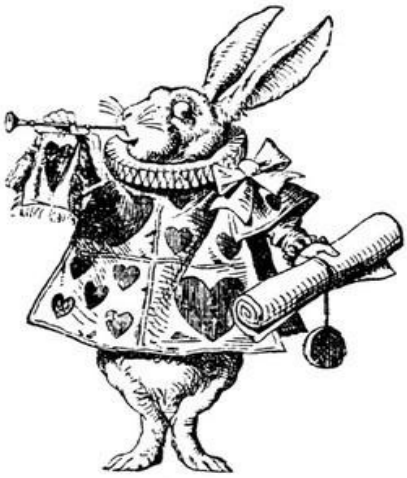
Nullius in verba



GUIDELINES FOR FORMULARY EVALUATIONS

[PROPOSED]

PROGRAM IN SOCIAL AND ADMINISTRATIVE PHARMACY

COLLEGE OF PHARMACY

UNIVERSITY OF MINNESOTA

Paul C Langley Ph.D., Adjunct Professor

Version 3.0: October 2020

For further information on implementing these guidelines please contact:

GFE, College of Pharmacy, University of Minnesota

308 SE Harvard Street, Minneapolis MN

55455 Tel: (612) 624-1000

Email: schom010@umn.edu

FOREWORD

The proposed Minnesota Guidelines for Formulary Evaluations (Version 3.0) are focused on credible, evaluable and replicable value claims for pharmaceutical products and devices. In common with the two previous versions of these guidelines, the emphasis is on real world evidence; not on the construction of incremental cost-per-I-QALY or similar imaginary worlds. While the construction of imaginary worlds is the standard advocated by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Institute for Clinical and Economic Review (ICER), this evaluation paradigm fails the demarcation test for normal science. These guidelines reject this paradigm. There are three main reasons: First, health care decisions should not be based on the construction of one of possibly many competing one-off imaginary worlds which are seen as providing 'approximate information; second, the failure by those committed to the construction of imaginary worlds to recognize that the I-QALY fails to meet the required axioms of fundamental measurement; this means that cost-per-I-QALY claims, which are the core of imaginary modeled claims, make no sense from a mathematical standpoint; and, third, claims must be focused on specific product attributes; to recognize dimensional homogeneity and the application of fundamental measurement..

If there is one overarching theme that drives these guidelines it is a belief in the need for recognizing and accepting the axioms of fundamental measurement. Rejecting fundamental measurement, the unfortunate hallmark of current standards in health technology assessment, means that value assessment claims for pharmaceutical products and devices are untenable. The paradigm is an analytical dead end.

ISPOR and ICER, and various academic centers promoting their reference case simulation models fail to appreciate that health care decisions should not be based on assumption driven imaginary constructs that look forward for 10, 20 or 30 years; a robust by unknown future assumption based reality.. Claims made must be empirically evaluable. Instruments must also recognize the axioms of fundamental measurement as unidimensional constructs. This is the standard for normal science: the discovery of new facts. Not a medieval dogma that only recognizes imaginary constructs; rejecting hypothesis testing Unfortunately this dogma, or to use a more modern term, meme, has held primary position in health technology assessment for the past 30 years. This is an absurd situation. Rather than mandating the discovery of new facts to support formulary decisions and the continued support for products and devices as part of ongoing disease area and therapeutic reviews, the health technology assessment meme asks us to suspend belief in hypothesis testing in favor of an endless creation of imaginary worlds. We must move on from imaginary world evidence to real world evidence.

The purpose of these guidelines is to propose acceptance of an evidence based, not a fantasy based, formulary. A formulary based on the tracking of outcomes in real time. Rejecting imaginative or fantasy worlds means a commitment to a research program to track and support treatment claims. This is unexceptionable, providing a firm basis for value based contracting over the patent lifetime of products, in particular biopharmaceuticals. Unfortunately, it will receive considerable opposition. After all, a fantasy cost-effectiveness claim is easier to create than a program of discovery for treatment impacts. While it may seem surprising, thousands of

researchers and health system decision makers have been prepared to accept imaginary claims rather than empirically evaluable claims in formulary decisions for over 30 years. It is time to consign imaginary claims to the rubbish bin and, hopefully, convince at least some of the folly of this meme. This is made easier by the failure of those promoting the construction of cost-per-I-QALY or imaginary QALY worlds to recognize the axioms of fundamental measurement and that the I-QALY is an impossible mathematical construct. The result: lifetime cost-per-I-QALY claims and thresholds are, apart from rejecting the standards of normal science, nothing more than mathematically impossible constructs which are expected to serve as the basis value assessment claims for 'fair' pricing and product access recommendations.

These guidelines are not intended just for formulary committees or health system decision makers in the private health sector. They have a global application in formulary decision making whether it is a group such as Anthem or public sector decision makers in Medicare, Medicaid and agencies such as the Veterans Administration.

CITATION

Langley P. Guidelines for Formulary Evaluations [Proposed]. Program in Social and Administrative Pharmacy, College of Pharmacy, University of Minnesota. Version 3.0. October 2020. https://www.maimonresearch.net/minnesota-guidelines/

THE ILLUSTRATIONS

The illustrations are by Sir John Tenniel (1820 – 1914) for *Alice's Adventures in Wonderland* (1865) and *Through the Looking Glass and What Alice Found There* (1871). As the principal argument for these guidelines is to abandon the wonderful and imaginary worlds of cost-utility value assessment in favor of real world evidence, the choice seems apt. We can all appreciate well-crafted imaginary tales and fantasy constructs. However, at some time we must exit the rabbit hole, thank Alice, and address evidence driven decision making in the real world.

TABLE OF CONTENTS

1. A NEW PARADIGM FOR VALUE ASSESSMENT

The preparation of submissions for formulary committees, insurers and other health system decision makers is a key step in the process of pricing and access negotiations for pharmaceutical products and devices. Unfortunately, guidelines proposed by agencies such as the National Institute for Health and Care Excellence (NICE) in the UK, the Pharmaceutical Benefits Advisory Committee in Australia (PBAC), the Academy of Managed Care Pharmacy (AMCP) in the US and, last but not least, the Institute for Clinical and Economic Review (ICER) in the US ask manufacturers to put the scientific standards for credibility, evaluation and replication of claims for cost-effectiveness to one side in favor of the construction of imaginary modeled claims. While this may seem odd, this approach is well established and has been accepted for 30 years. We are, literally, awash in modeled claims which are impossible to evaluate and were never intended to be evaluated. We are asked to take their word for it. This is unacceptable: *Nullius in Verba…take no man's word for it.*

These proposed guidelines present a completely different perspective, a new paradigm, the embrace of the standards of normal science; the process of the discovery of new facts and the assessment of credible claims for new and comparator products through a process of conjecture and refutation. Rather than creating an imaginary reference case world at product launch, on limited data and many assumptions, the focus here is on the creation of an evidence base to support product comparative claims for benefits and resource utilization. A commitment, not to imaginary world evidence, but to real world evidence; the rejection of pseudoscience (or bunk) in favor of the standards of normal science to meet critical evidence gaps.

If there is one term that captures the essence of the proposed new paradigm it is 'fundamental measurement'. All claims, whether they are patient reported for clinical endpoints or quality of life must meet the standards for fundamental measurement. This applies equally to claims for resource utilization and social impact. Claims must be dimensionally homogeneous and report therapy response on either interval or ratio scales. Ordinal scales, the bane be of health technology assessment claims, are non-starters.

These guidelines differ also in their focus on target patient populations within disease areas. All claims must refer to that population. Certainly, there may be wider societal issue that might be addressed, but not to the extent of creating imaginary cost-per-QALY simulations based on generic multiattribute raw scores in a vain yet impossible hope that these constructs might, in some way, contribute to decisions on resource allocation within health systems. The focus here is

on the patient and, where necessary, the caregiver. They are the primary target and hopefully beneficiaries of new and innovative therapies in rare and chronic diseases. From this perspective, attempts to build imaginary willingness to pay value frameworks are an analytical dead end.

The last 30 years have witnessed what may be described as an I-QALY disaster: a commitment to the creation of imaginary world evidence, in defiance of the standards of normal science, rather than a commitment to real world evidence [1]. The emphasis on the I-QALY modeled simulation as a basis for a comprehensive claim for cost-effectiveness is misplaced; we should instead focus on a basket of claims covering clinical, quality of life and resource utilization attributes.ds

1.1 MEETING THE STANDARDS OF NORMAL SCIENCE

If formulary decisions to admit or retain pharmaceutical products and devices are to be credible they must recognize the standards of normal science. The focus of these *Guidelines* lies in the recognition that if claims are made to support pharmaceutical products and devices, they must be credible, evaluable and reproducible [1]. Claims must be presented in a testable form that allows feedback to formulary committees as part of initial product assessments and ongoing disease area and therapeutic reviews. This evidentiary standard applies equally to claims for comparative clinical effectiveness as well as to claims for quality of life and resource utilization. If non-evaluable claims are presented by manufacturers then they should either be reformulated or put to one side.

1.1.1 Testable Hypotheses

The requirement for testable hypotheses in the evaluation and provisional acceptance of claims made for products and devices is unexceptional. Since the 17th century it has been accepted that if a research agenda is to advance, if there is to be an accretion of knowledge, there has to be a process of discovering new facts. Indeed, as early as the 16th century Leonardo da Vinci (1452 – 1519) in notes that appeared posthumously in 1540 for his *Treatise on Painting* (published in 1641) clearly anticipated the standards for the scientific method which were widely embraced a century later in rejecting thought experiments that fail the test of experience. By the 1660s, the scientific method, following the seminal contributions of Bacon, Galileo, Huygens and Boyle, had been clearly articulated by associations such as the Academia del Cimento in Florence (1657) and the Royal Society in England (founded 1660; Royal Charter 1662) with their respective mottos *Provando e Riprovando* (prove and again prove) and *nullius in verba* (take no man's word for it) [2].

By the early 20th century standards for empirical assessment were put on a sound methodological basis by Popper (Sir Karl Popper 1902-1994) in his advocacy of a process of 'conjecture and refutation [3] [4] . Hypotheses or claims must be capable of falsification; indeed they should be framed in such a way that makes falsification likely. Life becomes more interesting if claims are falsified because this forces us to reconsider our models and the assumptions built into those models. This leads to the obvious point that claims or models should not be judged on the realism or reasonableness of assumptions or on whether the model 'represents' for a public advocacy

research group such as ICER their belief in lifetime comparative cost-per-QALY outcomes future reality. A future reality that is unknown and unknowable, and is never intended to be known.

Popper's view on what demarcates science (e.g., natural selection) from pseudoscience (e.g., intelligent design) is now seen as an oversimplification involving more than just the criteria of falsification, the demarcation problem remains [5]. Certainly, there are different ways of doing science but what all scientific inquiry has in common is the 'construction of empirically verifiable theories and hypotheses'. Empirical testability is the 'one major characteristic distinguishing science from pseudoscience'; theories must be tested against data. Indeed, paradoxically, while the development of pharmaceutical products and the evidence standards required by the Food and Drug Administration (FDA) for product evaluation and marketing approval is driven by adherence to the scientific method, once a product is launched and claims made for cost-effectiveness and, in the case of ICER, modeled fair pricing and access recommendations, the scientific method is put to one side: pseudoscience succeeds science.

The rejection of a research program (the term is used loosely in technology assessment) that fails to meet the standards of normal science is possibly best exemplified by the latest version of the Canadian health technology guidelines where it is stated: *Economic evaluations are designed to inform decisions. As such they are distinct from conventional research activities, which are designed to test hypotheses* [6]. While this position puts modeled health technology assessment in the category of pseudoscience, it is also what may be described as a relativist position. Rather than subscribing to the position that the standards of normal science are the only standards to apply in health care decisions and value claims, the relativist believes that all perspectives are equally valid. Health care decisions are to be understood sociologically. No one body of evidence is superior to another. Results of a lifetime modeled simulation are on an equal basis with those of a pivotal Phase 3 randomized clinical trial. For the relativist, the success of a scientific research program, in this case one built on hypothetical models and simulations, rests not on its ability to generate new knowledge but on its ability to mobilize the support of the community. Basing decisions on models and simulations underpins the consensus view that evidence is constructed, never discovered. Instead of coming to grips with reality, science is about rhetoric, persuasion and authority [11]. Truth is consensus.

1.1.2 Assumptions and Imaginary Worlds

It is accepted that knowledge is provisional and permanently so. This stems from the obvious point that we can at no stage prove that what we 'know' is true. Attempting to believe or justify our belief in a theory is logically impossible. What we can do, by empirical assessment, is to try and demonstrate our preference for one theory over another (and apply it to the best of our knowledge).

Constructing imaginary worlds which were never intended to generate potentially falsifiable claims cannot, therefore, be defended by an appeal to the 'truth' of their assumptions. If a health technology assessment claim is built upon a series of assumptions, a reasonable question is to ask what is the

status of the various assumptions? Are they to be viewed as 'reasonable or 'realistic' metrics for an unknown future reality? Have they been selected from the literature because they seem appropriate? Are they the 'best available' from limited data? Have they been 'selected' to create imaginary claims? How much confidence can we have in ICER-type robust representation of an unknown future assumption driven reality?

More to the point, there is a belief that the fact that the selected assumptions are based, where feasible, on an empirical studies validates the choice of assumption. For example, if the model is intended to incorporate utilities that have been reported in one or two studies (usually as few as that) for progression and time spent in the stages of a disease over a hypothetical future lifetime, then there is an immediate logical issue. To claim that an assumption is valid is to revisit Hume's induction problem (David Hume 1711-1776): an appeal to facts to support a scientific statement. Unfortunately, as Hume pointed out, no number of singular observations can logically entail an unrestricted general statement. Certainly, there may be comfort in reporting that 'so far' the claim that all swans are white has not been contradicted (until that Qantas vacation in Western Australia) so that one fully expects the next swan to be white. But as Hume pointed out, this is a fact of psychology and does not entail any general statement. From a utility perspective, the fact that one hundred papers have agreed (within limited bounds) generic utilities from the same instrument for a target population in a disease state stage is immaterial. We cannot secure this assumption: it cannot be '*established by logical argument, since from the fact that all past futures have resembled past pasts, it does not follow that all future futures will resemble future pasts*' [7]. Claims, for the relevance of a constructed imaginary world built on the assumption, that the model elements have been validated by observation, is simply nonsensical.

Despite ICER's continued embrace, logical positivism is dead. It died some 80 years ago. All knowledge is provisional. Popper's contribution was to make clear that Hume's problem with induction can be resolved. We cannot prove the truth of a theory, or justify our belief in a theory or attendant assumptions, since this is to attempt the logically impossible. We can only justify our preference for a theory by continued evaluation and replication of claims. Constructing imaginary worlds, even if the justification is that they are 'for information' is, to use Bentham's (Jeremy Bentham (1748-1832) memorable phrase 'nonsense on stilts'. If there is a belief, as subscribed to by ISPOR, ICER and others, in the sure and certain hope of the relevance of approximate information created by imaginary worlds, a belief to drive formulary and pricing decisions, then it needs to be made clear that this is a belief that lacks scientific merit. Belief does not trump reality.

Certainly assumptions can be a critical element in models; the difference is that these models should support testable hypotheses. This is echoed by Newton (Isaac Newton 1642-1727), with Descartes as his target (René Descartes 1596-1650) in saying '*hypotheses non fingo*' (I do not feign hypotheses). Descartes in Newton's view had 'produced fantastic and untestable ideas, then assumed them to be true and used them as building blocks of his philosophy [8].

1.1.3 Approximate Information

A hallmark of the ISPOR value assessment framework is that it is designed to generate 'approximate information'. This follows from the rejection of hypothesis testing; all ISPOR has to offer are the imaginary and 'uncertain' lifetime value frameworks where, by definition, the

information content is imaginary. While there are attempts to remedy this through sensitivity analyses and scenarios, the fact remains that the information we are dealing with is not so much approximate as imaginary; to add a further layer of scenario and uncertainty analyses may seem a trifle redundant. While formulary committees might be willing to concede that some information is better than none if value claims are to be assessed, if it was made clear to them that the entire construct not only fails to meet the standards of normal science but that the crown jewel, the I-QALY is actually mathematically impossible, they may well put that assessment to one side.

While groups such as NICE in the UK and the PBAC in Australia might be willing to base formulary decisions on approximate information, with academic assessment groups apparently judging 'how approximate' the value claim is, it all seems rather ridiculous. After all, what is the 'approximate' imaginary information supposed to be approximate to? It is one argument to claim 'approximation' in the development of instruments to measure, directly (e.g., weight) or indirectly (e.g., temperature) in the physical sciences where successive instruments may be designed to be more accurate; it is another to claim that the value assessment ISPOR model is judged by assessors to be, following their ministrations, more approximate to an imaginary 'ideal' that stretches decades into an imagined future than an alternative simulation. Perhaps the 'truth is out there', a true value claim; if it is, it is entirely imaginary.

Information cannot be approximate if the value assessment framework, the ICER reference case, is a mathematically impossible construct. Utilities are raw scores on ordinal scales; this is incontrovertible. As such we cannot multiply time spent in a disease stage by an ordinal measure. Rather than approximate imaginary information we have a situation where the modeled information is impossible in the first place.

