MAIMON WORKING PAPERS www.maimonresearch.net

WORKING PAPER No. 14 JUNE 2020

INTRODUCING THE I-QALY: THE IMAGINARY CONSTRUCT SUPPORTING HEALTH TECHNOLOGY ASSESSMENT

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

Abstract

For over 30 years the concept of the quality adjusted life year (QALY) has played the central role in health technology assessment. Endorsed by agencies in single payer health systems such as the National Institute for Health and Care Excellence (NICE) in the UK and professional groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), cost-per-incremental-QALY models have been the vehicle for promoting cost-effectiveness of pharmaceutical products and devices. Unfortunately, as typically constructed, the QALY is a mathematical impossibility. Time spent in a disease state cannot be converted to quality adjusted time unless the adjustment 'utility' has ratio measurement properties; that is, a true zero where no adjustment utility can take a negative value. This is patently not the case as utilities can take negative values. This has been recognized and ignored for 30 years. The result is the publication of thousands of cost-per-QALY claims for cost-effectiveness which are worthless. The purpose of this commentary is to point out that the QALY is best rebadged as the imaginary QALY (I-QALY), joining other attributes of modeled claims for cost-effectiveness which put the modeling in the pseudoscience category; a commitment to pseudoscience and false claims that has characterized health technology assessment for over 30 years.

Introduction

Promoting the standards for health technology assessment, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has made quite clear the role of imaginary constructs as central to their cost-effectiveness message. For 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* [1]. This approximate value, or approximate information, is generated from the construction of lifetime imaginary worlds. Within a disease area, a hypothetical population is tracked over its assumed lifetime. The framework, which is typically a variation on a Markov process (he was shot by Stalin) represents a target disease categorized by stages with individual patients transitioning through these stages. The transition probabilities determine how long, on average, patients spent in each disease stage. Estimates of these probabilities are based on randomized clinical trial (RCT) results clinical trial results for response and relapse, with some auxiliary assumptions to support moving from a short term trial data to a lifetime response to therapy. Time spent in each disease state is translated to QALYs by multiplying by a utility value (notionally on a 0 to 1 scale). Utilities and costs of treatment in each disease stage are taken from the literature or simply guessed. Assumptions are typically justified on the grounds that they are 'realistic', with sensitivity analyses to suggest plausible variations to accommodate uncertainty over the lifetime which can be in excess of 30 years. The result is estimates of lifetime QALYs and costs. This gives estimates of incremental cost per QALY for the value case. This is the ICER reference case model.

The attraction of this model and the belief in its role in value assessment should not be understated: there have been thousands of studies published utilizing this framework with hundreds of academic researchers and others subscribing to it. Unfortunately, there are four issues with this model that groups like ISPOR, ICER and the various single payer health systems (e.g., National Institute for Health and Care Excellence, NICE, in the UK) who accept reference case models have yet to address satisfactorily. These are: (i) the reference case model fails to meet the standards of normal science; (ii) it is logically impossible to assume that assumptions from the past will hold in the future; (iii) QALYs are impossible mathematical constructs because they fail to meet the axioms of fundamental measurement and (iv) no one has defined or agreed on what the term 'approximate information' actually means.

Previous commentaries have considered these issues in detail [2]. As the focus here is on the I-QALY and fundamental measurement we can note briefly the concerns: (i) since the scientific revolution of the 17th century science has progressed through hypothesis testing and the discovery of new facts – this is alien to the reference case, described as pseudoscience, as the model claims (by construction) eschew hypothesis testing as they are not credible, evaluable empirically or replicable across possible target populations; (ii) it is logically impossible to argue that because a prior observation has been made that we can assume it will hold in the future – this is Hume's problem of induction (David Hume 1711-1776) which was only resolved by the rejection of logical positivism in the 1930s with Popper's (Sir Karl Popper 1902 – 1994) contribution to the philosophy of science in its emphasis on conjecture and refutation[3] [4]; we cannot build imaginary future worlds (outside of science fiction) on prior observations 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* [5] and (iii) when is approximate information approximate disinformation? The term is meaningless although it is center stage in justifying the construction of I-QALY imaginary claims for cost-effectiveness. This is stated quite clearly in the latest version of the Canadian formulary guidelines: *Economic evaluations are designed to inform decisions. As such they are distinct from conventional research activities, which are designed to test hypotheses* [6].

