Value Assessment, Real World Evidence and Fundamental Measurement: Version 3.0 of the Minnesota Formulary Submission Guidelines

Paul C Langley, PhD, Adjunct Professor, College of Pharmacy, University of Minnesota

Abstract

This latest version of the Minnesota guidelines is intended to reassert the application of the standards of normal science in formulary submissions for new and existing pharmaceutical products and devices. This represents a paradigm shift from the existing value assessment standards which are focused on imaginary or I-QALY modeling of lifetime claims. The proposed new paradigm rejects this as pseudoscience; a failure to recognize the standards of normal science, in particular a failure to recognize the constraints of fundamental measurement. As a result, current health technology assessment is dominated by value assessments that create claims that are neither credible, nor empirically evaluable or replicable. The fatal flaw is the failure to recognize that QALYS are an impossible mathematical construct (hence the term I-QALY). The proposed paradigm recognizes that if there are claims for product value then, regardless of whether the claim is for clinical impact, quality of life or resource utilization, all claims must be empirically evaluable. If not, then they should be rejected. The Minnesota guidelines propose a new evidence based approach to formulary assessment, together with ongoing disease area and therapeutic class reviews. The focus is on claims that are specific to target patient populations that are claims for specific attributes and are consistent with the axioms of fundamental measurement. Manufacturers are asked to support claims assessment through protocols detailing the evidence base for claims assessment, the timelines for those assessments and the process by which claims assessments are reported back to formulary committees. Value assessment leads naturally to value contracting, revisiting provisional prices as new information is discovered and delivered to the formulary committee.

Keywords: Minnesota Guidelines, abandoning I-QALYs, single attribute claims, claims protocols, evidence base

Introduction: The Impossible I-QALY Paradigm

To understand the importance of the Minnesota guidelines means understanding why the I-QALY paradigm for value assessment was doomed from the start [1]. In the early 1990s there was an agreed decision by the 'leaders' in health technology assessment to put to one side the standards of normal science in favor of supporting value assessments through the creation of approximate information [2]. This involved a commitment to reference case incremental cost per quality adjusted life year (QALY) models, best exemplified by the National Institute for Health and Clinical Excellence (NICE) reference case with cost-per-QALY thresholds [3]. These took center stage and NICE was emulated worldwide, including the Institute for Clinical and Economic Review (ICER) in the US and with support from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) through its practice standard guideline publications.

What is astounding is that those advocating I-QALY reference cases, with generic utilities (notably from the EQ-5D-3L scale) had no apparent idea of measurement theory and the limitations imposed on QALYs by the axioms of fundamental measurement on the ability to use utility scales to create I-QALYs. A review of the stream of ISPOR publications gives no hint whatsoever that measurement theory was a consideration. In fact, recognition of the standards of normal science to ensure that claims were credible, evaluable and replicable was absent. The result was, in retrospect, a disaster. A Faustian bargain had been struck to put normal science to one side in favor of the easy way out adoption of pseudoscience [4]. Rather than focus on evidence gaps in product value assessment, these gaps were filled by assumption and guesswork. Everyone joined in: academic institutions, formulary committees and health departments. For 30 years they went, not unchallenged, but in apparent ignorance of the unintended, long term consequences of Faustian bargains.

It is gradually beginning to dawn on those in technology assessment that they have wasted the past 30 years on a value assessment paradigm that fails the elementary mathematical logic of measurement theory. Utility scales are ordinal. To create a QALY you need a ratio scale. Those developing utility scales such as the EQ-5D-3L had, with few exceptions who were ignored, no idea that if you want a scale to have ratio properties it has to be developed to have those properties from the get-go. The EQ-5D-3L has neither interval, let alone ratio properties. It is an ordinal scale which cannot support multiplication and hence QALYs. To assume otherwise is sheer and utter nonsense [5]. The reason is obvious: the EQ-5D-3L algorithm (or utility equation) can generate negative values. This is made explicit in the standard technology assessment textbooks and labeled as 'states worse than death' [6]. By definition, a ratio scale has a true zero. It cannot have negative values. In instrument development recognition of this requirement is commonplace

Corresponding author: Paul C Langley, PhD
Adjunct Professor, College of Pharmacy
University of Minnesota, Minneapolis, MN
Director, Maimon Research LLC; Tucson, AZ
Email: langley@maimonresearch.com
Website: www.maimonresearch.net

DOI: https://doi.org/10.24926/iip.v11i4.3542

in the physical sciences and the mature social science such as economics and education. The mistake too many made was to put raw EQ-5D-3L scores on a number line with interval scores and assume this meant the EQ-5D-3L had interval scores.