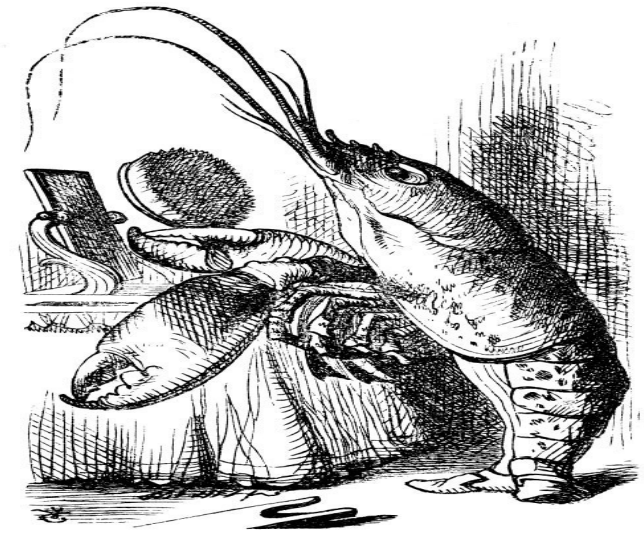
1.1.4 Validating Impossible Value Assessments

For those holding to a firm belief in the I-QALY paradigm, it is worth noting the process by which ICER academic modeling groups 'validate' their imaginary value assessment. The latest ICER evidence report on ulcerative colitis is a good example [9]. The authors of this report in the College of Pharmacy Modelling Group at the University of Washington, make the following statement:

> ***We used several approaches to validate the model and followed standard practices in the field. First, we provided preliminary methods and results to manufacturers, patient groups, and clinical experts. Based on feedback from these groups, we refined data inputs used in the model. We also shared the draft model with participating manufacturers during the public commenting period. We tested all mathematical functions in the model to ensure that they were consistent with the report. We also conducted sensitivity analyses with extreme input values to ensure the model was producing findings consistent with expectations. Finally, we compared results to other cost effectiveness models in this therapy area.***

This approach to 'validation' is at variance to the process of discovery in the physical and mainstream social sciences. There is no concept of hypothesis testing or of any attempt to generate claims that are credible evaluable and replicable. Any notion of real world evidence is absent. To defend this process by reference to 'standard practices' does nothing to convince the skeptic that this is anything more than a self-serving exercise to support a pseudoscientific paradigm [10].

1.1.5 The Patient Voice



The focus on the creation of imaginary worlds, the application of generic utilities to construct, in defiance of fundamental measurement axioms, quality of life claims has a major drawback. It ignores the interests and needs of patients within target populations. The EQ-5D-3L instrument, for example, operationalizes a clinician's view of population health. Health is defined in terms of symptoms and response levels within each symptom. The EQ-5D-3L comprises 5 symptoms (mobility, self-care, usual activity, pain/discomfort, anxiety/depression) and three response levels (no problem, some problems extreme problems). The utilities are constructed by an algorithm or equation that attaches weights to responses together with other data elements to produce raw scores.

The patient may respond but with the responses valued by community preferences. Claims for quality of life may have nothing to do with the patients 'valuation' of their quality of life. The final raw score utility is what the community feels your quality of life should be as defined by symptom and response levels. The needs of patients and caregivers and their ability to meet those needs are ignored.

Recent discussions within groups such as ISPOR have raised the question of alternative value frameworks to include issues that may be of concern to patients. Indeed, there is presently an effort underway in the UK to develop an E-QALY (extended QALY) to try and capture within one instrument other factors of interest to patients. Unfortunately, the 'old' I-QALY and its disdain for fundamental measurement will still be the core of the measure. If we are concerned with the patient voice specific to target populations in disease areas then this E-QALY endeavor is likely a waste of time; it will be generic and, again, will fail the standards not only of dimensional homogeneity but of needs fulfillment which should be central to quality of life response.

A strange feature of the ISPOR and ICER value assessment frameworks is that they are one-off. If you jump the imaginary hurdle, embrace the ICER fantasy world, then you can walk away from any attempts to track and report on the uptake of the therapy assessed and patient outcomes.

Of course, the ISPOR and ICER value assessment frameworks are designed to do this; we cannot track and report on imaginary claims and approximate information (even with Tarot cards). This is of course an easy way out: satisfy the formulary committee and you are home.

The ISPOR and ICER value assessment frameworks also deny the possibility of ongoing disease area and therapeutic class reviews. Regular evaluations by the health systems have a potentially pivotal role in reconsidering formulary placement and pricing. If this to occur then, as detailed in these guidelines, the health system requires agreed clinical, quality of life and associated outcome protocol measures than can be tracked and reported in real time. This, it is argued below, should be the responsibility of the manufacturer.

1.2 REJECTING IMAGINARY WORLDS

Central to the commitment for the fabrication of imaginary worlds is the belief in the role of 'approximate information' in health technology assessment decision making [proximus notitia: e.g., Centurion! The barbarians are somewhere north or south of the Alps]. The commitment to approximate information as opposed to hypothesis testing is unique to health technology assessment. As a proposed branch of economics, health technology assessment or pharmacoeconomics stands apart in its commitment to the construction of imaginary worlds. Mainstream economists, recognizing the importance of the distinction between positive and normative aspects of professional activities, would be somewhat bemused at the claim that the construction of imaginary worlds, on incorrect and illogical assumptions, are critical in generating, not claims for hypothesis testing, but the fabrication of approximate information to support formulary decisions.

The question, of course, is what the phrase 'approximate information' actually means, there is, no doubt, the opportunity to create, within the same reference framework and within a disease area, a multiverse of competing modeled 'approximate' claims. Do we have a 'my future assumption' is better than yours competition? Or do we, as in the case of NICE and the PBAC, nominate academic centers with years of experience in assessing imaginary claims to be the arbiters as real world referees or inquisitors for imaginary world claims. Do we establish in the US a cadre of inquisitors to judge the merits of competing models against Jesuitical standards set by a reference case? As self-appointed arbiter, ICER would, presumably, have a papal role. If the intention is to reduce uncertainty then we have no idea if uncertainty is reduced or increased. To paraphrase Wolfgang Pauli (1900-1958), a major contributor to quantum mechanics, "We don't know whether it is right of whether it is wrong; and we will never know". To which might be added "and we were never intended to know" [11].

1.2.1 Memes, ISPOR and ICER

If truth is consensus among those subscribing to the I-QALY paradigm, then how is this consensus, resting upon the construction of imaginary worlds, maintained; in this case for over 30 years of imaginary cost-per I-QALY modeled claims? The ISPOR consensus, embraced by ICER, on health technology assessment has been characterized in previous commentaries as a meme. This is deliberate, as it underpins the interpretation of ICER's continued unqualified acceptance of the reference case as its core business model, as a sociological phenomenon; a belief system or faith in elements such as the I-QALY which defy both reason and evidence. .

After all, it is unusual to find as the central pillar in an apparently academically respectable discipline, the construction of fictional imaginary worlds; in this case to support non-evaluable cost-outcomes claims with the creation of 'approximate information'. In this context, the ISPOR/ICER cost-per-I-QALY reference case can be characterized as a unit of cultural transmission or unit of imitation; as an analog of gene pool propagation 'by leaping from body to body via sperm or eggs' [12] .

One of the key health technology assessment meme tenets is the belief in the I-QALY; the venerated dogma which is central to value assessment. Utility scales have (or must have) ratio properties; ICER certainly believes it: ICER has, in their own words' an 'understanding, like most health economists', that utility scales are ratio scales'. By some mysterious alchemy, as we will discuss later, the utility raw scores on an ordinal scale are actually seen as ratio scales in disguise. While this is arrant nonsense, the key to this long-held belief lies in the transmission fidelity of the I-QALY meme. .

Human beings are good at imitation. The reference case meme, the faith in the I-QALY, appears to be adept in its infectivity, supported by an organizational infrastructure to defend it against competition in the technology assessment meme pool; ensuring survival through supporting propagation, longevity, fecundity (or acceptability) and, of particular note, high copying fidelity. The control exercised over the meme ensures few mutations. As Dawkins notes, few individuals brought up in a certain faith switch to other faiths or reject the 'faith and mysteries' of their parents' belief system[13]. With such an infrastructure, and punishment for transgression, belief systems can survive for hundreds if not thousands of years.

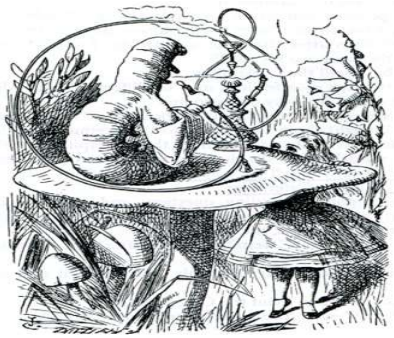
The widespread adoption and propagation of this meme in seen with literally thousands of imaginary world technology assessments published over the past 30 plus years (the latest cumulative PubMed count of references with the search term 'QALY' is from 1981 almost

19,000 as of 3 October 2020). Add to this the continued willingness of journal editors to publish imaginary claims, even if they are sponsored marketing exercises favoring the sponsor's product. In the case of this continued acceptance it is sufficient to point to the advocacy of the meme by organizations such as ISPOR with its global membership (some 20,000), its good practice guidelines for constructing imaginary worlds and defying fundamental measurement, training programs for newly arrived imaginary world apprentices, and conferences, together with endorsements from technology assessment agencies such as NICE, CADTH and the PBAC. Add to this its place in university post-graduate programs (including Colleges of Pharmacy) together with the contribution of textbooks that have rigorously supported the creation of imaginary worlds and attendant 'techniques' for non-evaluable likelihood claims (e.g., probabilistic sensitivity analysis) [14].

It is of interest to speculate, given the receptive audience for imaginary technology assessment claims, together with the 'technical' belief structure that underpins them, whether or not we are as students and academics receptive to pseudoscientific claims; is there a response bias toward accepting pseudoscientific claims as true? Is there an asymmetry between belief and unbelief? Is additional processing required if this bias is to be overcome? Is there an asymmetry that reinforces acceptance of the technology assessment meme and the acceptance of 'imaginary approximate information' even though it is a 'mystery' as to what this actually means? Perhaps we just accept it on 'faith'? Or do we subscribe to Tertullian (155-240? AD) an early Christian author: *Certum est quia impossible est* (It is certain because it is impossible) [11].

A further possibility is our inability to detect pseudoscientific constructs. Do we judge something as profound because we have failed to understand it? Are there measurable differences in the ability of individuals to discern or detect pseudoscientific statements including the more complex (and often obscure) modeling constructs supporting ICER imaginary worlds and attendant scenarios? Can we engage in analytic thinking? Do we understand the axioms of fundamental measurement? Have we ever been taught the meaning of fundamental measurement? To what extent is our ability to reflect on, rather than reflexively accept at face value, offset by our acceptance of a belief system that is central to our professional standing?

Perhaps we should not be surprised that the nature of the scientific method is not appreciated. After all, some 27% of Americans don't accept heliocentrism, 48% don't accept common ancestry (natural selection) and 61% don't accept the big bang. Even so, we should not be unduly pessimistic as a recent survey indicated that probably less than 2% of Americans believe in a flat earth, although globally traveling flat earth advocates seem active on the conference front. . There are, of course those who require visual evidence. One respondent remarked that he did not believe in gravity because he could not see it. Presumably he did not believe in imaginary worlds either.



1.2.2 The Imaginary World of ICER

ICER has taken upon itself the role of US national arbiter in value assessment for pharmaceutical products and devices. Fanfares accompany its media releases laying down the

Results of its latest product specific reference case recommendations for product pricing and access. Although recipients of ICER recommendations should know better, they are actually taken seriously. Those accepting ICER recommendations for pricing and access, let alone national imaginary budget impacts, have only the most rudimentary understanding of the imaginary ICER modeling framework, let alone the limitations imposed on measures such as the I-QALY of the axioms of fundamental measurement.

As will be detailed in the next few sections the ICER model, in focusing on incremental cost-per-I-QALY estimates and thresholds to determine 'value' fails on a number of counts: apart from asking the health decision maker to take seriously the information content of an assumption driven simulation tracking some decades into the future. The ICER modeling ignores the requirements of fundamental measurement. ICER is fixated on HRQoL measures represented by generic multi-attribute preference measures. As noted, ICER fails to recognize that not only are its threshold claims invalid in their neglect of fundamental measurement, but that with a potential multiverse of models which all conform to the ICER reference case, decision makers have a potential multitude of imaginary 'approximate information' or 'impossible information' creations to choose from.

ICER has just released, its latest version of the value assessment framework (January 31, 2020)[15]. This maintains ICERs commitment to the construction of assumption driven imaginary worlds. A recent review of the ICER revised value assessment framework provides a detailed critique [16]. It is congruent with the position taken here in the proposed Minnesota guidelines. The need is to make clear to a wider audience the lack of scientific method in the ICER value assessment framework. This was a major factor behind the release of these guidelines. Added to this is the failure to defend a key construct in value assessment: needs fulfilment. The ICER framework is focused on HRQoL, not to a broader patient centric framework that captures QoL in target patient populations. The entire exercise, to emphasize a major error, collapses because of a failure to recognize the constraints imposed by fundamental measurement.

1.2.3 Fair Price and Fair Value

In economics, fair value is a rational and unbiased estimate of the potential market price of a good, service or asset. Under the efficient market hypothesis this proposition holds in a well-organized and reasonably transparent market. The key is information symmetry and transparency. In terms of market structure, the pharmaceutical market diverges in a number of respects from the efficient market ideal. There are multiple stakeholders, it is assumed there is significant (undefined) information and incentive asymmetry between providers, third party payers and manufacturers, including patent protection [17]. The presumption is then made that existing markets do not resolve fair drug pricing; 'drug prices are not fair in the US or abroad'.

Certainly, on benchmark comparisons, prices are higher in the US than abroad (there are also significant global price differences for the Big Mac); this does not mean that there is any necessarily significant divergence between fair price and fair value in the US with its higher prices. More to the point is an obvious question: if we want to test this hypothesis how would we proceed? Given the role of the Big Mac in the US diet, it is surprising that the various health departments have not intervened to bring the Big Mac price in line with a basket of Big Mac country prices. Assuming, of course, that the exchange between the US and these countries is not

over or under valued for pricing comparisons. The same would apply to pharmaceuticals, which explains, in part, the objections to pharmaceutical cross-country average price comparisons. As the UK is often put forward as a price comparison it should be noted that NICE uses the I-QALY in its reference case threshold assessments for pricing.

Whether this conclusion on fair pricing is warranted is a moot point. At product launch there is clearly only limited information on product performance with at best phase 2 and phase 3 trials (the latter placebo controlled). This has led, as detailed above, to the I-QALY and the substitution of assumptions for evidence in fantasy worlds. Information is not, of course, a fixed quanta. Following market entry we see information feedback, some in the public domain, other proprietary. The key point is that transparency is probably dwarfed by the absence of information. This puts a premium on activities to identify and meet information gaps, which in the case of the Minnesota guidelines implies information feedback from manufacturers to formulary committees and other health system decision makers, with further release to patients and providers. Whether this should be seen as improving transparency (as though information was being deliberately withheld) as opposed to a general increase in the amount and the quality of information is a reasonable question. This points to the notion that markets are adaptive with the degree of efficiency, determined by information, changing over time [18]

It is at this stage that the construction of imaginary reference case simulation models becomes an option to judge fair price, as detailed in a recent White Paper produced by ICER and the Office of Health Economics (OHE) [19]. It is also at this stage that we are tempted to put the standards of normal science to one side and, notably, the axioms of fundamental measurement. In the US, ICER has become the arbiter of a 'fair' price. This is achieved by cost-per-I-QALY thresholds and constructing imaginary scenarios, by assumption, where community preferences for generic health outcomes drive a valuation. The ICER model is, in a sense, the surrogate efficient market of economic theory in price determination.

This is, of course, nonsense. The mathematically impossible I-QALY effectively demolishes this attempt to create a fair price. We are, in practical terms, forced to fall back on a simple prescription: negotiation on the basis of information available. Prices negotiated are provisional and, as detailed in these guidelines, subject to revision as evidence gaps are filled, including response to therapy in the real world.

1.2.4 Academy of Managed Care Pharmacy

ICER is not alone in the US in promoting the construction of 'for information only' health technology assessment imaginary worlds. The recently released Academy of Managed Care Pharmacy (AMCP) Format for Formulary Submissions (Format 4.1) joins ICER in proposing standards for economic evaluation which also fail the demarcation test: again, the approach described clearly meets the standards for pseudoscience [20]. This argument

has been made earlier in respect of Format 4.0 [21] [22]. The principal difference in the newly released format is that it is extended to unapproved products and unapproved uses of approved products.

Although, unlike ICER, AMCP does not mandate a reference case to inform the creation of imaginary modeled worlds, the role of the AMCP Format is to inform: *to communicate clinical and economic evidence and information to health care decision makers*. There is no concept within the AMCP Format of the role of the scientific method in evaluating claims. Certainly the AMCP Format encourages an ongoing dialogue between manufacturers and heath care decision makers in imaginary information provision, which could presumably be extended to cover the evaluation of credible claims, but there is nothing to suggest that, if the process of discovery is to be applied in health care decision making, that hypothesis testing has any role to play.

AMCP clearly endorses the construction of imaginary lifetime model worlds, it also points to the importance of I-QALYs in cost-outcomes modeling. What it fails to consider, together with ICER, are issues which point to the objections raised here in terms of, not only the pseudoscience of imaginary information and modeling claims by assumption, which can generate a multiverse of competing yet elegant (with claims to realism) modeled claims, but the I-QALY as a meaningless construct.