The I-QALY as Center Stage

The recognition that the I-QALY is an imaginary construct (hence the "I" for imaginary in I-QALY) rests on the axioms of measurement theory, the role of measurement scales. There are four measurement scales; putting to one side conjoint simultaneous measurement which underpins Rasch Measurement Theory (RMT)[7]. 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 a manifest 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[8] [9]. In the physical sciences this has been long recognized; 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.

If we assume for the moment that with the reference case model we can put to one side concerns that it fails the standards of normal science, can be absolved from the induction problem and that the concerns with approximate information are overstated, then the problem is that the typical multiattribute instrument that creates utilities (e.g., the EQ-5D-3L, EQ-5D-5L, HUI Mk3, TTO) is actually generating manifest scores [10]. That is, the utility score is on an ordinal scale: it has magnitude in an ordered relationship of one multiattribute utility score to another but we have no idea of the difference between values. It lacks the invariance of comparisons that characterize an interval scale and also fails the required standard of the ratio scale, a true zero. The EQ-5D-3L, which is constructed from symptom ordinal scales, yields utility values in the range -0.59 to 1.0. It is therefore possible to have negative I-QALYS. As the I-QALY requires time (which has a true zero) to be multiplied by a ratio scale with a true zero, the resulting I-QALY is an impossible construct. The only escape is to ignore the negative values and assume the zero (death) is a 'true zero'. After all, with the plethora of assumptions characterizing the ICER reference case, what is one further assumption? This is only approximate information. Manufacturers, formulary committees and journal editors can, presumably, live with this 'necessary' assumption even though the result defies the axioms of measurement theory and any modelled claims build on I-QALYs are nonsensical. Put simply: ICER recommendations for pricing and product access lack any pretense to credibility.

ICER's position could be defended on two grounds: first, the ICER staff deeply and truly believe that the EQ-5D-3L as a representative generic utility truly has ratio properties or, second, it recognizes the problem but as its business case rests on imaginary cost-per-I-QALY worlds it just 'holds its nose'. Neither option is defensible. The purpose of this commentary is to make abundantly clear that the ICER position is indefensible.

The Road Not Taken

It is of interest to speculate as to why ISPOR and professional groups in various academic institutions latched onto the construction of imaginary I-QALY worlds to support claims for cost-effectiveness. In the early 1990s evidence for product impact at product launch was limited (it still is). Certainly, administrative claims data in the US provided some insights, but their ability to support claims for cost-effectiveness were limited. What were the options: (i) to establish evidence frameworks for capturing and reporting product impacts (e.g., registries for target patient populations) or (ii) to short-circuit the question by proposing the construction of imaginary lifetime reference case worlds to create evidence to support imaginary cost-effectiveness claims. The latter was chosen, nature abhors a vacuum with considerable hype surrounding decision models with 'realistic' assumptions and likelihood estimates of cost-effectiveness. Within ten years this commitment to the construction of imaginary cost-per- I-QALY models had morphed into the NICE reference case and the global acceptance of imaginary lifetime value assessment frameworks [11]. This puts 'pharmacoeconomics' in a unique position: alone among the social sciences it has promoted the construction of imaginary models to fabricate evidence. Model building, with assumptions, to support hypothesis testing was rejected in favor of non-evaluable, by design, claims, which could never be challenged empirically. An undeniably unique and perhaps enviable situation. Any possible objection was put to one side by the application of sensitivity analyses for the constructed decision frameworks; scatter diagrams were a standard tool with probabilistic sensitivity analysis providing further cover. Although a few authors challenged the absence of any consideration of measurement theory they were, essentially, ignored [12]. Interestingly, these focused on the question of cardinal measurement properties of utility scales rather than on the question of ratio properties which would have shown that the I-QALY was a mathematically indefensible construct from day one.

