratio properties this should have been central to the development of the various systems; it was not. The meaningless 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 [22] [23]. Abandoning the existing technology assessment paradigm clears the field for a new paradigm, a patient centric, disease specific paradigm that recognizes the need for meeting the fundamental axioms of measurement theory. Patients are the principal beneficiary, including caregivers, from new and innovative therapies. This should not be denied. We can abandon the absurdity of lifetime imaginary constructs and the creation of impossible information. Our attention needs to be on patients and caregivers. Once this evidence is before us, we can

discuss pricing and access; these discussions must be driven by real world evidence, not imaginary and impossible claims.

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

[3] Pettitt D, Raza S, Haughton B et al. The Limitations of QALY: A literature review. *J Stem Cell Res Ther*.2016;6:4

[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] 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)

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

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

[12] Wootton D. The Invention of Science: A new history of the scientific revolution. New York: Harper Collins, 2015

[13] Langley P. Modeling Imaginary Worlds: Version 4 of the AMCP Format for Formulary Submissions, *Inov Pharm*. 2016;7(2): No.11

[14] Bond T, Fox C. Applying the Rasch model: Fundamental measurement in the human sciences. 3rd Ed. New York: Routledge, 2015

[15] Neumann P, Willke R, Garrison J. A Health Economics Approach to US Value Assessment Frameworks – An ISPOR Task Force Report [1]. *Value Health*. 2018;21:119-123

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

[17] McKenna S, Heaney A. Composite outcome measurement in clinical research; the triumph of illusion over reality? *J Med Econ*. 2020doi:10.1080/13696998.2020.1797755

[18] Langley P. Sunlit Uplands: The genius of the NICE reference case. *Inov Pharm*. 2016;7(2): No. 12

[19] Langley P. Dreamtime: Version 5.0 of the Australian Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC). *Inov Pharm.* 2017;8(1): No. 5

[20] Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS)—explanation and elaboration: a report of the ISPOR Health Economic Evaluations Publication Guidelines Good Reporting Practices Task Force. *Value Health*. 2013;16(2):231-250.

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

[22] Wilburn J, McKenna SP, Twiss J et al. Assessing Quality of Life in Crohn's Disease: Development and validation of the Crohn's Life Impact Questionnaire (CLIQ). *Qual Life Res*. 24(9):2270-88

[23] Wilburn J, Twiss J, Kemp K et al. A qualitative study of the impact of Crohn's disease from a patient's perspective. *Frontline Gastroenterol*. 2017;8(1):68-73