1.3 MEETING THE STANDARDS FOR REAL WORLD EVIDENCE

Presumably, the beneficial target for any therapy intervention is the patient. While a clinician may be focused on clinical markers, these may say nothing about the benefit that is perceived by patients in the target population. The EQ-5D-3L is a case in point. This multiattribute instrument was designed with limited patient input, but had to be from the clinical perspective, simple and easy to respond to by patients. Its relevance to specific disease states is questionable; it is doubtful if it correlates with the patient's assessment of benefit. The patient's perspective is, of course, ignored as the various responses are weighted to match community preferences; the value the community places on the responses by patients in a disease area. The community decides on the 'impact' of a disease state; a central health planning perspective for resource allocation; an ordinal metric which is unlikely ever to be applied in practice.

Contrast the EQ-5D-3L with the concept of meeting the needs of patients in disease areas. Forget trying to develop a generic measure applicable across all known disease states, focus rather on a patient-centric measure that looks to a specific disease state and the needs of patients and caregivers in a target patient population. This can be achieved by taking as the latent construct for instrument development the fulfillment of needs which patients and caregivers identify in their disease state. Some needs may be relatively easy to meet with a new therapy, compared to

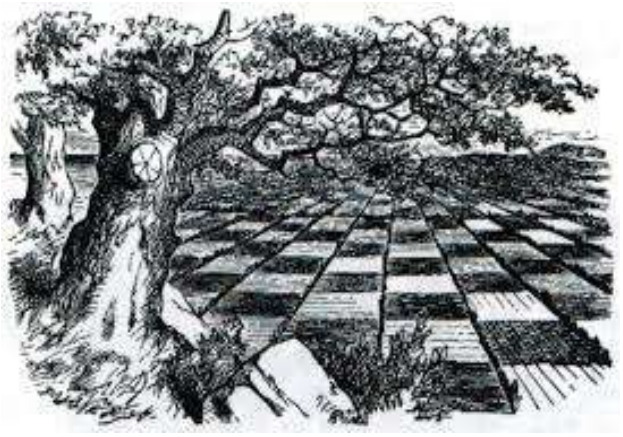
standard of care, others may be more difficult. At the same time patients will differ in their ability to meet those needs.

Unidimensionality or dimensional homogeneity is critical. In the physical sciences care is taken to measure one attribute as a time. Certainly response to therapy is a focus, but rather than trying to jam different attributes into one measure, we must aim to capture one attribute at a time. Multiattribute scores are a less useful summary (analysts will insist on trying to tease out the attributes contained in the score). We can report on instruments capturing individual attributes as part of a reporting protocol and we can even add qualitative assessments as complements. For those wedded to aggregate ordinal rankings, we might even apply Multi-criteria Decision Analysis (MCDA) to create an overall subjective ordinal weighted score of individual unidimensional interval scales and various add-on items (a hope index?) [23]. Any attempts beyond ordinal rankings to effect comparisons would, obviously, fail the axioms of fundamental measurement. This is the mathematical brick wall that attempts to 'enhance' value assessment frameworks face, including those where the I-QALY is still center-piece and their supporters have yet to recognize the impossibility of the construct.

The needs fulfillment model in therapy intervention assessment is well established and it has been debated in the literature from the past 25 plus years. Central to the development of such instruments is Rasch Measurement Theory (RMT). As detailed below, RMT rests on an acceptance that there is single non-physical attribute or construct that is to be measured. If the focus is on the patient, then the focus should be on a needs-fulfillment quality of life (QoL) construct. If health status is the major contributor to an individual's overall quality of lie, then the impact of new therapies need to assessed in terms of the needs of patients being met.

Needs fulfillment is quite removed from the clinical and operations elements that are found in the majority of HRQoL measures. Not only do these measures lack an appreciation of the notion of a latent construct that may sit behind their attempts to 'capture' the construct, but the individual symptoms they attempt to capture may represent a separate latent construct from the other symptom measures. It is truly a dog's breakfast of attributes and latent constructs.

1.3.1 Fundamental Measurement



To appreciate the contribution of RMT to the construction of needs-fulfillment quality of life instruments, it is important to make clear the axioms of fundamental measurement. Surprisingly, groups such as ISPOR and ICER seem to have a singular lack of awareness of fundamental measurement. This has led to their promotion of I-QALY constructs which fail these axiom.

The quality of measurement in health technology assessment should match the quality of measurement in the physical sciences. The properties of the scales of measurement are detailed in Figure 1. Patient value requires an

instrument that allows single attribute changes to be interpreted through the calibration of a unique interval score for the disease in the target patient population. The score might allow also, if required, co-calibration across disease states. If the disease specific measures embody the same theoretical measurement basis and common items, this co-calibration can occur. Unfortunately, in health technology assessment, ordinal raw scores rather than interval and ratio measures hold center stage. Where these scores, in the case of utilities for example, are constrained to an interval (e.g., 0 – 1) there will be a distortion of intervals to the margin, let alone floor and ceiling effects. This means that depending on the starting point in the scale, changes in the raw score will differ. This is not acceptable. Add to this the fact that the utility scores can take negative values (e.g., EQ-5D-3L has a range from -0.59 to 1.0). For a scale to be accepted it must conform to the two general axioms of measurement theory: invariance of comparisons and sufficiency[24]. These requirements have been recognized and applied through the application of RMT in education and psychology for the past 60 years [25]. There is a growing acceptance of the necessity of interval measurement in health technology assessment, although it has yet to achieve wide recognition.

There is no PRO database that appears to even contemplate fundamental measurement. MAPI values fails to flag the measurement status of the hundreds of PROs it has faithfully tabulated; there is no mention of RMT. Indeed. As the premier database for PRO instruments, MAPI says nothing as to whether or not the instruments meets required RMT measurement standards for calibrating response to therapy (i.e., has demonstrable unidimensional interval properties) [26]. This failure to recognize measurement standards is also seen in the Tufts utility emporium, a Cost-Effectiveness Analysis (CEA) Registry of over 8,000 papers, supported at the Tufts Medical Center by the Center for the Evaluation of Value and Risk in Health. Those creating and updating the database appears not to have recognized for over 30 or more years that utility scales are ordinal raw scores [27]. Rather, it provides summaries of cost-utility studies, detailing constructed I-QALYs and cost-per-I-QALY claims. Reporting on I-QALYs and imaginary cost-per-I-QALY claims from published studies seems a complete waste of time.

In the case of health technology assessments, where generic utilities are calibrated to support I-QALY claims, the various multi-attribute instruments and preference scales fail to meet Rasch measurement standards. The consequent I-QALY claims are, by definition, mathematically impossible [28]. Needless to say the same conclusion applies to the overwhelming number of PRO instruments. While they might meet classical test theory (CTT) standards and Item Response Theory (IRT) for instrument development, they fail to achieve Rasch standards; again, they must be rejected. Care must also be taken in accepting at face value claims by instrument developers that their instrument meets RMT standards and should be considered to have cardinal and unidimensional properties. This is seldom the case. As McKenna et al point out, while a few PRO measures attempt to claim they meet RMT standards, those that do fail to demonstrate that this is in fact the case. Obviously, the majority of those either using or developing PRO instruments, whether generic or disease specific, missed the memo. Fortunately, however, there is a growing appreciation of the need to develop RMT standard PRO measures with over 30 now available. These have all taken a patient centric approach; that is, the focus has been on needs fulfilment to drive quality of life (QoL) rather than health related quality of life (HRQoL) as their point of departure.

FIGURE 1

SCALES OF MEASUREMENT

The measurement scales used in statistical analysis are nominal, ordinal, interval and ratio.

Each scale of measurement meets one or more of the following properties:

- ***Identity: Each value of a scale has a unique meaning (i.e., a descriptive category) with no inherent numerical value in respect of magnitude such as gender, race***
- ***Magnitude: values on the measurement scale have an ordered relationship to each other (i.e., larger or smaller) but we don't know the distance between them; it has the properties of identity and magnitude***
- ***Interval: scale units are equal to each other in an ordered relationship where the distances are known (i.e., by how much larger or smaller; but not how far from zero it is); it has the properties of identity, magnitude and equal intervals; the presence of zero is arbitrary)***
- ***Ratio Scale: the scale has a 'true zero' or a minimum value of zero (e.g., on a weight scale there can be no weight less than zero). A ratio scale has all four properties of identity, magnitude, interval and ratio.***

Implications for statistical analysis:

- ***Scales with identity and magnitude properties can support only median and modal measures but no other operations***
- ***Scales with identity, magnitude and interval properties can only support the operations of addition and subtraction from any point on a real integer line (i.e., change the point relative to where it was before)***
- ***Scales with all four properties (a true zero) can support the further operations of multiplication and division (i.e., they can change the point on the integer line relative to zero)***

Unless designed to have interval or ratio properties, instruments will exhibit ordinal properties: they will be a ranking of raw scores. Applied to other quantities they will still generate ordinal scores (the operation makes no sense from a mathematical standpoint).

If you wish to create QALYs then the utilities must have ratio properties (a true zero) where the utility values (on an arbitrary scale of 0 – 1) meets all four measurement properties.

Unfortunately, utility values are ordinal: they only have identity and magnitude. They are not ratio scales, let alone interval scales.

==========

1.3.2 Composite Outcomes

A common characteristic of PRO measures is their composite nature. That is, they combine variables capturing different dimensions of health experience. This applies to both generic HRQoL instruments as well as to the majority of disease specific instruments. While composite PROs may be defended by their ability to capture and summarize in a single score a range of health outcomes they fail the axioms of fundamental measurement. Typically, the instrument fails to meet the standard for dimensional homogeneity, where variables can only be combined if they have the same dimension of health experience [29]. Otherwise they lack construct validity. They are multidimensional rather than unidimensional with ordinal scale properties. As such, they cannot assess response to therapy other than by ordering scores and applying non-parametric statistics.

There are some obvious limitations on composite measures. What is driving a change in the aggregate score? Do we need to disaggregate to obtain any meaning? How transparent is the score? How are score changes to be interpreted by clinicians? How were the items comprising the score selected and reported on by patients? Are the items weighted? On what basis were weights employed? Will different weights produce different outcomes? Are the items comprising the composite measure interrelated? How useful are composite measures in decision making? Is the composite measure effectively redundant? While these issues are relevant to all composite measures, the problems do not arise when we accept the role of single attributes and construct a unidimensional or dimensionally homogeneous measure.

1.3.3 Meeting Rasch Standards

If claims are to be accepted to meet the required measurement standards of these guidelines then it has to be made quite clear what the Rasch process involves. In a recent paper McKenna et al have detailed the step-by-step requirements (in this case for oncology patients) to develop an RMT outcome measure [30]. The important point to note is that the attribute measure has to be constructed to conform to the axioms of fundamental measure. In practice, this means creating an interval scale. This is where utility advocates fall down: they assume just because you can place raw scores on an interval number line that the scores themselves have interval properties. Of course, if there are negative utilities then the number line has to encompass negative as well as positive scores.

The steps are:

- Agree a coherent unidimensional latent construct measurement model to guide the selection of items from in-depth unstructured patient interview for those in the target disease state that cover all aspects of patient lives and not just those impacted by health interventions and services

- Ensure that the focus is disease specific so that the content is directly tailored to patient experiences while independent of the nature of the intervention
- Ensure that the measures are patient-centric, with the objective of identifying factors that most affect patients' lives, including the language used by respondents
- Create an index of outcomes rather than a profile on individual items so that the combined impact of illness and interventions is captured.
- Create a one-dimensional (unidimensional) scale where the items (selected from the pool of items) measure the latent construct and generate a total interval score that meets the standards of the Rasch modelling framework..

Where the object to be measured is a psychological or non-physical construct (e.g., quality of life) the situation in the social sciences is more complex. It was not until the early 1960s that recognition of the fundamental axioms of conjoint simultaneous measurement provided a framework for going beyond the notion of simple interval and ratio scales, to propose a framework for detecting, if they exist, measurement structures in non-physical attributes with constructed interval properties. That is, unlike ordinal scales where only the order of values matters and we can say nothing about the difference in the values, the interval scale is one where we know the order and the exact difference. Unlike the ordinal scale which allows only statements regarding the mode or median and the application of nonparametric statistics, interval scales allow addition and subtraction. This permits calculation of measures of central tendency and dispersion (e.g., effect size). However, as the interval scale does not have a true zero, we cannot compute ratios (i.e., multiplication and division). It is only with ratio scales that we have a true zero. The seminal contributions are those by Luce and Tukey, and Rasch [31] [32].

The critical step is to recognize the contribution of Rasch Measurement Theory (RMT) to constructing outcomes instruments in health technology assessment [24]. As noted below, the criteria for designating a scale as meeting the axioms of fundamental measurement, is to develop the instrument by application of Rasch model standards. Otherwise the instrument will generate nothing but ordinal scores.

The Rasch contribution is to recognize the need, if we are to develop the analog to measurement in the physical science, to produce the data (items in a questionnaire) to fit the Rasch model, not as in, for example Item Response Theory (IRT) and classical test theory (CTT), to fit the model to the data. As an example, for a mathematics test, a matrix may be defined by the ability of examination candidates (row elements) and the difficulty level of the various items in the test (column elements). Patterns of relationships between the cells, where each cell gives the probability of an outcome (Yes/No) as the difference between the difficulty of an item and the ability of the student can be determined by applying the axioms of conjoint simultaneous measurement.

The Rasch model, although developed independently of Luce and Tukey, utilizes a modified form of the axioms of conjoint simultaneous measurement, to assess patterns in a matrix of expected response probabilities; again as a function of differences between ability and difficulty [17]. The unidimensional Rasch model, a focus on a single attribute or homogeneous dimension captured in a latent construct, rests on two 'order' premises:

- The easier the item, the more likely it is to be affirmed; and
- The more able the respondent, the more likely are they to affirm an item

If the data items fit the Rasch model, they are necessarily translated from ordinal scores to interval scores where the unit of measurement is the logit or logs odd unit. The Rasch model rejects raw scores. Rather, a log-odds transformation is applied to these ordinal attribute measures to create a Rasch relative distance or interval measurement scale. This scale avoids the 'clumping' of raw scores around the middle scores and enhances the contrast in results for, in the case of ability, those at the extreme values of the scale. The purpose of the Rasch model is to build a measurement tool (a list of items, tasks, questions) that will make a meaningful assessment of a latent construct. Difficulty is relative to the other items in the scale. Each item on a unidimensional scale should contribute meaningfully to the construct being evaluated.

1.3.4 Rasch Confirmatory Analysis

Over the past 20 years there have been many examples where Rasch analysis has been used to evaluate a PRO instrument for potential interval properties and the creation of summary scores. In some cases the extent to which the original PRO ordinal scales fit the Rasch model has involved minimum item reduction. One example is the Gibbons et al report on a Rasch analysis of the Motor Neurone Disease (MND) Social Withdrawal Scale (SMS) [33]. Recognizing that the original instrument developed with classical test theory (CTT) would always, by definition, be ordinal, the 24 items in the original were assessed for their factor structure and evaluated for model fit, category threshold analysis, differential item functioning, dimensionality and local dependency. The four factor solution of the original instrument was confirmed with Mokken scale analysis suggesting the removal of one item and Rasch analysis a further three. Following this, each of the four scales exhibited excellent Rasch model fit. A 14-item summary scale was shown to fit the Rasch model after dropping one of the sub-scales. This provided a total measure of social withdrawal.

Other examples include the Hospital Anxiety and Depression scale (HADS) in MND where Rasch analysis led to minimum item reduction for the two constituent scales [34]; an analysis of the Mini-Mental Health Adjustment to Cancer Scale (mini-MAC) which required more extensive item reduction [35]; and minimum item reduction for the Depression Anxiety and Stress Scale [36,37] Against these, a Rasch analysis of anxiety scales in Parkinson's disease where it was concluded that none of the currently used anxiety scales had satisfactory measurement properties [38].

1.3.5 Rejecting Classical Test and Item Response Theory

Conjoint simultaneous measurement was proposed by Luce and Tukey in the early 1960s as a new type of fundamental measurement, which subsumed the other types. A specific application

was to provide a framework for detecting measurement structures in non-physical attributes (e.g., quality of life). In a slightly modified form this is now applied as RMT. Rasch measurement standards, following those of the physical sciences are designed to create instruments that have interval measurement properties. As noted above RMT is not compatible with either classical test theory (CTT) or item response theory (IRT). They are, as Bond and Cox point out, competing paradigms. RMT takes the perspective that if the instrument is to meet fundamental measurement standards then we should adopt the Rasch *data-to-model* paradigm. If we are not concerned with, or are happy to ignore, questions of fundamental measurement, then we can follow the CTT or IRT *model-to-data* paradigm. The key distinction is that *RMT uses the measurement procedures of the physical sciences as the reference point.* We can aim for the standards in the physical sciences by, as Stevens pointed out in the 1940s, allocating numbers to events *according to certain rules* [37]. It is these rules that comprise RMT. To reiterate: RMT is designed to construct fundamental measures. CTT and IRT focus on the observed data, these data have primacy and the results describe those data. As Bond and Cox emphasize: in general, CTT and IRT are *exploratory* and *descriptive* models; the Rasch model is *confirmatory* and *predictive.*. If RMT is ignored then, by default, instruments utilizing Likert scales or similar frameworks will fail to meet the required axioms of fundamental measurement and remain ordinal scales.