Without doubt, there was a willing (and apparently forgiving) audience for the imaginary I-QALY reference case model. After all, it gave a lot of people something to do and publish. Rather than waiting for real world evidence the analysts, often lacking formal training in economics, could build quickly a lifetime Markov model (or equivalent) and create the necessary evidence for a cost-effectiveness claim. Manufacturers needed claims. Indeed, it has been noted on many occasions that published I-QALY model claims appear disproportionately to favor the manufacturer whose product is being modeled. Journal editors seem oblivious to this 'marketing exercise' where creating evidence trumps (sorry!) hypothesis testing. Add to this the presence, in countries such as the UK and Australia of academic groups who are contracted to review and referee manufacturer's I-QALY models by NICE and the Pharmaceutical Benefits Advisory Committee (PBAC). While this may seem a singularly pointless activity, after all we can envisage a multiverse of 'realistic' models, they persevere to concur, modify or create their own imaginary models. None are apparently aware of the axioms of fundamental measurement. There is, one might note, a certain Jesuitical quality to the role of referees to pronounce upon the 'truth' of imaginary beliefs. This is, it should be emphasized, not a conspiracy of silence. The various actors seemed to be genuinely unaware of measurement theory. This applies even to the groups responsible for developing the various generic instruments. Certainly they were aware of the uncomfortable fact of negative utilities, but this was never pursued to its logical conclusion in measurement theory.

Thirty Wasted Years

The ability for groups such as ISPOR to sustain over decades the belief in the construction of imaginary cost-per-I-QALY worlds to support formulary decisions may seem unreal. After all there have numerous papers published and warnings to ISPOR in the intervening years, even in its own journal, *Value in Health*, that the axioms of fundamental measurement were being violated [10] with more recent critiques in the *Journal of Medical Economics* [13] [14]. There are three possible and not mutually exclusive explanations: first, that the supporters of generic multiattribute utility instruments were simply unaware of the axioms of fundamental measurement and the recognition, held in the physical science for over 300 years, that an instrument has to be designed to capture single attributes from the ground up and with the required measurement properties determined from day one; second, the perceived need to fill an information vacuum meant analysts were not prepared to wait for evidence platforms to test claims, opting instead for approximate imaginary information; and third, that once embarked on constructing the imaginary world paradigm there was no turning back, leaders in the field could hardly announce that they had built a castle on sand; that the failure to recognize the limitations of their approach and the impossibility of meeting the required measurement standards were, quite simply, wrong.

Over the next few years the belief in the lifetime reference case paradigm hardened; it became a mainstream belief system. Endorsed by leaders in the field, leading academic institutions and the dominant professional group, ISPOR, it has become an unquestioned 'analytical' tool for imaginary constructs. ISPOR is now a global organization with strict rules on value assessment where the I-QALY is center stage. Opposing this is a challenging prospect.