The Minnesota 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 [7]. 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, 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, notably in respect of the application of fundamental measurement [8]. The focus is on 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 within a methodology that meets the standards of normal science, this paradigm shift is a rejection of a pseudoscientific paradigm (e.g., intelligent design) to a mature 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.

The Minnesota paradigm brings together three elements: value assessment, real world evidence and fundamental measurement. These are subsumed under the umbrella requirement that the claims for any pharmaceutical product or device must, if value is to be assessed, meet the standards of normal science. Approximate information is rejected in favor of hypothesis testing. Utility scores are recognized, but only in respect of their ordinal properties; patient reported outcomes (PROs) are recognized, but their application is restricted by their measurement properties; claims are recognized but only if they have interval or ratio properties; and all claims must be for single value attributes and dimensionally homogeneous. All claims are disease specific and relevant to the target population in that disease area.

In practical terms this represents a major shift in value assessment. Perhaps the most striking feature is the requirement that manufacturers, if they are to have a claim recognized, must provide a protocol to the formulary committee detailing how that claim is to be assessed, a description of the evidence base and the timeframe for reporting back to the formulary committee.

Guidelines: Structure

The Minnesota Guidelines comprise six main sections. These are:

- Section 1: A New Paradigm for Value Assessment
- Section 2: The Target Patient Population
- Section 3: Clinical and Evidence Standards
- Section 4: Quality of Life: Patients and Caregivers
- Section 5 Claims and Value Assessment
- Section 6: Checklist for Formulary Submissions

Section 1: A New Paradigm for Value Assessment

The term paradigm is applied to emphasize the importance of the rejection of what is described as the I-QALY paradigm, although it is not to be interpreted, as Kuhn does, as representing the overthrow of activities within normal science. Rather, it is the rejection of a paradigm (or possibly more appropriately a meme) that accepts a pseudoscientific framework for non-evaluable value claims in favor of the Minnesota paradigm that returns to normal science with a commitment to credible and empirically evaluable value claims. This is by far the most lengthy yet most important part of the Minnesota guidelines as it sets out why the I-QALY paradigm is well past its use by date, if it ever had one. Section 1 comprises five sub-sections. These are:

- Meeting the Standards of Normal Science
- Rejecting Imaginary Worlds
- Meeting the Standards for Real World Evidence
- Exeunt QALYs and Thresholds
- Meeting the Standards for the Minnesota Value Assessment Paradigm

Meeting the Standards of Normal Science

This sub-section focuses on the failure of the existing technology assessment paradigm to meet the standards of normal science. Issues covered include the central role of hypothesis testing in the discovery of new facts. Since the 17th century science has progressed through claims development and the evaluation of those claims [9]. If this standard was applied, as it is in product development, to the impact of competing therapies in real world treating environments then the focus should be on creating the evidentiary environment and to continue to evaluate claims for therapies. Rather, the existing approach is to bring together, within a lifetime model, evidence from pivotal trials and assumptions based on the literature. The limited data at product launch is subsumed in the model so that, by assumption, non-evaluable claims can be made. This approach is shown to be nothing more than pseudoscience. Modeling to create non-evaluable claims has to be rejected. It is pointed out that in validating modeled claims in, for example, ICER evidence reports, no account is taken of empirical evaluation. The focus is on creating approximate imaginary information, which in the emphasis on multiattribute

DOI: https://doi.org/10.24926/iip.v11i4.3542

utility measures, ignores completely the interests of patients and caregivers. The modeling exercise collapses because it fails to create credible, evaluable and replicable claims. It is pseudoscience.

Rejecting Imaginary Worlds

This section raises the question of why so obvious a rejection of the standards of normal science in favor of the creation of approximate evidence to make formulary committee claims has survived for over 30 years. Why has there not been a groundswell of opinion to reject this impossible paradigm? The arguments presented in the guidelines focus on the concept of a technology assessment meme and the transmission fidelity of the I-QALY dogma. A key point is that leaders in this field, typically academic, have never been successfully challenged (or at least challenges are brushed aside by gatekeepers such as journal editors). Beginning in the late 1980s and early 1990s the decision was made by so-called leaders in technology assessment that if they were to make a case for the cost-effectiveness for a new product then, with limited data on market entry, they would reject hypothesis testing in favor of creating 'approximate' evidence with uncertainty captured by sensitivity and scenario analyses. The notion of fundamental measurement was absent. In large part due to the ease of constructing imaginary modeled claims and a ready market among manufacturers, the I-QALY paradigm gained widespread acceptance. This was reinforced from the late 1990s by professional associations such as ISPOR and the adoption of the imaginary reference case frameworks by organizations such as NICE in the UK and other single payer health systems.