1.3.6 Quality of Life: Needs Fulfilment

Following McKenna et al, the construct that is considered relevant in health technology assessment is one that hypothesizes that the benefits patients derive from a therapy intervention is the extent to which it supports needs fulfilment [39]. Within disease states, QoL, the value placed by individual lives, is dependent on the extent to which their human needs are met. The presence of disease and the impact of interventions drive QoL. Through the presence of disease (and stage of disease) the effectiveness of interventions ameliorates the impact of the disease through supporting needs fulfilment.

Needs fulfilment needs to be assessed directly from patients in the disease state. Attempting to infer indirectly, through the impact of interventions on HRQoL, may have little to do with therapy impact on needs. A clinical focus on symptoms and functional response to interventions, while of interest to clinicians, may not reflect the contribution of those interventions to needs. Non-clinical factors may modify the impact of therapeutic interventions. Needs fulfilment as a latent construct sets it aside from instruments that take a narrower view of HRQoL This is not a question of the number of items. Rather, it is the difference between an instrument that measures symptoms and functional status and one that focuses on the extent to which impairments and disabilities impact needs fulfillment and hence the quality or value of patients' lives. This does not mean that we necessarily reject PROs that capture functions and symptoms. These can certainly be reported as part of a therapy evaluation.

This underscores a fundamental feature of developing what we may call 'patient centric' outcomes (PCO) instruments: the initial generation of items for a unidimensional PCO instruments from qualitative interviews with patients, with the final selection governed by the conceptual model. This sets aside the PCO instrument with its commitment to RMT from the hundreds of generic and disease specific PROs. A PCO instrument brings the patient voice to

center stage. After all, it is the patients who are experts in the impact of a disease, not a clinician or other group who are focused on clinical markers.

Similarly we can focus on caregivers. In many, if not most, rare diseases the caregiver plays a key role for pediatric patients, to include late adolescence as well as the older patient An instrument can be developed for caregivers to assess whether their needs are being met by new therapies or interventions. A recent example of the caregiver approach, as it employs RMT in a needs-based instrument development, is the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue) [40] . Of particular interest here are the different experiences of caregivers identified by whether they are spousal as opposed to non-spousal who, in qualitative interviews reported different experiences of caregiving.

1.4 EXEUNT QALYS AND THRESHOLDS

Is there any role for cost-per-I-QALY claims, or even claims for I-QALYs themselves, in health technology assessment? The answer is no, unless a utility scale to support I-QALY estimates can be shown to have ratio properties. As it stands the I-QALY is an impossible construct and an analytical dead end. It cannot support empirical evaluation as a claim for therapy response. Even if a ratio utility scale could be constructed, its contribution to lifetime reference case models would be irrelevant as the models themselves would fail the standards of normal science.

MODELED AND COMPARATIVE CLAIMS

Health care systems that adopt the Minnesota guidelines should make it quite clear to manufacturers and other making submissions for formulary approval that the only modeled or comparative claims they will accept to support product value will be those that (i) meet required standards of normal science, including fundamental measurement and (ii) provide credible evaluable and replicable unidimensional claims that can be reported to the formulary committee within a meaningful and agreed timeframe.

Specifically, the following claims are unacceptable:

- **Claims that fail the axioms of fundamental measurement which means only interval or cardinal scales and, if possible, ratio scales.**
- **Claims based on composite measures (e.g., multiattribute instruments) that lack dimensional homogeneity**
- **Claims from patient reported outcomes instruments that do not meet Rasch**

measurement standards

- **Claims based upon quality adjusted life years (I-QALYs)**
- **Claims for life years**
- **Claims for equal value of life years gained**
- **Claims that are non-comparative or exclude a comparator product agreed with the formulary committee**

1.4.1 I-QALY Thresholds

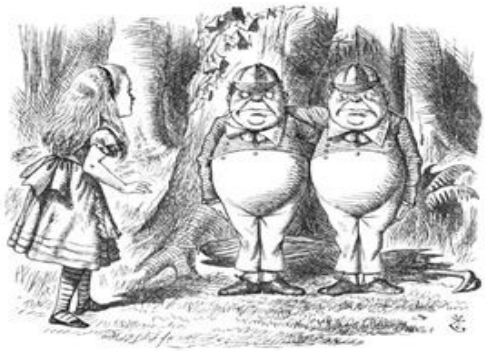
I-QALY thresholds are a key step in ICER's recommendations for pricing, the mythical fair price, and access to pharmaceutical products and devices. Given the mathematical impossibility of an I-QALY completely invalidates any notion of a cost-per-I-QALY threshold for pricing and access even without the failure of the cost-per-I-QALY model to meet the standards of normal science. It might be added, for those unwilling to overthrow years of belief in the ICER reference case, that different models and different utility scales will lead to different imaginary threshold based pricing and access recommendations. As a value tool, it is not only nonsensical but is self-defeating as it admits of a range of possible fair price recommendations for the same product comparisons. This is a ridiculous situation, pointing once again to the irrelevance of cost-per-I-QALY threshold claims to support a value assessment framework. While it may be useful to speculate on disease specific quality of life (QoL) the point to emphasize is that if the utility is ordinal then the I-QALY is a mathematically impossible construct. End of story. Any further discussion of lifetime I-QALYS and thresholds is simply a waste of time. Formulary committees should reject out of hand any modeled value assessments involving thresholds and ICER claims for a fair price in formulary design.

It should be apparent by now that those adopting the Minnesota proposed guidelines will have little interest in claims based on metrics that fail to meet the general axioms of measurement theory. This would include any reference to previous product comparisons based on submissions to agencies that mandate I-QALY modeling. As noted above, this would mean abandoning ordinal utilities as a core metric in technology assessment. This embrace of the health technology assessment meme (see Section 1.8) is all the more remarkable with the explicit disavowal of hypothesis testing by 'experts' in the field. Perhaps the 'latent' view is that *if a theory is sufficiently elegant and explanatory, it need not be tested experimentally* [41]. Which, as the authors continue *is breaking with centuries of philosophical tradition of defining scientific knowledge as empirical*.

It has always seems odd that there are debates over the use of threshold willingness to pay thresholds when their application is built on imaginary cost-per-I-QALY worlds which fail, not only the measurement standards of normal science, but the demarcation test [42]. To claim, as ICER does along with NICE and other single payer health systems, that the appropriate value assessment framework is the application of willingness to pay thresholds to imaginary lifetime

reference case models is, of course, nonsense. Evidence reports and their associated threshold-based recommendations by ICER should be abandoned.

1.4.2 Accepting Claims



Those health systems who are willing to implement the Minnesota proposed guidelines will not be interested in claims for products or devices that lack credibility, which are impossible to evaluate, let alone replicate. If claims are presented, they will only be accepted if they meet the standards of normal science. This should not present concerns for claims from pivotal randomized clinical trials (RCTs) or those from observational studies, to include administrative claims databases and registries. The caveat, of course, is that where clinical trials or observational studies include patient reported outcome instruments, acceptance requires they meet standards for fundamental measurement. Modeled claims would also be accepted, such as extrapolations from RCTs, as long as they could be implemented, evaluated and reported to a formulary committee in a meaningful timeframe.

It is understood that RCTs may focus on clinical endpoints and may also include clinical functions and symptoms as endpoints. If they are secondary endpoints and lack power, the Minnesota guidelines would consider they lack credibility although potential candidates for subsequent claims assessment. The point is that patient-centric claims are 'primary' endpoints; and are admissible as long as they meet the standards of fundamental measurement.

If a manufacturer submits a needs-fulfillment quality of life claim, then the claim must be appropriately powered and meet Rash measurement standards. The manufacturer should be in a position to justify any claim for an instrument that meets RMT standards. Otherwise, the claim should be rejected. If a quality of life instrument that meets RMT standards has not been included or there is no validated quality of life instrument in that disease area, then the manufacturer could be asked to develop a claims assessment protocol that supports needs assessment. This may require the development of a *de novo* RMT standard instrument or the modification of an existing PRO instrument to meet RMT standards. If manufacturers are aware of this requirement, then it can be included as part of the product development process. The incentive is there. There is no certainty that attempting to modify an existing instrument will succeed.

1.4.3 Claims Protocols

While guideline specifications are the responsibility of the formulary committee, a crucial common element must be a request for a protocol for each claim detailing how the claim is to be assessed and reported to a formulary committee in a meaningful time frame. This is different from protocols that are detailed for RCTs, as the protocol would need to

detail the evidence base for claims assessment. As noted above this may be an existing or a *de novo* registry, but one that can support long term evaluations, including ongoing disease area and therapeutic class reviews. This raises the issue of manufacturer specific evidence bases where, to support a product with (possibly) a number of unique claims, the manufacturer has invested in a customized evidence base. Of particular interest here is the situation where a manufacturer has invested in a need-fulfillment quality of life instrument to support claims that cannot be replicated by existing and potential competitors.

The importance of a protocol in claims assessment is that it makes the manufacturer focus on the need to meet the standards of normal science, the need to meet, at least, interval measurement standards in the development of unidimensional single attribute instruments. It is important that manufacturers are given fair warning that the world has changed.

1.4.4 Real World Evidence

Evaluation and ongoing replication of clinical and other claims are an essential part of the initial and continued acceptance of new and hopefully innovative therapies. Replication of protocol driven claims is an interesting question as there has been a long standing concern that the likelihood of replicating a protocol driven RCT claim is low. All too often attempts to replicate pivotal phase 3 claims have failed. Whether this is a general conclusion or one specific to a disease area, the Minnesota guidelines ask manufacturers to propose how claims are to be evaluated in the proposed target population, appropriate to that health system. This applies to clinical claims as well as to claims for quality of life and for other possible claims involving issues such as compliance behavior, adverse events and medical resource utilization.

The recommendations in previous versions of the Minnesota guidelines for evidence platforms are not new. Similar recommendations had been proposed some 15 years earlier in draft guidelines for WellPoint (now Anthem) [43]. The position was clear. A manufacturer submitting claims for formulary listing and price, should commit (or already have committed) to the development of a platform for claims assessment. Manufacturers, if they subscribe to the process of discovery with real world evidence, should consider how the claims made for their product are to be assessed with feedback in a relatively short time frame to formulary committee members. This is not something that should be left until product launch. Unfortunately, manufacturers have been able to put the question an evidence platform to one side, relying instead on imaginary world value assessment frameworks. This is not only an easy way out (and low cost) but the fact that the fabricated value claims cannot be evaluated is an added bonus.

1.4.5 The Role of Registries as an Evidence Base

Attempting to track outcomes and resource utilization from existing data sources is fraught with difficulties, not to mention critical evidence gaps. Linking 'de-identified' patient records is problematic, although claims have been made for algorithms to match such patient records probabilistically (a likely match) for a

'coherent' target patient population. The evidence for this and its interpretation is not convincing. Certainly, administrative claims data can be a valuable source for linking defined target patient populations to drug utilization and comorbidity profiles. Again, they are not sufficiently customized to capture end points such a clinical outcomes, and quality of life, with the ICD-10-CM codes often being insufficient to capture target populations. Most importantly, however, as we engage more with precision medicine tools to identify target populations, there are no data bases that provide this level of detail.

Given the limitations of 'big data' fishing, the one answer for an evidence platform is a registry, particularly when it is possible to 'piggy-back' on an existing registry which captures, or can be expanded to include, the required target patient population. Establishing an evidence platform for monitoring and claims assessment should be an integral part of product development. This would apply in particular to rare disease groups and patients with disabilities and their caregivers where real world evidence is limited.

A registry-based evidence platform that is maintained for a target patient population has a potentially significant role to play in supporting therapy targeting through genetic profiling. The previous version (Version 2.0) of the Minnesota guidelines focused on next generation sequencing and the importance of the choice of profiling assay in evaluating therapy response. The guidelines raised issues in respect of choice of assay, a necessary first step in claims assessment for a target patient group [44]. A test that lacks sensitivity and specificity (e.g., unacceptable false negatives) should be rejected.

Putting aside imaginary evidence in favor of real world evidence gives the opportunity for a more comprehensive assessment involving patient and caregiver centric quality of life assessments together with real world treatment data. Rather than discouraging investors with adverse pricing recommendations driven by imaginary worlds, a registry data base would give a firm basis for guidelines and value contracting between manufacturers and health systems that involves patients and caregivers, yielding provisional pricing by negotiation satisfying all sides.

1.4.6 Disease Area and Therapeutic Class Reviews

As well as supporting claims for a new product at launch, a registry evidence platform can also support ongoing disease area and therapeutic class reviews. These reviews, which should be on a regular (say) three-year cycle, present manufacturers with the opportunity to provide comprehensive comparator data on outcome claims over and above feedback to formulary committees for product entry claims to support pricing and access. Access to micro-data from the registry will facilitate the review process, providing a robust and reliable source of real world evidence, with the added advantage that data collection from registry participants can be modified as our understanding of the disease area and the positioning of new therapies for genetically defined sub-groups becomes more precise

.

It should be remembered that we cannot, in logic, prove a claim. Rather, and this is the point of ongoing disease area and therapeutic class reviews, we can continue to assess, on a regular basis, the evidence base for continuing acceptance of claims, the formulary position and pricing of the product or device. Relying on academically contrived devices

such as probabilistic sensitivity analyses to claim the likelihood of cost-effectiveness at different prices is no substitute for continued assessment.

1.4.7 Evidence to Decision Frameworks

It is not the intention of these proposed guidelines to supplant decision frameworks such as those proposed by the GRADE system, notably the Evidence to Decision (EtD) frameworks developed within the European Union [45]. Rather, the intent is to focus on the quality of the evidence presented to formulary committees and other decision makers and claims for resource utilization within the framework. If these claims lack credibility then it seems somewhat pointless to waste time arguing over their contribution. This is seen in the focus in value assessment frameworks proposed by ISPOR and NICE, with their uptake by ICER, where the claimed values or benefits are entirely imaginary. This does not mean that the GRADE framework for evaluating clinical data should be put to one side; far from it [46].

1.4.8 ICHOM and Fundamental Measurement

Rejecting -QALYs means we are back to square one: within each disease state. We have to revisit the merits of all PROs within disease states to determine whether there are any instruments that meet the required standards. There is no shortage of candidates for review; although we could quickly point out that if the PRO instrument is not supported by RMT it can be rejected. There may be exceptions subject to a review for item reduction. The International Consortium for Health Outcomes Measurement (ICHOM) has produced over the past decade a set of standardized patient-centered outcomes specific to (at the last count) 28 disease states. In depression and anxiety, to give one example, the standardized outcomes for the patient population cover (i) symptom burden; (ii) functioning; (iii) recovery speed and health sustainability and (iv) medication side effects. ICHOM is not in the business of creating 'new' outcomes measures, but takes existing measures which by agreement are included in the set for that disease state. This has the unfortunate consequence of including measures that may fail to meet the axioms of fundamental measurement, notably the non-functional measures that attempt to capture patient-reported outcomes. The further point is the number of measures proposed and the frequency of their administration. In practical terms it is doubtful if a health system would have the resources or the willingness to introduce the ICHOM set for any disease state, let alone for 28 across the board.

ICHOM is not alone. As Porter el al note, progress on outcomes measurement has been slowed dramatically by the 'let a thousand flowers bloom' approach [47]. A patchwork of inconsistent outcomes, measures and definitions. While the authors might press for standardization the progress has been slow. Yet there should be an agreement on a minimum common set of outcomes for disease states which meet the required measurement standards.

The position taken here is that the health system must determine the outcome measures considered appropriate to a target patient population. This may be seen as avoiding the question, but there is no other approach. Attempting an I-QALY-type metric is a waste of time. If value based outcomes are to be the metric, then the elements of that metric must be defined and, if the basis for value based contracting, they must be quantified and tracked through an agreed evidence base. In the case of the Medicaid population, as an example, care must be taken to ensure compatibility with value-based programs implemented by the Centers for Medicare and Medicaid Services (CMS).

At the same time, there must be core measures. Contracting for value must focus on a few measures, specific to disease states and selected attributes. These must be agreed and advertised. It would be unfortunate if the accepted standardized measures were not captured in pivotal clinical trials. One core measure that is proposed for the Minnesota guidelines is needs fulfillment quality of life.. Not, is must be emphasized, a multiattribute symptoms and response rat-bag, but a need-fulfillment measure that is specific to disease states, and meets RMT measurement standards for unidimensionality and interval scoring. It is of interest to note that this core measure is absent from the ICHOM outcomes lists even though RMT measures haven been available in a number of the disease states where standardized sets have been proposed. The notion of needs fulfillment does not appear to be on the ICHOM agenda.

In the case of anxiety and depression, for example, the Patient Health Questionnaire PHQ-9 that is commonly used to screen for depression and for monitoring symptoms while having good classical psychometric properties fails to meet RMT standards [48]. It lacks a unidimensional (i.e., single attribute) structure as well as lacking interval scale properties. It cannot be used, if we subscribe to the axioms of fundamental measurement, for measuring change. The failure of the PHQ-9 stands in contrast to the Quality of Life in Depression Scale (QLDS), a needs-based instrument that meets RMT standards and, with an interval scale, has been used widely in clinical trials and evaluating response to therapy[49]. Not surprisingly, ICHOM does not mention the QLDS.