Of course, the defense is that the object is to create approximate information and not to test hypotheses. This admits that the approach fails to meet standards for normal science; it is, as noted, pseudoscience (or bunk) [15]. The term approximate information is meaningless; it relies on a naïve belief in the realism of future assumptions. But the *coup de grace* is the manifest utility scale. If a utility scale is to have any credibility for QALY creation, multiplication by a scale on a zero to unity number line, then the scale must have ratio properties. That is, it must, in addition to latent interval properties, have a true zero. This is measurement theory 101. Unfortunately, none of the generic or even disease specific utility measures have this property. We have known since the 1960s of the role of axioms of measurement and with the extension of the axioms for measurement into the concept of conjoint simultaneous measurement. The work of Luce and Tukey [16], and Rasch [17] made it quite clear that we needed a competing paradigm. Rather than attempting to fit models to data (and creating negative utilities as an unwelcome and fatal side effect) we should determine the attribute to be measured, typically a latent attribute in the non-physical sciences, and identify those items that could support a non-physical based measure of that underlying attribute. The Rasch measurement model was a recognized solution. It translated ordinal scales to an interval scale, focusing of ordering items by their difficulty and the ability of respondents, to create an instrument that could capture response to therapy interventions. Unfortunately, even today Rasch measurement theory (RMT) is ignored. One of the leading textbooks, now in its fourth edition after some 28 years fails to mention of discuss RMT and it is only in the latest edition (2015) that the axioms of fundamental measurement are addressed (and then confusing interval and ratio scales to support measurement) [18]. If we abandon the I-QALY, Rasch offers a way forward to develop disease specific patient centric needs-based quality of life instruments that meet the axioms of fundamental measurement [19]. As these are interval scales they cannot, of course, support creating QALYs. The key point is that they offer a patient-centric unidimensional credible measure of response to therapy. These Rasch instruments are presently available for a wide range of disease states [20]. This does not mean that generic multiattribute instruments should be abandoned; nor the majority of PROMS that also fail fundamental measurement. Manifest scores can be reported just as the responses to Likert scales. Going beyond these measures is to fail the axioms of fundamental measurement. The heyday of generic multiattribute measures, which are poor if not misleading measures of response to therapy, is over.

A Short Cut Bites Back

There is no evidence to suggest, even now, that the primacy of the I-QALY has been challenged in the technology assessment literature. The leading journals continue to publish I-QALY studies. While both ISPOR and ICER have reviewed proposed alternative (or more properly complementary) value assessment frameworks, the I-QALY reference case model is still central to comparative product claims. The focus over the last 30 years on a societal resource allocation framework to drive cost per I-QALY metrics within health systems was always of academic interest but of no practical importance in health care decisions where political considerations are paramount. One explanation for this resolve (or more properly refusal) is that to admit a fundamental error would not only undercut their position as thought leaders (ISPOR) but their business case (ICER). No doubt there will be attempts to demonstrate that generic multiattribute utilities have undeniable ratio properties. There may even be attempts to rescale utilities to create an artificial zero. Most likely academic interest will switch to other metrics for technology assessment with the past 30 years conveniently forgotten. Yet, the I-QALY may linger. The technology assessment old guard may defend the I-QALY mystery: *prorsus credibile est, quia ineptum est* (it is certain because it is absurd; Tertullian 155 – 240 AD)

The issue that the leaders of health technology assessment have yet to face is how to explain why, for 30 years, there has been an explicit endorsement by academic centers, journal editors and reviewers of the I-QALY. Perhaps they have been seduced by the chimera of approximate information where the model claims can accommodate any number of dubious assumptions, including the assumption, or closely held belief, that the utility has ratio properties. Only a psychologist could answer the question; perhaps a specialist in memes, the mysteries of strongly held belief systems and their transmission fidelity.

Even so, this realization, after 30 years, puts ISPOR and ICER in an awkward position. It is not as though it is a sudden Damascene revelation; the axioms of fundamental measurement have been formalized for almost 80 years. There have been numerous warnings that technology assessment was paying insufficient (if any) attention to measurement scales. The result is that in a successful attempt to put the standards of normal science to one side, technology assessment and the embrace of the reference case has finally run into a brick wall. An easy way out, a short cut, to manufacture approximate evidence to justify imaginary cost-effectiveness claims has proven to be a monumental mistake.

A Commitment to Honesty

Looking back over 30 years of claims for cost-effectiveness, the observer might feel that all too many readers have been shortchanged. There are thousands of studies, catalogued in the Tufts emporium for cost-per I-QALY claims that are, to be honest, junk [21]. How this bone-yard of claims has been allowed to accumulate, promoted by academic centers, by ISPOR and by journal editors is remarkable. A situation that is unheard of in the physical sciences and by those in the social sciences, including mainstream economics. And yet it continues; ICER continues to create evidence reports that are nonsense; journals continue to accept imaginary cost-effectiveness claims and ISPOR continues to propagandize for the ratio I-QALY and the wonderful worlds of the value assessment imagination.