A further point that underpins the lack of awareness in the standards of normal science is the belief that assumptions can drive future imaginary claims in creating simulation models. This is logically indefensible. 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*' [10]. 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.

Certainly there were 'voices in the wilderness' attempting to point out the importance of fundamental measurement [11 12 13]. Again, they were drowned out by organizations promoting the various multiattribute utility scores such as the EQ-5D-3L and later the EQ-5D-5L; fundamental measurement was ignored. The continued acceptance of the I-QALY dogma, the guidelines emphasize, appears to be a postmodernist sociological phenomena where truth is malleable not universal. Truth in technology assessment is consensus.

Overturning an entrenched paradigm (or meme) is always a tall order. The I-QALY meme, the technology assessment belief system, is well entrenched [14]. Faced with arguments based on fundamental measurement, practitioners simply refuse to believe that the EQ-5D-3L is an ordinal score. It is almost a dogma; a belief system which after 30 years is still resistant to criticism. ICER, as a case study, understands, but cannot prove, that the EQ-5D-3L is a ratio scale; it 'understands' or 'believes'. Perhaps, as the guidelines note, ICER and others believe in the I-QALY because it is an impossible construct.

The guidelines also point out that the I-QALY meme is bolstered by the application of cost-per-QALY thresholds to create 'fair pricing' recommendations. While these are clearly nonsense they have a simple appeal to decision makers; many of whom are unaware of the I-QALY disaster. When ICER has been sked to prove that the utility scale has ratio properties it, as noted, has evaded the question [15]. This is not surprising as it is impossible to prove that a scale with negative values has a true zero.

Meeting the Standards of Real World Evidence

The patient is, presumably, the beneficiary of improved therapies together with the caregiver and the patient's family. Given this, as the guidelines point out, it is puzzling that the EQ-5D-3L focuses on capturing the community assessment of therapy response; not the patients. The patient may respond in terms of the five health dimensions and three response levels that characterize the EQ-5D-3L, but this misses entirely the question of whether, in their own frame of reference, the patient and their caregiver benefit.

As the guidelines note, there are two issues here: (i) are the patient and caregiver needs represented by generic multiattribute measures and (ii) do these measures meet the required standards of fundamental measurement. In respect of generic multiattribute instruments and the majority of patient reported outcomes (instruments) the answer to both questions is that they do not.

The guidelines make clear that, if the framework for technology assessment is to be taken seriously, then the quality of measurement, the development of measures to calibrate response to therapy, must meet the standards that are common in the physical sciences. Patient reported outcomes (PROs) and other instruments must be designed to meet the standards of fundamental measurement. If response to therapy is to be assessed than the measure must have interval properties. The guidelines make these requirements clear by providing a brief introduction to measurement scales.

The guidelines also point to the limitations of composite measures [16]. These are commonplace in PRO measures in capturing different dimensions of health experience and then adding these together to create a composite score. This, once

DOI: https://doi.org/10.24926/iip.v11i4.3542

again, invalidates the axioms of fundamental measurement in failing to meet the standard for dimensional homogeneity. They are multidimensional rather than unidimensional; they have only ordinal properties. This lays the foundation for the focus in the guidelines on instruments that capture single attributes in therapy response.

The answer, according to the guidelines, is to recognize the role of Rasch measurement theory (RMT) [17] . If this is not recognized, then PRO instruments will have only ordinal characteristics. 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. The Rasch model, utilizes a modified form of the axioms of conjoint simultaneous measurement, to assess patterns in a matrix of expected response probabilities. 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 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 scores 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.

The guidelines, in the context of RMT, consider as a key measure of response to therapy, needs fulfillment quality of life (QoL); a single attribute relevant to patients as well as caregivers [18] . The hypothesis is that the benefits patients (and caregivers) derive from a therapy intervention is the extent to which it supports greater needs fulfilment in the target patient population. Within disease states, QoL, the value placed by individual lives on competing therapy interventions, is dependent on the extent to which their human needs are met; the presence of disease and the impact of interventions drive QoL.

The QoL defined by needs fulfilment needs to be assessed directly from patients in the disease state. Attempting to infer indirectly, through the impact of interventions on health related quality of life (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 meeting patient needs [19] [20]. Non-clinical factors may modify the impact of therapeutic interventions. Needs fulfilment as a latent construct sets it aside from instruments that take the 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 such 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 as long as they meet the required measurement standards.