1.5 MEETING THE STANDARDS FOR THE MINNESOTA VALUE ASSESSMENT PARADIGM

For those unfamiliar with Kuhn's work, a paradigm shift is a fundamental change in the concepts and experimental practices of a scientific discipline [50]. Characterized as a scientific revolution, it occurs with the overthrow of activities within normal science, where these activities are rendered incompatible with new phenomena. Unlike technology assessment where there is no intention, in making incremental cost per I-QALY claims, or of any need to meet the standards of normal science, the paradigm shift that brings

technology assessment 'in from the cold' is a return to normal science. Exemplified by the formulation of product claims that are credible evaluable and replicable; not imaginary information constructs that fail the standards of fundamental measurement. It is noteworthy that instead of the paradigm shift occurring with a methodology that meets the standards of normal science, this paradigm shift is a rejection of a pseudoscientific paradigm (e.g., intelligent design) to one that recognizes the standards of normal science (e.g., natural selection). Once this new paradigm is accepted, then the term 'cost-effectiveness' loses any relevance in formulary decisions, being replaced by a focus on a formulary committee defined set of core unidimensional attributes defined by RMT..

1.5.1 An Analytical Dead End

The ICER reference case and the support for lifetime cost-per-I-QALY models by ISPOR and other agencies is, as noted, an analytical dead end. There is not only the failure to meet the evidentiary standards of normal science, but the willful neglect of the axioms of fundamental measurement. Modelled claims built on assumptions to create approximate yet impossible information must be rejected. Health technology assessment seems alone as a branch of the social sciences that focuses on imaginary as opposed to real world evidence. While this may be defended on the grounds that at product launch evidence for therapy response is limited, it is no justification for creating evidence. Certainly, evidence may be limited at product launch, but that is no excuse to put aside the notion that claims are always provisional in favor of an imaginary construct that effectively shuts the door to future claims evaluations. A black box I-QALY lifetime simulation is a rat-bag of assumptions of claims on the future. Rather than relying on this, the focus should be on single claims representing defined attributes.

1.5.2 A New Paradigm

The new paradigm must be disease specific; it must focus on real world evidence for target patient populations. It must recognize the necessity of meeting the axioms of fundamental measurement. Certainly, it can accommodate ordinal measures while recognizing that if the objective is to assess response to therapy, then the instruments must have latent if not ratio measurement properties. This requirement applies equally to all PRO claims. At the same time there is the need to take the patient voice and needs seriously. Instruments should focus on specific attributes and not attempt to cram multi-attributes into a single scale. Instruments must be shown to have unidimensional properties, dimensional homogeneity and meet RMT standards. Finally, there must be an ongoing commitment to reassessing and if necessary rejecting previous provisional claims. This points to the requirement for ongoing disease area and therapeutic class reviews.

1.5.3 A Dynamic Claims Platform

The paradigm proposed here takes us from an essentially static claims one-off model framework to one that is dynamic. That is, rather than relying solely on claims driven by pivotal phase 3 RCTs, the ability to focus on real world data and real world evidence gives manufacturers the opportunity to consider the opportunities for supporting effectiveness claims that are targeted to specific patient populations. One example would be needs-fulfillment QoL. This is unlikely to be

captured in pivotal trials and even if relegated to a secondary outcome it would lack power to support labelled claims. A protocol driven process of claims assessment, covering varying timelines and reporting requirements is a critical requirement; formulary committees should determine for a target population and disease specific basis, the appropriateness of single attribute claims.

1.5.4 Value-based Contracting

These proposed Minnesota guidelines are in complete alignment with value contracting, to include risk-sharing agreements and outcomes based contracts. This is seen in the rejection of attempts through imaginary modeling to determine a fair price for a pharmaceutical product or device. The bankruptcy of the ICER value assessment framework is a salutary lesson. There is no magic framework to set fair prices for products indicated for target populations in disease area. Pricing is seen as provisional subject to negotiation between the parties with agreed measurable claims that are to be evaluated as the product of device is taken up by the target population. The key is the evidence base, where claims supported by evaluation protocols, including as part of the protocol possible limitations on access and prescribing within the indication for the product are implemented, assessed and reported to the formulary or other decision makers in a meaningful time frame. The evidence base provides for ongoing disease area and therapeutic class reviews, preferably tracking patients over the course of the disease until replaced by a more effective therapy intervention. Price is always subject to re-negotiation as the response to therapy is better understood. Performance targets can be aligned to price increases (or decreases) with possible ceilings for annual or semi-annual manufacturer proposed price increases. This is possible through the commitment to an evidence base infrastructure that supports real world evidence for claims evaluation and therapy response.

1.5.5 Falsification and Feedback

Proposing quantitative claims goes hand in glove with protocols describing how the claim is to be assessed in the target treating population and the findings reported back to the formulary committee. This is the foundation of real world evidence: the accumulation of knowledge regarding treatment effects which meet the demarcation standard. If we accept that science is about problem solving, the discovery of new facts, then claims made must be capable of falsification. A position resolutely opposed by ISPOR and ICER. Claims which have a low probability of 'failing' should be avoided. Claims (or hypotheses) must be stated in such a way that they can be exposed unambiguously to refutation. This satisfies the criterion for demarcation between science and non-science. If there is no observable difference between a claim being true and being false then it conveys no scientific information. Unless a claim can be falsified it is not a scientific statement; hence the objection to modeled imaginary worlds.

1.5.6 Value Assessment Standards

The proposed vale assessment standards are detailed in Figure 2. The intent is to sweep away the dross that currently dominates value assessment under the ISPOR/ICER technology assessment paradigm. We need a clean break and not an attempt to continue with the I-QALY under a different guise.

Figure 2: Value Assessment Standards for the Minnesota Paradigm

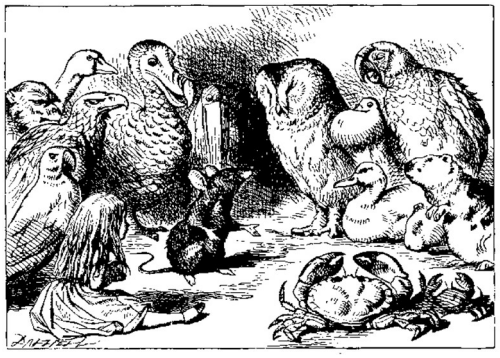
EMBRACING A NEW PARADIGM FOR VALUE ASSESSMENT

The Minnesota Guidelines represent a new paradigm in value assessment. They represent, not an extension of the imaginary world paradigm, the creation of evidence from incremental cost-per-I-QALY lifetime or reference case modeled claims, but a complete rejection of that paradigm. We have to start from an entirely new foundation: the acceptance of the standards of normal science and, in respect of value claims for response to therapy, instruments that recognize and are developed to meet the requirements of the axioms of fundamental measurement. This paradigm shift is 30 years overdue.

In summary:

- **All value claims should meet standards of normal science for credibility, evaluation and replication**
- **All value claims should meet standards set by axioms of fundamental measurement**
- **All value claims should be unidimensional and be specific to a response attribute**
- **All value claims should meet interval or ratio measurement properties**
- **All value claims should be disease specific, reflecting the interests of patients, caregivers and clinicians**
- **All value claims should be supported by a protocol detailing how the claim is to be evaluated and reported**
- **Proposed value claims that fail to meet any of these criteria should be rejected**

2 THE TARGET PATIENT POPULATION



A coherent framework supporting real world evidence is critical to formulary submissions and decisions. Central to this is the notion of an 'evidence registry'; a registry capturing a representative sample of the target patient population which can support initial and ongoing assessments of claims. The evidence registry would not only support replication of claims from pivotal clinical trials and attempts to draw indirect comparisons between comparator therapies, but a platform for exploring a range of value assessments required by the formulary committee; values that the pivotal trials have either not captured as powered primary endpoints or which have not been part of the trial protocol. An obvious candidate in the latter group is a patient centric needs fulfillment Rasch standard, quality of life instrument.

Manufacturers should demonstrate, apart from the claims generated by pivotal phase 2 and phase 3 trials:

- their awareness of the characteristics of the target population
- the relevance of their protocol population to the target population
- how the target population will be identified in treatment practice
- their proposed evidence platform for tracking and reporting (e.g., registry design)
- how the selection of patients, adherence and outcomes are to be assessed
- the unmet clinical and social needs of the target population (including caregivers if appropriate)
- the extent to which claims for meeting unmet clinical and social need will be resolved with the proposed intervention
- the clinical and social benefits of their product over and above those of comparators

2.1 THE VALUE ASSESSMENT EVIDENCE PLATFORM

The insistence by a formulary committee on an evidence platform for the assessment of product claims is a key element in the Minnesota guidelines. If we reject the fabrication of imaginary worlds with their claims for 'valid' approximate information to drive formulary decisions, then we must have a platform which provides, for the target patient group, a vehicle for both initial and continuing claims assessment. This is the responsibility of the manufacturer. The manufacturer should provide a detailed description of the evidence platform that is either in place or proposed for claims assessment. It should be made clear by the formulary committee that the evidence platform, ideally an evidence registry, should support not only claims made for product entry with feedback to the formulary committee but also prospective disease area and therapeutic class reviews for products indicated for the target patient population.

The hundreds of rare and chronic diseases present significant challenges for value assessment; evidence is limited and manufacturers, perhaps understandably seek a return on investment with

prices that many see as exorbitant. While the response by ICER has been to create imaginary fair price value assessment models and apply significantly higher thresholds for cost-per-QALY claims, the exercise is meaningless. If we want to discover new facts, including the effectiveness of innovative therapies in rare disease such as sickle cell disease and Duchene Muscular Dystrophy in treatment practice, then an evidence registry is essential. With a registry in place, provisional pricing and access protocols can be negotiated while awaiting feedback to formulary committees.

Reasonably, formulary committees will be aware of unsuccessful efforts to replicate phase 3 clinical claims; let alone claims when protocol restrictions are lifted and new claims protocols proposed for 'real world evidence'. Manufacturers may have these data in place given the delays in formulary assessments across different health systems following product launch. If so, then the manufacturer is in a position to report claims assessments as part of the formulary submission.

The importance of establishing, at early stages in product development through Phase 2 and 3 RCTs, the evidence base for replicating claims and tracking patient response is seen in the failure of retrospective studies to replicate clinical trial evidence. A recent assessment of 220 trials reported in the top 7 high impact journals in calendar 2017 found that 15% of the US-based clinical trials could be feasibly replicated through the retrospective analysis of administrative claims or electronic health record data [51]. While there is the suggestion that real world evidence could be the basis for retrospectively complementing randomized trials through observational studies, we are a long way from achieving this objective. This raises the obvious conclusion: if existing data sources are unlikely to be modified in data collection and classification (e.g., administrative claims data) to allow for replication of trial results, then we must look to alternative data sources. Hence the emphasis in the Minnesota guidelines on protocols proposed by manufacturers to support and complement with additional disease specific value measures the replication of trial based claims in target treating populations.

Irrespective of the hyperbole that attaches to the presumptive role of real world data to create real world evidence, seen in the notion of 'big data', existing real world data sets fall far short of the mark in supporting any move from imaginary world data to real world data in value assessment. Certainly, the FDA is focused on constructing from electronic health records the creation of a medical data enterprise system. Unfortunately, that is only useful if the measures of patient outcomes embedded in the records meet acceptable measurement standards. The case presented here is that formulary committees must be proactive in requiring manufacturers to demonstrate and support evidence registries, both to assess existing trial claims but to complement these with patient centric quality of life measures of response

2.1.1 Identifying the Target Patient Group

Given the focus on credible and evaluable claims, the first step in a submission should be to define the target patient group. If a coherent diagnosis is the basis for identification then this needs to be spelled out. Diagnoses that are not precise may fail to take account

of unwanted variation in diagnoses which can adversely impact value claims. The diagnosis should be capable of timely assessment in treatment practice, representing a consensus view (e.g., in guideline implementation) which can support not only evaluation of response claims but replication of assessment across designated treating populations. If the target group is a sub-set of a more broadly defined group (e.g., stage of disease) then this needs to be specified and the appropriate diagnostic procedures applied. This applies also to the application of precision medicine.

If a potential target population is to be defined by a genetic assay then the submission should detail the recommended assay, giving reasons for this choice and its availability in treatment practice, including assessments of diagnostic accuracy. At the simplest level, if a binary classification is appropriate, this assessment should include evidence for the sensitivity and specificity of a test, false negatives and false positives, in the anticipated treating environment. It is not sufficient to just reproduce the protocol defined target population from pivotal RCTs. The concern is with a test that lacks external validity. Defining a target treating population too narrowly, may give a false impression, not only of the clinical impact of a new therapy (i.e., effectiveness), but also a false reading of the extent to which, under the proposed gold standard of needs assessment, the intervention meets the needs of the target population. Application of a patient centric needs-fulfilment instrument must be relevant to the needs and concerns of the target population. A too broadly defined target population sample may not represent the needs of a sub-population defined by stage of disease or selected genetic markers.

Where guidelines have been developed by professional associations, these should be detailed together with the proposed place in therapy of the product and relevant comparators within the guideline framework.

2.1.2 Epidemiological and Social Profile

As part of the development of an 'evidence registry' the manufacturer should be in a position to report, as part of the submission to the health system, the characteristics of the potential target population. Required data elements to provide a social and epidemiological profile (i) at a national (US) level; and (ii) for the health system receiving the submission are:

- **Data sources**: detail the data sources, codes and possible algorithms that are considered necessary to identify the target population
- **Population Estimates**: provide estimated target population counts for the last 5 years detailing the data sources and potential sources of error
- **Incidence**: given prevalence estimates provide annual incidence counts of patients diagnosed with the target disease
- **Basic Demographics**: provide a profile identifying the target population by age (5 years groups), gender, ethnicity and race (US census definitions),
- **Socioeconomic Status**: provide a profile identifying the target population by work status, (including unemployed/retired) and family income (US census definitions)
- **Insurance Status**: provide a profile of the insurance or health system coverage for the target population (commercial/private, Medicaid, Medicare, no insurance)

- **Drug Utilization**: the distribution for each of the past 3 calendar years of drugs utilized for the proposed indication in the target population detailing compliance patterns, switching to comparators and average/median time to discontinuation
- **Polypharmacy**: the distribution of all prescription drugs identified for the target population in the past three years
- **Clinical Status**: if there are defined disease stages provide a profile of the target population by disease stage (including the elements detailed above)
- **Genomic profile**: identify subpopulations within the target population that may respond differently to the target therapy or excluded from treatment
- **Comorbidity Status**: provide a profile of the five (5) most prevalent co-morbidities in the target population
- **Caregivers**: provide a profile (if appropriate) of the prevalence of caregivers (e.g., for pediatric patients) in the target population
- **Social Factors**: extent to which environmental, income and lifestyle factors impact drug access and utilization

2.1.3 Patient Reported Outcomes

One of the obstacles to evaluating response to therapy has been the reliance on PRO measures that fail to meet the axioms of fundamental measurement. It is important that a systematic review is undertaken of the PROs that have been utilized to evaluate response to therapy in the target patient population to assess, not just the extent to which in their development they have met the required standards in classical test theory but, more importantly, the extent to which in this development they have met the required axioms of fundamental measurement. This is not just an *ex post facto* assessment of item selection and claims for unidimensionality, but a commitment from day one that the PRO measure meets Rasch measurement standards. The focus is on response to therapy not imaginary simulations. To achieve this PRO instruments should have demonstrated interval level properties for the attribute they are proposing to measure; otherwise they should be rejected..

The importance of a detailed critique of the dominant PRO measures in the target disease and patient population refers not only to those used in RCTs but those commonly used in clinical practice. It is probably fair to say that the majority of these PROs almost certainly fail to meet RMT standards. It is no defense to argue that the COSMIN standards are met; a more detailed measurement standards assessment is required. After all, we know that the overwhelming majority of PROs belong to the 'add em' up' school of scoring from Likert or similar scales which inevitably lead to ordinal scales. These, unfortunately, as noted above, fail to meet standards for assessing response to therapy. They will not be able to support credible claims for response to therapy. As the basis for comparator claims they will be worthless. It is important that the formulary committee is aware of these various instrument and their shortcomings.

Needless to say the formulary committee will not be interested in claims submitted that rely on such instruments.

At the same time a systematic review may prompt the formulary committee to request response claims from newly developed PROs that meet required measurement standards. One example would be, if quality of life is a proposed claim, that an instrument is developed that meets RMT needs fulfillment criteria.

2.1.4 Unmet Medical Need

The extent to which a new therapy contributes to meeting an 'unmet' medical need(s) in a target population is a critical aspect of formulary assessment. As part of the formulary submission package, manufacturers should present a clear assessment of unmet needs in the target population group. This should be based on a systematic literature review supported by a statement as to what the company perceives to be the unmet need that their product or device is intended to meet. Clearly, other products may be promoted as meeting unmet medical needs in the target population. The submission should include a statement that details the extent to which specific unmet needs are addressed by these products.

2.2 PROTOCOL AND SUBMISSION ASSESSMENT

2.2.1 Protocol Reconciliation

For a formulary committee to judge the commitment of a manufacturer to a product, it is important to have a detailed profile of completed, ongoing, and proposed RCTs and observational studies. The latter would include links to patient advocacy groups and possible joint projects underway or anticipated.