When the emperor has no clothes, there are few options. One is to admit that, indeed, I have no clothes and hurriedly depart for the nearest men's outfitters; another is to take a defensive position which argues that, yes, there are issues but the landscape of value assessment clothing is changing. The previous pseudoscientific constructs are pushed under the carpet with a commitment for a new dawn of alternative frameworks. The I-QALY fades into the background. Looking to the future of a technology assessment golden age; the past can be buried.

Conclusions

There can be no doubt that the past 30 years of health technology assessment and the thousands of published studies rest on assumptions that are undeniably false. The I-QALY is center stage in this charade. This is unlikely to be acknowledged; instead we will have a fudge. But this creates problems? To dispense with the I-QALY effectively destroys ICER, ISPOR and NICE reference case models. There is no halfway house. All those consulting dollars are in vain. Manufacturers supporting ICER are actually supporting a fantasy product; they might be better taking out a subscription to *Psychic News*.

We are talking about a paradigm shift which does not build on a previous accepted framework of analysis and accommodate it, but destroys it. It says, quite clearly, we were wrong. Careers have been devoted to an analytical dead end; degrees have been awarded; consulting income has flowed, manufacturers have emptied their pockets – but to accomplish nothing but imaginary I-QALY claims that have been taken, mistakenly, at face value. Perhaps one comparison is the shift from believing in a geocentric to a heliocentric model of the solar system. This was not without its risks; the church was prepared to stand firm regardless of logic or the evidence as Giordano Bruno (1548 – 1600) and Galileo Galilei (1564 – 1642) found out. In both cases the Vatican has yet to make an effective apology. Perhaps we need a new scientific revolution in health technology assessment; a revolution that endorses the move from imaginary non-evaluable claims to hypotheses and the discovery of new facts; a move from imaginary to real world evidence. We have accepted imaginary I-QALY claims for products for far too long. As the motto of the Royal Society (1600; charter 1662) makes clear *nullius in verba*: take no man's word for it; let alone a paradigm built upon a denial of the axioms of fundamental measurement.

REFERENCES

[1] Neumann P, Willke R, Garrison J. A Health Economics Approach to US Value Assessment Frameworks – An ISPOR T99ask Force Report. Value Health. 2018;21:119-123

[2] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-20234 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12

[3] Popper KR., The logic of scientific discovery .New York: Harper, 1959.

[4] Lakatos I, Musgrave A (eds.). Criticism and the growth of knowledge. Cambridge: University Press, 1970.

[5] Magee B. Popper. London: Fontana 1973

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

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

[8] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-680

[9] 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

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

[11] Langley PC. Sunlit uplands: the genius of the NICE reference case. *Inov Pharm*. 2016;7(2): No.12.

[12] Richardson J, Hawthorne G. Negative utility scores and evaluating the AQoL worst health state. Working Paper 113. Centre for Health Program Evaluation. Monash University. June 2001

[13] 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

[14] McKenna S, Heaney A, Wilburn J. Measurement of patient reported outcomes 2: Are current measures failing us? *J Med Econ*. 2019;22(6):S23-30993

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

[16] Luce R, Tukey J. Simultaneous Conjoint Measurement: A new type of fundamental measurement. *J Math Psychology*. 1964;1(1):1-27

[17] Rasch G. Probabilistic Models for some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedagogiske Institut, 1960

[18] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes. New York: Oxford university Press, 2015

[19] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21(5):475-80

[20] 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

[21] Tufts Medical Center. Center for Evaluation of Value and Risk in Health (CEVR). Cost-effectiveness Analysis (CEA) Registry. https://cevr.tuftsmedicalcenter.org/databases/cea-registry