Exeunt QALYs and Thresholds

The guidelines are quite clear: there is no place for the I-QALY in lifetime simulation models. More to the point the guidelines list claims that are unacceptable:

- Claims that fail the axioms of fundamental measurement which means accepting only interval 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 from imaginary simulations
- Claims for equal value of life years gained from imaginary simulations
- Claims that are non-comparative or exclude a comparator product agreed with the formulary committee

As noted, rejecting the I-QALY means the rejection of I-QALY thresholds. The guidelines point out those attempts to define a 'fair' or acceptable price by the imposition of cost-per-QALY (or I-QALY) thresholds (e.g., $100,000 per QALY) are mathematically impossible constructs [21]. If the I-QALY is impossible then the threshold lacks any practical application; the entire exercise and any recommendations for pricing and access are just nonsensical. Indeed, ICER has attempted to defend its position by claiming that utility scales have ratio properties or, more correctly, are in fact ratio scales in disguise.

For claims to be accepted by a formulary committee they must be accompanied by a protocol detailing how the claim is to be evaluated in a real world environment, the timeframe for evaluation and how it will be reported to the formulary committee. The only exception here is where claims have already been evaluated following the guidelines protocol in other health jurisdictions for the same target population.

DOI: https://doi.org/10.24926/iip.v11i4.3542

A key element in evaluating claims is access to an acceptable evidence base. This could include registries or other observational frameworks. Where a number of claims are proposed they should all relate to the target patient population. If manufacturers propose to assess claims for, say, resource utilization then they have to demonstrate, if the data are from administrative claims, that the patient characteristics match those for clinical and quality of life claims. Ideally, the evidence base should be able to support ongoing disease area and therapeutic class reviews. If so, then details should be presented regarding how the viability of the evidence base is proposed to be maintained over time.

Meeting the Standards for the Minnesota Value Assessment Paradigm

Unlike technology assessment where there is no intention, in making incremental cost per I-QALY claims, to appreciate the 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 within a prior value framework 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).

The guidelines propose seven value assessment standards that are at the core of the new paradigm. These are

- 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

Section 2: The Target Patient Population

A coherent framework supporting real world evidence assessment is critical to formulary submissions and decisions. Central to this is the notion of an ongoing evidence base. An example would be an 'evidence registry'; a registry capturing a representative sample of the target patient population which can support initial and ongoing assessments of claims. To establish such a registry manufacturers should be able to demonstrate:

- 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

It is critical that manufacturers show a detailed understanding of the target patient group. As part of the formulary submission manufacturers should provide a comprehensive quantitative evaluation of the target patient group, for both the overall US population and for the target patient group in the health system represented by the formulary committee. Proposed elements of this profile 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

DOI: https://doi.org/10.24926/iip.v11i4.3542

- **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 be 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; patients with dementia) in the target population
- **Social Factors**: extent to which environmental, income and lifestyle factors impact drug access and utilization

At the same time, the manufacturer should be asked to provide a systematic review of the generic and disease specific PROs that have been used to support existing claims for the therapies common to the target patient population together with their measurement properties. The intent here is to separate those PROs that meet the accepted standards of fundamental measurement from those that fail. The formulary committee is interested only in the former category.

For a formulary committee to judge the commitment of a manufacturer to a product, it is important to have a detailed profile on completed, ongoing, completed and proposed RCTs and observational studies. The latter would include links to patient advocacy groups and possible joint projects underway or anticipated. Again the RCT protocols should be subject to a review of the measurement standards of all primary and secondary outcomes.

External validity of trial based claims is a perennial concern to formulary committees. A submission should detail the protocol exclusion and inclusion criteria, by relevant clinical trials, 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 should cover trials that have been completed, ongoing and proposed, together with prospective competitors for their product.

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 that the trial protocol population 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

Section 3: Clinical Evidence Standards

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).

The issue that is emphasized in the Minnesota guidelines is that any claim for comparative response to therapy is provisional. A framework must be in place, agreed to with the formulary committee, for monitoring and reporting therapy response. Certainly, attempts to replicate pivotal clinical trial claims are important; but only if the instrument(s) used to evaluate response to therapy meet the standards of fundamental measurement. It seems somewhat of a waste of time to replicate claims with instruments that fail to meet these standards. This means that only claims that are based on instruments that meet single attribute measurement properties are accepted.