External validity of trial based claims is a perennial concern to formulary committees. A submission should detail for each RCT the protocol exclusion and inclusion criteria, to include those trials for comparator products. Of particular interest are proposals for (i) active comparator trials and (ii) trials where it is proposed to relax the exclusion criteria. This review must cover trials that have been completed, ongoing and proposed. Specifically:

- A summary of all relevant completed, ongoing and proposed trials for the product and its comparators detailing:
  - NDC Code (www.clinicaltrials.gov)
  - Trial designation
  - Trial objectives
  - Primary and secondary outcomes
  - RMT assessment of outcome claims
  - Inclusion and exclusion criteria
  - Status (completed, ongoing, proposed)
- Match the inclusion criteria for subject selection to the
  - demographics (age, gender, ethnicity, race)

  - socioeconomic status
  - stage of disease
  - comorbidity profile

- Estimate the proportion of the target population characteristics at (i) the national and (ii) the health system level that the individual trial protocols capture

2.2.2 Protocol Replication

It is possible that a manufacturer may claim that it is feasible (at least in the US) to replicate the trial protocol from existing data sources: registries, administrative claim data and electronic health records. While this does not provide an excuse to put issues of access to an evidence registry to one side given the trial protocol population it is likely to be a subset of the target population, it may provide a useful opportunity to assess the replication of trial claims. Specifically:

- Assess the likelihood that each of the individual trial protocols could be feasibly replicated from existing data sources (e.g., electronic health records, administrative claims data, registries)
- Describe for each feasible protocol the data source(s) and accessibility

2.2.3 Pipeline Product and Competitor Therapies

Manufacturers should detail the prospective competitors for their product in the target patient population within a time horizon of expected approvals within the next five years. This should include anticipated product enhancements and other products within the manufacturer's own pipeline.

To avoid future disappointments, the formulary committee might advise manufacturers that given their pipeline, either for extended indications for established products or for new products in a therapy area, that RCT protocols should be reviewed (alongside possible observational studies) and that instruments should meet RMT measurement standards. Future claims rejection should be anticipated.

2.2.4 Submissions to National Evaluation Agencies

Manufacturers should provide a list of all submissions made for their product to foreign national health authorities such as NICE. Links should be provided to the relevant website if, as in the case of NICE, material is posted to the public domain. Manufacturers should provide a summary of the case made for their product (including any reference case imaginary worlds), critiques of their submissions and outcomes for pricing and formulary listing. If publications are associated with these assessments a list should be provided.

2.2.5 ICER and other US Submissions

Manufacturers should provide a list of all submissions for pricing and access made to agencies in the US. This would include ICER evidence reports and other proposals to agencies such as the VA. The results of each submission should be detailed together with any publications associated with the submission and relevant website links.

3 CLINICAL EVIDENCE STANDARDS

Two questions are paramount in a formulary assessment of a new product: first, the clinical status of claims vis à vis comparators (not placebo) and (ii) real world evidence for the performance of the product in treatment practice following FDA marketing approval. Clinical claims should be specific to attributes and meet the standards detailed in Figure 2 for the Minnesota paradigm. A range of clinical response attribute claims can be presented, supported by protocol for each claims proposing how it will be (or may have been) evaluated. A key element is the evidence for the replication of claims. Formulary committees are understandably reluctant to rely on a few randomized clinical trials where the protocol may effectively exclude any claims for external validity. Replication of claims is crucial. A failure to address the question of external validity in target patient populations is unlikely to impress.

There is any number of guides for presenting the clinical case for a new or existing product, where the latter may be part of a disease area or therapeutic class review. These include a summary of the pivotal clinical trials, together with spreadsheets dealing with the comparator products in the disease area and the results of meta- or indirect assessments of treatment effect. The intent here is to focus on clinical evidence that is directly relevant to the formulary decision. Certainly, detailed spreadsheets can be prepared. The likelihood of anyone reviewing them is slight. Importantly, the formulary committee will make its own decision. It is not interested in groups, such as ICER, who may have determined what they seen as the 'value' of competing therapies. The committee is perfectly capable of coming to its own conclusions. This does not exclude network meta-analyses undertaken by reputable groups (e.g., Cochrane collaboration).

It is unusual for a product to be reviewed by a formulary committee immediately following product marketing approval. By the time committees get round to evaluating a product there will be (or should be) real world evidence for product performance in the target population in treatment practice. Manufacturers should be on notice, given those professing the importance of real world evidence, to prepare for formulary committees requesting this information. Supplementary real world evidence that will also be important is the potential for compliance and adherence (i.e., what patterns are there) and experience with the product where access barriers are in place.

The obvious point, if we consider Popper's contribution to the philosophy of science, is that any claim is provisional. This holds for clinical response claims as it does for any other hypothesis. This means that formulary committees require, indeed it is in their duty of care for the interest of the patient, real time feedback for all claims whether these are in clinical terms, quality of life or elements such as product uptake and discontinuation, together with units of resource utilization (but not aggregate direct medical costs).

3.1 Systematic Reviews and Meta Analyses

As a first step submissions must provide and report on a reporting list that conforms to the latest PRISMA statement. The first PRISMA item list was released in 2009 [52]. This has now been substantially revised and updated [53]. While each item should be addressed in the PRISMA scheme, particular attention should be given to (i) the synthesis of results for the individual meta-analyses and (ii) the strength of evidence for each primary outcome. Where appropriate submissions should meet the PRISMA-P standards for systematic reviews in adults and the pediatric extensions: PRISMA-P-C (Protocol for Children) and PRISMA-C (Children) [27 28].

PRISMA, unfortunately, takes no account of measurement standards in its results assessment. This implies that many if not the majority of PRISMA claims are redundant. When a manufacturer reports on systematic reviews and meta analyses it is important to qualify the assessment by evaluating the extent to which the respective reviews and analyses recognize the importance of meeting the axioms of fundamental measurement. Meta-analyses of ordinal scores are clearly of little interest as the nature of the scale precludes such assessments.

3.2 Reporting Randomized Trials

Reporting of results from randomized clinical trials of test performance should conform to the Consolidated Standards of Reporting Trials (CONSORT) [54]. This is a standard format for reporting on trial organization, analysis and interpretation. The CONSORT Statement comprises a 25-item check list and flow diagram to record the progress of patients through the trial. As well as the CONSORT framework, care needs to be taken as to whether or not study results are generalizable to the target population.

In addition it is important, given the focus in these guidelines on patient reported outcomes to note the CONSORT PRO extension [55]. Five CONSORT-PRO checklist items are recommended for randomized clinical trials where the PRO is a primary or secondary endpoint:

- Identification of the PRO measure as a primary or secondary endpoint
- A description of the PRO measurement hypothesis and relevant domains if multidimensional
- Evidence for validity and reliability (classical test theory)
- Procedures for missing data
- Discussion of study limitations and generalizability to other populations and clinical practice.

Once again, those developing the CONSRT standards failed to address the issue of the measurement properties of the instrument. It is clear that this concern was not apparent to the CONSORT PRO checklist authors. This is a major oversight. The first question should be:

- Does the PRO (or PROs) meet the required axioms of fundamental measurement (i.e., has interval or even ratio properties for a single attribute been demonstrated)
- If NO, then the RCT should be rejected or, if the PROM is a secondary endpoint and the primary endpoint meets required standards, then that should be accepted

It is recognized that this may put the manufacturer in the awkward situation with phase 2 or 3 pivotal RCT results failing to meet the required measurement standard. Claims that may have been accepted for review and marketing approval by the FDA may fail simply because the FDA does not appreciate the importance of fundamental measurement. This raises an interesting dilemma: comparator products may have been accepted on the basis of claims created by instruments which would also have failed to meet the required measurement standards. There is, of course, a solution: the manufacturer could commit to a comparator RCT with instruments that meet the required standard. This has the added advantage that it avoids the need for indirect statistical assessments for comparators with only placebo controlled trials. As an interim solution the formulary committee may be prepared to accept a claim on an interim basis subject to an RCT protocol that meets its required real world evidentiary standards, together with a protocol for implementation and reporting timetable.

3.3 Evidence Hierarchy

Claims for the efficacy or effectiveness of claims in clinical practice must be founded on high quality and bias-free evidence. Where a submission has undertaken a systematic review or relies upon individual studies to support credible, evaluable and replicable claims that meet the required standards for fundamental measurement, the evidence presented should be assessed against the standards established within the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working groups. The GRADE framework has superseded earlier proposals for the ranking of evidence (which typically ranks from randomized trials through to observational studies and anecdotal, key opinion leader evidence) to a more flexible evidence hierarchy addressing the quality of evidence for individual outcomes. Specifically: bias, inconsistency, indirectness, imprecision and publication bias [56]. All that is missing is an appreciation of fundamental measurement..

The GRADE framework is intended to apply to meta-analyses from systematic reviews but can be applied to individual studies or non-quantitative syntheses. The essence of the GRADE approach is that, within each hierarchy level, it allows the downgrading or upgrading of evidence. Downgrading, for example in the case of randomized clinical trials, occurs if there is a

risk of bias, inconsistency, indirectness, imprecision and publication bias. Upgrading, for example in the case of non-randomized studies can occur if there is a large magnitude of effect, evidence of a dose response effect and if all plausible confounding factors have been taken into account. The application of the GRADE framework is a 4-level quality rating hierarchy. This is detailed in the Cochrane Collaboration handbook [57].

1. *High Quality Rating*: Randomized trials; or double-upgraded observational studies
2. *Moderate Quality Rating*: Randomized trials; or upgraded observational studies
3. *Low quality rating*: Double-downgraded randomized trials; or observational studies
4. *Very low quality rating*: Triple-downgraded randomized trials; or downgraded observational studies; or case series/case reports.

The GRADE evidence approach has figured largely in the Agency for Healthcare Research and Quality *Methods Guide for Comparative Effectiveness Research* to support the Evidence-base Practice Center (EPC) Program [58] . The EPC framework grades the strength of evidence from RCTs as well as observational studies in a systematic review through assessing specific domains: study limitations, directness, consistency, precision and reporting bias. Potential additional domains are: dose-response association, plausible confounding for observed effect and strength of association. Scoring these domains yields four strength of evidence grade:

1. *High*: The reviewers are very confident that the estimate of effect lies close to the true effect
2. *Moderate*: The reviewers are moderately confident that the estimate of effect lies close to the true effect
3. *Low*: The reviewers have limited confidence that the estimate of effect lies close to the true effect
4. *Insufficient*: The reviewers have no evidence, they are unable to estimate an effect, or we have no confidence in the estimate of effect for this outcome

Once again, application of the GRADE framework must be qualified by the application of the axioms of fundamental measurement. PRO measures that fail to meet these standards for unidimensionality or dimensional homogeneity should be rejected. While effect size may be claimed, the absence of response defined by RMT based interval scores invalidates any claims.

3.4 Provisional Response Claims

The formulary committee is not interested in clinical or associated claims for therapy response that rely on PRO measures that fail to meet the axioms of fundamental measurement. This includes the majority of PRO measures, outside of utility instruments, which are only capable of creating ordinal scales. This is a significant limitation which points to the need to reconsider the relevance of instruments with a commitment to creating, at least, unidimensional PRO measures with interval measurement properties to capture response to therapy. It is surprising that after 30 years and thousands of RCTs the question of meeting the axioms of fundamental measurement has been overlooked. Even so, there should be scope for manufacturers to build value claim

proposals from RCTs where the response from pivotal trials are taken as benchmarks, provisionally accepted prior to a protocol driven claims evaluation that meets fundamental measurement standards.

.

4 QUALITY OF LIFE: PATIENT AND CAREGIVER NEEDS

The position taken in the Minnesota guidelines is that claims for quality of life must be disease specific and patient centric. This means an instrument that is unidimensional and has interval calibration, to assess response to therapy, and which is focused on needs fulfilment within a Rasch measurement framework. Claims for quality of life that do not meet these requirements should be rejected. Dimensionally heterogeneous instruments such as multiattribute utility instruments are not acceptable.

4.1 Prior Quality of Life Claims

It is of no interest to present detailed assessments of previous HRQoL or quality of life QoL claims for the target patient population when the instruments involved fail to meet RMT or the axioms of fundamental measurement. Failure to meet the requirement of dimensional homogeneity with interval response is the key criteria for rejection. As the instrument will usually be one of the multiattribute utility scales (the EQ-5D-3L is by far the most popular) it is sufficient to point this out. A detailed evaluation of previous submissions or research that includes imaginary modelled cost-per-QALY claims is a waste of time.

An example of spurious claims for creating QALYs is provided by a recent response by ICER to questions regarding a 'proof' of the required ratio properties of utility scales to create QALYs [10]. ICER's strange response was:

> *We (and most health economists)* ***have the understanding*** *(emphasis added) that the EQ-5D (and other multiattribute instruments) do have ratio properties. The EQ-5D value sets are based on time trade-off assessments (which are interval level) with preference weights assigned to different attributes. We fail to see why this should be considered as an ordinal (ranked) scale. ICER believes that the dead state represents a natural zero point on a scale of health related quality of life. Negative utility values on the EQ-5D scale represent states considered worse than dead.*

The rebuttal made clear that these were nonsensical assertions:

This is a truly amazing response; and one that is demonstrably false. For ICER everything in constructed simulations is by assumption. ICER and others may assume anything; in this case to assume the TTO tariffs of the EQ-5D algorithms have ratio properties is complete nonsense. Unfortunately, ICER does not provide a proof of this bizarre assertion. Similarly the TTO technique does not yield interval let alone ratio properties [59]. The TTO tariffs created by the EQ-5D scoring formulas from the econometric modelling have only ordinal properties.

If ICER continues to insist that they can defy the axioms of fundamental measurement they are entitled to do so; hoping presumably that they will be believed. If, as discussed in the text, ICER insists on this ratio property then the EQ-5D-3L with a range from -0.59 to 1.0 must have (somewhere) a true zero. However, ICER, in their reformulation of measurement theory must prove that in the absence of a true zero multiplication (to create QALYs) is possible. Can we see this proof? This proof must support all arithmetic operations (but not be assumption). However, we do have the intriguing but weird possibility of negative QALYs! I suppose there is an upside.

Consider the phrase 'have the understanding'. Can health economists demonstrate that the EQ-5D-3L, even with negative values, has ratio properties which requires a true zero? Can ICER show that time-trade off has unidimensionality and interval properties? The answer is that it does not: to claim that the EQ-5D-3L scale has ratio properties because the TTO has interval properties is just nonsense. ICER might demonstrate how an interval scale can be (and apparently has been) transformed to a ratio scale. We might have the understanding that the moon is made of green cheese; this does mean it has. At least this claim can be empirically assessed unlike ICER claims.

Indeed, ICER admits that there can be states worse than dead (i.e., negative utilities) which means that the scale does not have ratio properties. Perhaps ICER should make its mind up

Where a QoL instrument has been used that meets the required measurement standards for the target population, this should be detailed with a summary of the steps in instrument development and the respective claims for therapy response.

More recently, in response to a request for a 'proof' that the EQ-5D-3L or similar multiattribute systems has ratio measurement properties led ICER to reference a paper that attempted to make the case [60]. Much as they try, the authors of the study cannot demonstrate they have found a 'truw zero' for an ordinal scale with negative utilities. Nor can they demonstrate that the utility scales have interval properties. Perhaps, even with negative utilities the model builder could assume that the raw scores generated by the EQ-5D-3L algorithm were ratio scales 'in disguise'. This is clearly nonsense. Indeed, asking ICER to provide a proof is a rhetorical question as we know that it is impossible. ICER together with other analysis may have a belief in the ratio properties of ordinal raw scales, but this demonstration of faith does not trump the axioms of

fundamental measurement. It is a an *ex post facto* attempt to justify 30 years of applying I-QALYs.

If further evidence is required to dismiss ICER's belief that the EQ-5D-3L and other utility scales are ratio measures in disguise, it is useful to point out how the EQ-5D-3L raw score for health states is constructed. The EQ-5D-3L algorithm involves starting from unity (1 = perfect health):

- subtract a constant term (for any dysfunctional state) [- 0.081]
- subtract five dimension scores for mobility level, self-care level, usual activities level, pain or discomfort level, anxiety or depression level (3 levels: no problems, some problems, extreme problems) [no problems = 0; some problems range 0.069 to 0.123; extreme problems range 0.049 to 0.386];
- subtract N3 level (where level 3 occurs within at least one dimension) [– 0.123}

As an example of the raw score calculation consider a situation where the respondent reports extreme problems on each of the five symptoms (health state 33333). The dimension/response scores are generated by the time trade off (TTO) technique and the scores are defined as community TTO tariffs. The equation is:

$$1 – (0.081 + 0.314 + 0.214 + 0.094 + 0.386 + 0.236 + 0.269) = 1 – 1.594 = -0.594$$

For the more dysfunctional states the algorithm will yield negative scores within the possible range of 1 to -0.594 The negative values (below death) are simply health states worse than death. If the axioms of fundamental measurement are applied then it is quite clear that the utility value is an ordinal raw score. No consideration was given in instrument development to the requirement for interval let alone required ratio properties.

4.2 Preparing for Quality of Life Claims

As noted, an undoubted advantage of the ISPOR (and ICER) recommendations for imaginary modeled cost-per-QALY claims is that these can be safely ignored until marketing approval has been received. It is then relatively easy (and inexpensive) to cobble together a model that is driven entirely by a few data points from clinical trials but with an abundance of assumptions. It is an easy 'one-off hurdle' that any number of private consultants offer as their stock in trade; after all, who pays any attention to it later? Such an approach is not conducive to

the belief, if any, that a formulary committee might have in a quality of life claim.

While it might be wishful thinking, the position take in the Minnesota guidelines is that if quality of life is considered, from a patient (including caregiver) centric and needs perspective, as a critical issue in therapy claims for specific rare and chronic diseases, then a manufacturer should address this in the context of product development. If product claims are focused on quality of life impact then these need to be articulated at an early stage in product development with an underwriting of the appropriate needs fulfillment instrument (or instruments) for phase 3 trial protocols.

A manufacturer in making a submission under the Minnesota guidelines framework should demonstrate that a systematic QoL review has been undertaken of potential instruments that meet Rasch measurement standards, capturing patient centric and needs fulfillment criteria. There has been a significant literature over the past 20 years on Rasch instruments, including efforts to modify existing ordinal instruments to meet Rasch interval standards. At the same time a number of Rasch needs-fulfillment instruments have been developed across a range of disease areas in multiple language versions and utilized in clinical trials. These include:

- Pulmonary hypertension
- Alzheimer caregivers
- Atopic dermatitis
- Psoriasis
- Growth hormone deficiency
- Crohn's disease
- Recurrent genital herpes
- Migraine
- Multiple sclerosis
- Depression
- Asthma
- COPD
- Arthritis (osteoarthritis, psoriatic, rheumatoid)
- Incontinence

Instruments cover pediatric patients, caregivers and adults. Samples can be viewed at (www.galen-research.com)

4.3 Reporting Quality of Life Claims

Following the Rasch model, the instrument will create therapy response on an interval scale. Because it is dealing with needs fulfillment the question of minimal clinical difference does not arise. A single score is reported with the contribution to needs-fulfillment quality of life calibrated over baseline. This is a 'true' interval scale unlike the ordinal raw scale for the generic measures such as the EQ-5D-3L which is a 'mashup' of clinically determined symptoms with ordinal responses for symptom severity.

QUALITY OF LIFE CLAIMS

The formulary committee is only interested in quality of life claims that meet the following standards:

- **The QoL claim must focus on the needs of the target patient group (including caregivers)**
- **The QoL claim must be demonstrated to have been developed for the target patient group with a documented audit trail**
- **The instrument should meet Rasch measurement standards'**
- **The instrument (or instruments) must report on a single attribute (e.g., needs fulfillment quality of life) with, as a consequence, unidimensional or dimensional homogeneity with interval response properties**

4.4 Tracking Quality of Life Impact

Claims for quality of life driven by either phase 3 or phase 4 trials are only a first step. They establish a baseline for response assessment. They should be tracked over the lifetime of a patient in the case of rare and chronic diseases, or for the period over which the patient is compliant with therapy. This raises the question of an evidence platform. Manufacturers must be in a position, as part of their claims assessment protocol, to propose how QoL claims are to be monitored and reported.

If there is no QoL instrument that meets Rasch measurement standards suitable to support QoL response claims in the target patient population (e.g., a needs assessment instrument) then the manufacturer should propose whether or not they intend to underwrite development of the instrument, with a meaningful timeframe for reporting to a formulary committee. If QoL is not considered a relevant outcomes measure for that target population, the manufacturer should detail why this is the case.

5. CLAIMS AND VALUE ASSESSMENT

One of the more egregious mistakes that technology assessment has promoted is the belief that there is a single gold standard metric that can be proposed for formulary evaluation. This is not only naïve but false. The I-QALY paradigm is a disaster. It is not in the interest of the formulary committee to entertain modeled claims that fail the standards of normal science. This means that claims generated from lifetime reference case models will be automatically rejected. Further, the committee is not interested in submissions which propose to apply cost-per-I-QALY thresholds to support pricing and access recommendations. Threshold based value-for-money assessments claims are, given the mathematical impossibility of an I-QALY, clearly nonsensical.

A claim that a product is cost-effective is not acceptable. This, again, is a term that has exceeded its use by date. This implies the application of a nonsensical single metric of direct medical costs. The formulary committee sets its own standards for judging whether the claimed cost of therapy for a specific product is consistent with response to therapy claims at a price proposed by the manufacturer. The committee should not be interested in presumptive claims from the manufacturer that the product is 'cost-effective'. This is up to the committee to decide once the required data elements have been submitted to the committee. Until then pricing must be provisional (and, indeed, may continue to be provisional given ongoing claims for product impact and utilization).

5.1 Claims Protocols in Practice

THE GIGANTIC GOOSEBERRY.

A noted in Section 1, all claims for product performance in the target patient population must be supported by a claims evaluation protocol. One role of the formulary committee is to review protocols submitted and agree with the manufacturer on protocol implementation and time lines for reporting. This applies not only to clinical claims but to claims for quality of life, resource utilization costs and other value assessments agreed with the formulary committee.

Clinical claims must be relevant to the target patient population. Manufacturers must be able to defend claims appropriate to the target population and expressed in terms that make them credible and evaluable for that population.

Claims must not be presented that rely exclusively on pivotal phase 3 trials; nor should claims be presented that are generated by modeled indirect comparisons from pivotal trials. Claims should be comparative and presented as protocols for evaluation in target populations. If such protocols have already been applied in target patient populations then the results should be presented. In either case, the protocol must be justified in terms of the characteristics of the target population (see Section 2.2 above). and the axioms of fundamental measurement. Dimensionally heterogeneous claims are not acceptable.

The insistence on the provision of one or more protocols to assess claims for comparative product effectiveness does not mean a commitment to RCTS; these can be time consuming and expensive. They have been used as an excuse to do nothing. Claims based on pivotal clinical trials, to include those that attempt through comparative meta-analyses to reconcile response claims for nominally comparator therapies are clearly of limited interest to formulary committees who are looking to robust claims for therapy response in target patient populations.

Protocols should look to establishing a permanent evidence base, possibly a registry, to support well-designed observational studies and continuing clams reassessment. Formulary committees are in a position to demand protocol driven claims assessment that capture the characteristics of the target patient population. Pivotal trial claims are only a first step; a tentative one at best. The protocol, apart from the essential requirement of a viable evidence base, should detail how the manufacturer proposes to assess and translate pivotal claims to those that have external validity.

Time is of the essence. It is in the interests of both the formulary committee and the manufacturer to establish claims for product effectiveness. This can be driven by the simple expedient of provisional pricing. All claims must be empirically evaluable in a timeframe that is meaningful for the formulary committee. Claims must be, in short, credible, evaluable and replicable to meet the standards of normal science. The clams must be specific to the target patient population. Where necessary for specific claims, these should be supported by a systematic review of the literature and included within the protocol. If successful claims assessment supports pricing strategies, then these must be detailed for the formulary committee.

Finally, the protocol should detail for claims, some of which may be based on instruments with interval measurement properties, others with ratio properties, the analysis that is proposed to assess the therapy response and track that response over the course of treatment. Protocols that are submitted as part of the formulary evaluation process should meet appropriate FDA standards or recommendations for RCT protocols as well as those for real world data and real world evidence. Attention needs to be given to the possibility that the as a result of the protocol implementation a manufacturer may attempt to submit additional product claims., This might apply, for example, to needs-based quality of life assessment for comparative product claims. In this case any contracted development for a new quality of life instrument should be required to meet FDA audit and reporting standards. A similar caveat would apply where other PROs are proposed that meet required measurement standards.

CLAIMS EVIDENCE REQUIREMENTS

Evidence required to support formulary decisions is entirely at the discretion of the formulary committee. The key requirement should be that the claims submitted for prospective product or device performance should be credible, evaluable and replicable. Each claim must be supported by a protocol detailing how that claim is to be evaluated and reported to the formulary committee [Section 5.1]; otherwise the claim should be rejected. Rather than a submitted reference case imaginary world being the philosopher's stone of true value, it should be regarded as just one of any number of imaginary constructs that have no place in formulary decisions. Product pricing and formulary placement should be driven by existing and prospective real world evidence. Provisional acceptance should always be conditional on prospective claims assessment. Claims may be usefully categorized for designated target populations as:

- **Clinical claims for therapy response [Section 5.2**
- **Patient centric quality of life claims [Section 5.3]**
- **Supporting clinical claims for co-morbid conditions [Section 5.4]**
- **Product Entry, Uptake and Discontinuation [61]claims[Section 5.5]**
- **Claims for impact on medical resource utilization [Section 5.6]**
- **Societal impact claims [Section 5.7]**

5.2 Clinical Claims for Therapy Response

Pivotal phase 3 trials are typically placebo referenced, often lacking external validity for a target patient population. One reason for a formulary committee insisting on a detailed profile and systematic review of clinical and associated studies for the target population is to provide a basis for evaluating submitted protocols for product claims. These must be comparative, with the usual standard of care the reference point. As well, the protocol must recognize the possibility of adverse events, detailing how these are to be detailed and reported.

While these requirements might seem daunting, this is the only basis for tracking and assessing product performance and possible guidelines for product use. Ideally, a manufacturer, as part of product development (and no later than initial planning for phase 3 studies) will have reviewed options for an evidence platform (e.g., establishing or modifying an existing registry) to support post-marketing approval formulary submissions. It would only require one well-designed evidence platform for the target population to support clinical observational studies of effectiveness, taking account of target patient population characteristics.

Additional product claims that result from access to an evidence base may have relevance in the selection of tests and the application of genomic tests to identify target patient sub-populations. The advantages of this to a health system are obvious: less wastage and improved outcomes. If genetic profiling is a component of protocol development for a product then the preferred test should be identified with evidence for its performance against other tests and the required genetic markers for patient selection detailed.

5.3 Patient and Caregiver Quality of Life Claims

If there is a requirement by a formulary committee for quality of life impact of a new product, then the manufacturer should consider developing a patient centric needs-fulfillment instrument for the target patient population. This applies both to patients and, where applicable, their carergivers. These could be family members involved with pediatric and adolescent patients, as well as the partners of older patients with dementia or other chronic co-morbid conditions.

Designing a patient centric instrument is time consuming, with best estimates of 12-15 months for completion. Again, this requires forward planning by the manufacturer with development initiated at least at the end of Phase 2 trials. Where the presence of a caregiver is an integral part of assessing response to therapy, the manufacturers should consider (and the formulary committee requesting) instruments that focus of the needs of both patients, spousal and other caregivers. In some instances the caregiver would respond on behalf of the patient. Usually, however, instruments should be developed separately for the patient and the caregiver as their needs may be distinct yet complementary.

5.4 Supporting Clinical Claims with Co-Morbid Conditions

Claims for therapy response in the presence of comorbidities involve a number of possible scenarios. The profile presented for the demographic and clinical characteristics of the target population should provide a comprehensive co-morbidity and drug utilization profile for the target population. Where a protocol is presented it must detail also how the presence of co-morbidities, including the stage of disease by comorbidity, its severity and presence of polypharmacy, are to be accommodated in evaluating therapy response. This may involve the exclusion of specific groups within the target population or targeting groups with specific characteristics. Exclusion may involve genetic profiling; in which case an adverse genetic markers (or markers) will sit alongside specific co-morbid conditions as exclusion criteria.

Certainly there is a substantial literature on co-morbidities and therapy response in many disease states. These, unfortunately, often tend to be piecemeal in their approach with a focus on

academic medical centers. The purpose of an evidence base is that it is comprehensive, capturing not only clinical and biomarker characteristics but potentially environmental and lifestyle factors that may impact therapy response.

5.5 Product Entry, Uptake and Discontinuation Claims

Claims by manufacturers for the anticipated budget impact of their product are of no interest to the formulary committee. The committee can make its own assessment. The information that is required is the manufacturer's estimates (or guesses) of: (i) product uptake; (ii) displacement of comparator therapies and (iii) product discontinuation (compliance patterns) including reasons for non-compliance. Manufacturers are asked to detail for each quarter following anticipated formulary listing and for possible 8 consecutive quarters their 'best' estimates of:

- Initial (new) prescriptions for target patients defined by dosage, pill count and formulation, separately identifying Medicare, Medicaid, other public sector and commercially insured populations
- The proportions in each of the above categories that are 'new' prescriptions as opposed to 'switch' prescriptions where a 'new prescription' is defined as one where there is no previous prescriptions for a comparator product in the six months preceding the 'new' prescription and a switch prescription where there is no coverage gap between the new prescriptions and the last or previous comparator prescription
- Anticipated displacement of comparator therapies (or therapy), defined for each therapy by the characteristics of the displaced therapy by quarter for each of the populations defined above
- Anticipated patient profiles of adherence and discontinuation detailing the average and median duration of compliance with therapy in days of prescription coverage with no gaps in new prescription uptake for each of the populations defined above
- Distribution of time to discontinuation of therapy for patients receiving a first prescription of the target therapy for Q2 through Q8 with a profile of reasons for discontinuation
- A profile for the major comparator therapies of discontinuation patterns for both individual patients by population group as well as by aggregate patterns of those introduced to therapy

These claims are to be specific to the target patient population in the target health system. National estimates are not acceptable. Manufacturers should detail the evidence bases and

procedures for extracting these data (e.g., data extraction algorithms) with procedures for reporting these to formulary committees.

Discontinuation of therapy is a critical variable. It makes little sense for patients to be introduced to new therapy if a high percentage discontinue within 6 or 12 months. If this is the projected discontinuation rate then a formulary committee may reasonably ask why patients should be introduced to this therapy in the first place. Of course, this may be a redundant argument if the reasons for discontinuation reflect a perception that the therapy is not achieving claimed responses. Manufactures, as part of their submission and prospective claims should assess the reasons for discontinuation of therapy by time since therapy initiation

5.6 Claims for Impact on Resource Utilization

An assumption driven lifetime claim for "discounted" direct medical costs makes no more sense than lifetime discounted I-QALYs; the cost assumptions lack any credibility, along with the underlying resource utilization assumptions and the attached unit costs. Claims for direct medical costs should be rejected as they not only assume unit costs that may be irrelevant to a formulary committee but in aggregating those costs to a single metric create a measure that lacks dimensional homogeneity. The focus must be on credible claims for resource utilization that are, once again credible, replicable and evaluable within a meaningful timeframe. This means claims for defined individual resource units.

Apart from estimated drug utilization, manufacturers should detail resource utilization required to support new therapies in terms of claims for measureable units: physician visits by type; urgent care visits, and hospitalizations. These are readily assessed from administrative claims databases and hospital claims databases. Protocols for resource claims must detail the recommended databases and the data extraction protocols. These should be matched to claims for comparator therapies following drug switching claims and the uptake of the new therapy.

The resource utilization protocol should detail both claims for resources to support the new and comparator therapies and the net resource utilization impact. This impact should be detailed for each of 8 quarters following formulary listing of the new therapy. The formulary committee is then in a position to consider the appropriate unit costs to apply to provide a basis for costs.

Clearly, claims for anticipated resource unit utilization must match claims for product uptake, displacement and discontinuation. Manufacturers should detail the procedure for matching these claims. In cases where resource utilization is a function of product uptake with 'new' patients initially consuming more resource units (e.g., titrating to an optimum dose) than 'established' patients the patterns of discontinuation matching to new patient enrollment need to be clearly stated.

5.7 Societal Impact Claims

Whether societal impact claims should be part of a submission to a formulary committee is a moot point. They can be part of an imaginary ICER-type reference model which is attempting, unsuccessfully to make a generic cost-per-I-QALY case. Otherwise, there seems little scope given the virtual impossibility of evaluating those claims

It is at the discretion of the formulary committee as to whether or not decisions on formulary listing and pricing should include the wider potential societal impact costs of acceptance of the target therapy. Claims for societal impact are typically vague. The most commonly cited are productivity costs, although in imaginary modeled claims there is no attempt to propose how they are to be assessed.

If the formulary committee believes that societal claims for product performance are relevant for decisions for formulary acceptance and pricing, then the manufacturer should develop protocols to report on those claims. In the case of productivity claims, the manufacturer should provide a protocol detailing how therapy pact might be evaluated to include both reduced productivity as well as absence from employment given the employment characteristics (e.g., occupation, industry, hours of work, wage rate) of the target patient population.

5.8 Claims Uncertainty

In the imaginary world of lifetime models, uncertainty is a security blanket. It is a fog that is called into play to disguise the fact that the claims are driven by assumption, lacking credibility and any possibility of empirical evaluation. These are typically accompanied by likelihood estimates and tornado diagrams to point to the dominant assumption impacts.

Understandably, manufacturers may balk at the request for claims at the level of detail and the timeframe proposed. There is no objection to manufacturers admitting to degrees of uncertainty with credible claims made for the new therapy. Uncertainty bounds may even be proposed. In some cases the claims made may be nothing more than educated (or less educated) guesses. The point is that if a formulary committee is to consider claims made for a new therapy then the committee is quite reasonably asking details for the claims and, through protocols, how these claims are to be assessed.

6 CHECKLIST FOR A FORMULARY SUBMISSION

Accepting a paradigm shift, the rejection of the existing technology assessment paradigm, exemplified by the ICER and AMCP modelled reference case frameworks, will not be easy. A situation, perhaps paradoxically, made that more difficult by the incontrovertible fact that the multiattribute utility scales that are the foundation for I-QALY claims, are ordinal scales. This seems such an obvious observation that many will be concerned that this objection had not surfaced earlier or why it had been ignored. It is incumbent on formulary committees to make clear that manufacturers invited to make formulary submissions must recognize this rejection of imaginary approximate information claims.

6.1 REQUEST FOR A FORMULARY SUBMISSION

Unsolicited submissions to a formulary committee should be rejected. It is up to the formulary committee (or equivalent decision group) to determine the timing of a request, the claims content (i.e., the attributes of primary interest to the committee for the target patient population) and the timing of the review process.

The request for a submission should detail the criteria under which a submission will be received. It should be made clear that the committee is not interested in product claims that are multidimensional and lack credibility. It should also be emphasized that individual claims should be accompanied by a protocol to demonstrate have it is to assessed and reported to the committee in a meaningful timeframe.

It should be made clear that the committee is certainly not interested in claims that are driven by imaginary cost-per-I-QALY simulations as these fail the demarcation test that characterizes normal science. The bankrupt ICER reference case, although supported by contracted academic centers, is out. Formulary review and decisions will be based on real world evidence. This may be limited in the period following marketing approval so that formulary decisions will be provisional The committee will focus on how claims are to evaluated, the evidence platforms to support those claims and the protocols proposed by the manufacturer to evaluate those claims.

6.1.1 Draft Request for Submission Letter

DRAFT REQUEST FOR SUBMISSION LETTER

Dear ……

You are invited to make a submission for [your product] to enable our formulary committee to assess whether or not, on the available and hopefully prospective evidence, that [your product] may be considered and possibly provisionally approved for pricing negotiations and listing.

As you may be aware, this formulary committee has endorsed a new approach to the provisional and continued approval of pharmaceutical products and devices. The focus is on real world evidence and claims for pharmaceutical products and devices that are credible, evaluable and replicable for the target patient population; claims should meet the standards of normal science, recognizing the constraints of fundamental measurement. Claims must be for specific attributes for the indicated target population.

The committee is not interested in simulated approximate non-evaluable claims, but in manufacturers being able to demonstrate that claims can be, and how they are proposed to be, evaluated in the target patient population. This means that modeled incremental cost-per-I-QALY claims will be rejected or any reference to such claims to support value assessments.

Claims that the product is 'cost-effective' are not required. The committee will make a decision on the evidence presented. All that is required of manufacturers is to provide the committee with real world evidence to support its decisions.

Attached please find a checklist that you should complete and provide to the committee as part of your submission.

Sincerely …..

6.2 FORMULARY SUBMISSION CHECKLIST

This checklist is both an *aide-mémoire* to emphasize the focus on real world evidence and claims assessment through protocols as well as a checklist for the required elements that should be submitted. Manufacturers should respond to each question, providing, if necessary, additional details to clarify their response or any perceived obstacles to reporting on claims. As a first step manufacturers should summarize the value attributes they propose to evaluate in the target patient population together

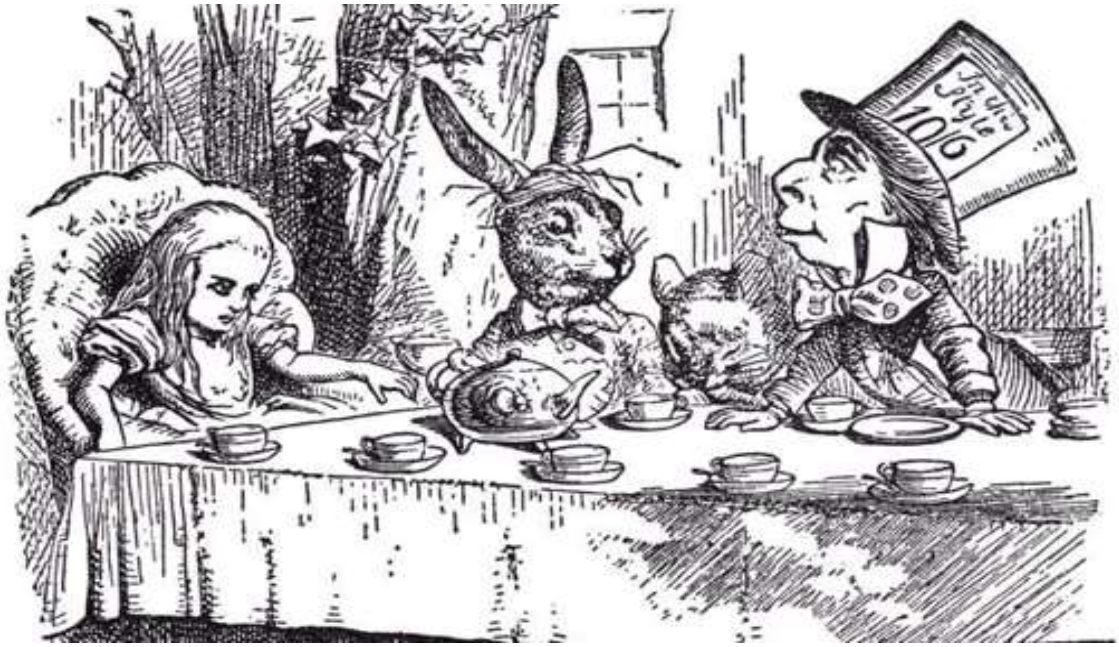
with a timeline for evaluating these attributes and reporting to the formulary committee.

Required Elements	Response by Manufacturer
Value Claims and Evidence Base	
Have you provided a summary list of the attributes you propose to evaluate (or have evaluated) in the target patient population?	
Are all value claims made for your product either supported by evidence or capable of empirical evaluation?	
Are all your value claims comparative and have you detailed the comparators for each claim?	
Are all value claims for your product that require clinical evaluation capable of being reported to the formulary committee within 18 months?	
Are the value claims for your product based on instruments that meet the standards for fundamental measurement including dimensional homogeneity?	
Have you provided evaluation protocols for proposed product value claims?	
Do your proposed evaluation protocols detail the evidence base for their evaluation?	
If your proposed evidence base is a registry have you provided details on the structure and management of the registry?	
If your evidence base involves administrative claims or other 'big data' sources have you detailed agreements with vendors for access, data extraction and reporting?	
Have you provided for each claims protocol a timeline for reporting the results of the evaluation to the formulary committee?	
Will the evidence base for any claim support future requests from the formulary committee for revisiting claims as part of ongoing disease area and therapeutic class reviews?	
Target Patient Population	
Have you described and defined (e.g., ICD-10-CM codes) the target patient population to provide a common algorithm for defining the population in different data sets?	
Where claims for your product and comparators are presented in your submission (e.g., from RCTs) do the protocols define the target patient population that would	

be encountered in treatment practice?	
If protocols to define the target treated population describe only a subset of that population have you identified those characteristics (e.g., age, gender, co-morbidities) that have been excluded?	
Have the protocols you are proposing for claims assessment included characteristics of the target population that would maximize the external validity of your assessment?	
Target patient population; Have you provided an epidemiological profile covering the following characteristics of the proposed claims assessment: • **Data sources** • **Population estimates** • **Incidence** • **Basic demographics** • **Socioeconomic status** • **Insurance status** • **Drug utilization** • **Clinical status** • **Genomic profile** • **Comorbidity status** • **Caregivers**	
Have you provided a summary and critique of the PRO instruments and their measurement properties that have been applied in studies focusing on response to therapy for the target patient population?	
Have you provided a review and possible reconciliation of the various RCT and observational study protocols and their measurement properties that have supported response claims for products in the target patient population?	
Have you provided a review and summary of the unmet medical needs for the target patient population?	
Have you provided a summary of pipeline products and prospective comparator therapies that may enter the market in the next five years for the target patient population?	
Clinical Evidence	
Where systematic reviews and meta-analyses are	

presented, have you excluded those studies where the PRO instruments fail to meet the required standards for fundamental measurement?	
Where an evidence hierarchy is proposed for the various observational studies, have you discounted those studies where PRO instruments fail to meet the standards for fundamental measurement?	
In retrospect, have any of the outcomes measures (primary and secondary) used in your use of pivotal RCTs failed to meet the standards of fundamental measurement?	
Quality of Life	
Have you ensured that any quality of life claim for your product in the target patient population meets standards for fundamental measurement?	
Have you proposed a needs-fulfillment QoL instrument for claims that meets the requirements of Rasch Measurement Theory?	
If you have not proposed a Rash needs fulfillment instrument, why not?	
How do you propose to report on QoL claims for your product and comparator products in the target patient population?	
Outcomes and Resource Utilization	
Have you provided claims assessment protocols for each of the following outcomes and resource utilization categories in the target patient population: • **Clinical claims for therapy response** • **Patient and caregiver centric QoL claims** • **Clinical claims with co-morbid conditions** • **Product entry, uptake and discontinuation** • **Impact on medical resource utilization** • **Societal impact**	
Disease Area and Therapeutic Class Reviews	
Has the company provided a commitment to support prospective disease area and therapeutic class reviews for the target patient population?	

THANK YOU



REFERENCES

[1] Langley P. The Great I-QALY Disaster. *Inov Pharm*. 2020;11(3): No. 7 https://pubs.lib.umn.edu/index.php/innovations/article/view/3359/2517

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

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

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

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

[8] Briggs R. The Scientific Revolution of the seventeenth century. Longman, 1971.

[9] Ollendorf DA, Bloudek L, Carlson JJ, Pandey R, Fazioli K, Chapman R, Bradt P, Pearson SD. Targeted Immune Modulators for Ulcerative Colitis: Effectiveness and Value; Evidence Report. Institute for Clinical and Economic Review, September 11, 2020

[10] Langley P. The Impossible QALY and the Denial of Fundamental Measurement: Rejecting the University of Washington Value Assessment of Targeted Immune Modulators (TIMS) in Ulcerative Colitis for the Institute for Clinical and Economic Review (ICER). *InovPharm*.2020;11(2): No 17 https://pubs.lib.umn.edu/index.php/innovations/article/view/3330/2533

[11] Langley P. Validation of modeled pharmacoeconomic claims in formulary submissions. *J Med Econ*. 2015;18(12):993-99

[12] Dawkins R The Selfish Gene. 30th Anniversary Edition. New York Oxford University Press, 2006

[13] Dawkins R. A Devil's Chaplain. New York: Houghton-Mifflin, 2003

[14] 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

[15] ICER. 2020 Value Assessment Framework. 31 January 2020 https://icer-review.org/wp-content/uploads/2019/05/ICER_2020_2023_VAF_013120-1.pdf

[16] Langley PC. Nonsense on Stilts Part 1: The ICER 2020-2023 Value assessment Framework for Constructing Imaginary Worlds. *InovPharm*. 2020;11(1).No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444

[17] Campobell JD, Kaló Z. Fair global drug pricing. *Expert Rev Pharmacoeconomics Outcomes Res*. 2018;18(6):581-83

[18] Tran VL, Leirvik T. A simple but powerful measure of market efficiency. *Financial Res Letters*. 2019;29:141-51

[19] Pearson SD, Lowe M, Towse A et al. White Paper: Cornerstones of "Fair" drug coverage: Appropriate cost-sharing and utilization management policies for pharmaceuticals. Institute for Clinical and Econo9mic Review and Office of Health Economics. September 2020 . https://icer-review.org/material/cornerstones-of-fair-drug-coverage/

[20] Academy of Managed Care Pharmacy. AMCP Format for Formulary Decisions (Format 4.1). AMCP, 2020 (released December 2019) https://www.amcp.org/sites/default/files/2019-12/AMCP_Format%204.1_1219_final.pdf

[21] Academy of Managed Care Pharmacy. *Format for Formulary Submissions (Version 4)*. AMCP: April 2016

[22] Langley P. Modeling imaginary worlds: Version 4 of the AMCP Format for Formulary Submissions. *Inov Pharm*. 2016;7(2): Article 11 https://pubs.lib.umn.edu/index.php/innovations/article/view/434/429

[23] Baltussen R, Marsk K, Thokala P et al. Multicriteria Decision Analysis to Support Health Technology Assessment Agencies: Benefits, Limitations, and the Way Forward. *Value Health*. 2019;22(11):1283-88

[24] Grimby G, Tennant A, Testo L. The use of raw scores from ordinal scales: Time to end malpractice (Editorial) *J Rehab Med*. 2012;l44:97-8

[25] Bond T, Fox C. Applying the Rasch Model . New York: Routledge, 2015

[26] Mapi Trust. PROQOLID. https://eprovide.mapi-trust.org/about/about-proqolid

[27] Tufts Medical Center. Cost-effectiveness Registry https://cevr.tuftsmedicalcenter.org/databases/cea-registry

[28] Langley PC, McKenna SP. Measurement, modeling and QALYs [version 1; peer reviewed] *F1000Research* 2020, 9:1048 https://doi.org/10.12688/f1000research.25039.1

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

[30] McKenna S et al. The limitations of patient reported outcomes measurement in oncology., *J Clin Pathways*. 2016;2(7):37-46

[31] Luce R, Tukey J. Simultaneous conjoint measurement: A new type of fundamental measurement. *J Math Psychol.* 1964;1(1):1-27

[32] Rasch G. Probabilistic models for some intelligent and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960

[33] Gibbons C, Thornton E, Ealing J et al. Assessing social isolation in motor neurone disease: A Rasch analysis of the MND Social Withdrawal Scale. *J Neuro Sci*. 2013;334:112-118

[34] Lambert S, Pallant J, Boyes A et al. A Rasch analysis of the Hospital Anxiety and Depression Scale (HADS) among cancer survivors. *Psychol Assess*. 2013;25(2):379-90

[35] Zucca A, Lambert S, Boyes A et al. Rasch analysis of the Mini-Mental Adjustment to Cancer Scale (mini-MAC) among a heterogeneous sample of long-term cancer survivors: a cross-sectional study. *Health Qual Life Outcomes*. 2012;10:55

[36]Shea T, Tennant A, Pallant J. Rasch model analysis of the Depression, Anxiety and Stress Scale (DASS). *BMC Psychiatry*. 2009;9:21

[37] Stevens S. On the Theory of Scales of Measurement, *Science*. 1946;103:677=80

[38] Fiorjaz M, Martinez-Martin P, Dujardin K et al. Rasch analysis of anxiety scales in Parkinson's disease. *J Psychosom Res*. 2013;74(5):414-9

[39] 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):474-80

[40] McKenna S, Rouse M, Heaney A et al. International development of the Alzheimer's Patient Partners Life Impact Questionnaire (AAPLIQue). *Am J Alzheimer's Dis Oth Dementias*. 2020 35:1-11

[41] G Ellis, J Silk. Scientific Method: Defend the integrity of physics. *Nature* 16 December 2014 https://www.nature.com/news/scientific-method-defend-the-integrity-of-physics-1.16535

[42] Cameron D, Ubels J, Norström F. On what basis are medical cost-effectiveness thresholds set? Clashing opinions and an absence of data: a systematic review. *Global Health Action*. 2018;11:1447828

[43] Langley P. Recent Developments in the Health Technology Assessment Process in TR Fulda and A I Wertheimer, *Handbook of Pharmaceutical Public Policy*, New York, Haworth Press, 2007, pp. 457-477.

[44] Langley P. Nullius in Verba: Version 2.0 of the University of Minnesota, School of Social and Administrative Pharmacy Program, Proposed Guidelines for Formulary Evaluation. *Inov Pharm.* 2016;7(4): No 16 https://pubs.lib.umn.edu/index.php/innovations/article/view/473

[45] Aloinso-Coello P, Schünemann H, Moberg J et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction.. *BMJ*.2016;353:i2016

[46] Guyatt G, Oxman A, Aki E et al. GRADE Guidelines 1. Introduction – GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64:3483-94

[47] Porter M, Larsson S, Lee T. Standardizing patient outcomes measurement. *N Eng J Med*. 2016;364:6

[48] Doi S, Ito M, Takebayashi Y et al. Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. *PLoS ONE*. 2018; 13(7):e0199235

[49] McKenna SP, Whalley D, Doward LC. Improving the measurement properties of the quality of life in depression scale. Poster presented at the 2002 ISPOR 5th Annual European Congress; November 2002. Rotterdam, The Netherlands. [abstract] *Value Health*. 2002 Nov 3; 5(6):522.

[50] Kuhn T. The Structure of Scientific Revolutions. Chicago: The University of Chigo Press, 1962,

[51] Bartlett V, Dhruva S, Shah N et al. Feasilbility of using real-world data to replicate clinical trial evidsence. *JAMA Network Open*. 2020;2(10):e1912869

[52] Moher D, Liberati A, Tetzlaff J et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Ann Int Med*. 2009;151(4): 264-270

[53] Page M, McKenzie J, Bossuyt P et al. Mapping of reported guidance for systematic reviews and meta-analyses generated a comprehensive item bank for future reporting guidelines. *J Clin Epidemiol*. 2020;118:60-68

[54] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Ann Int Med*. 2010:152(11). See also: The CONSORT Statement 25-item check list [http://www.consort-statement.org/checklists/view/32-consort/66-title] and flow diagram [http://www.consort-statement.org/consort-statement/flow-diagram] to record the progress of patients through the trial.

[55] Calvert M, Blazeby J, Altman D et al. Reporting of patient reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA*. 2013;309(8): 814-22

[56] Meader N, King K, Llewellyn A et al. A checklist designed to aid consistency and reproducibility of GRADE assessments: development and pilot validation. *Systematic Rev*. 2014;3:82

[57] Cochrane Handbook: [http://handbook.cochrane.org/part_2_general_methods_for_cochrane_reviews.htm].

[58] Berkman MD, Lohr KN, Ansari M et al. Grading the strength of a body of evidence when assessing health care interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An update. Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13 (14)-EHC130EF. Rockville MD: Agency for Healthcare Research and Quality. November 2013.

[59] Lugnér AK, Krabbe P. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. *Exp Rev Pharmacoeconomics Outcomes Res*. 2020; 29(4):331-342

[60] Roudijk B, Donders R, Stalmeier P. Setting dead at zero: Applying scale properties to the QALY model. *Med Decis Making*. 2018;38(6): 627-34