One of the more unfortunate aspects of clinical assessment is the failure by 'authorities' to recognize the critical role of meeting the axioms of fundamental measurement. The authors of the Consolidated Standards of Reporting Trials (CONSORT) for the evaluation of trial organization, analysis and interpretation were, apparently, unaware of fundamental measurement constraints [22 23]. The question to be addressed is whether the RCT primary and secondary endpoints meet the standards for fundamental measurement. Unfortunately, trial protocols and consequent marketing approvals by the FDA may have accepted measures which failed these standards. The manufacturer is thus in an awkward position where the formulary committee may reject trial results that fail to meet its requirement for fundamental measurement even though the FDA has approved the product in ignorance of those constraints.

DOI: https://doi.org/10.24926/iip.v11i4.3542

Similar issues arise in the case of grading evidence where the manufacturer, in presenting a submission, applies the Grading of Recommendations Assessment Development (GRADE) framework [24]. The GRADE rankings pay no attention to the axioms of fundamental measurement. Again, a GRADE ranking may be rejected by the formulary committee if the relevant measures fail to meet the required standards.

In summary, the formulary committee should not be 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 measurement standards.

Section 4: Quality of Life - Patient and Caregiver Needs

The stand taken by the Minnesota guidelines is quite clear: quality of life claims that fail to meet the standards or fundamental measurement are unacceptable. This includes multiattribute as well as other ordinal scales. Claims for QoL 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.

As detailed in Section 1 of the guidelines spurious claims, such as those made by ICER that there is an 'understanding' that the EQ-5D-3L for example, has ratio properties which allows it to generate I-QALYs are also unacceptable. The corollary of this is that modelled simulated lifetime I-QALY or cost per I-QALY claims is also unacceptable.

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, which is 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. This applies to QoL claims developed as separate instruments for patients and caregivers.

A manufacturer in making a submission under the Minnesota guidelines framework should demonstrate that a systematic review has been undertaken of potential instruments that meet Rasch measurement standards, capturing patient centric and needs fulfillment criteria. As detailed in Section 1 of the guidelines, 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. But we still have a long way to go as the overwhelming majority of PRO instruments, both generic and disease specific fail the standards of fundamental measurement for interval properties to assess response to therapy.

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 or for the period over which the patient is compliant with therapy. In the context of an evidence platform the manufacturers should be in a position, as part of their claims assessment protocol, to propose how QoL claims are to be monitored and reported.

The guidelines summarize the requirements for an acceptable QoL claim:

- 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

Section 5: Claims and Value Assessment

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 effectiveness (e.g., incremental I-QALYs). The formulary committee should set 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).

DOI: https://doi.org/10.24926/iip.v11i4.3542

As noted, 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 and other value assessments agreed with the formulary committee.

Protocols should look to establishing a permanent evidence base, possibly a registry, to support well-designed observational studies. 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. One framework is to support claims assessment through value contracing.

Finally, the protocol should detail 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.

Claims are considered under six categories. These are:

- Clinical claims for therapy response
- Patient centric quality of life claims
- Supporting clinical claims for co-morbid conditions
- Product entry, uptake and discontinuation claims
- Claims for impact on medical resource utilization
- Societal impact claims

Two points should be noted. First, attempts to evaluate the extent to which co-morbid conditions, stage of a comorbid disease and the presence of polytherapy, modify claims for response to therapy are a critical part of product reviews in target populations. Second, the focus on elements of resource utilization puts to one side any attempts to make claims for expected costs. The formulary committee should be interested in the claims for resource utilization and whether resource sparing is anticipated which are relatively easy to track. The committee can then apply what it considers to be the appropriate unit costs to arrive at an estimate of overall costs.

Section 6: Checklist for a Formulary Submission

The Minnesota guidelines conclude with a proposed formulary submission request from the health system to the manufacturer and a checklist for formulary submissions to be completed by the manufacturer. 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. The key first step is for a manufacturer to 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. The key questions are:

- 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?

DOI: https://doi.org/10.24926/iip.v11i4.3542

Conclusions

The proposed guidelines for formulary submissions represent a distinct break with the past; a break characterized as a paradigm shift. It overturns some 30 years of creating approximate information; a denial of the standards of normal science; a denial that claims for products and devices should be credible, evaluable and replicable. The result has been a disaster of mega-proportions with the promotion of the mathematically impossible QALY (the I-QALY) to an untutored audience; an audience of academics and consultants, let alone health system decision makers, who have no understanding of the standards of fundamental measurement or, indeed, of the standards of normal science and the process of discovery of new facts recognized for the past 400 years. The result is almost 19,000 cost and QALY publications based on a PubMed search (October 2020). None apparently have shown any awareness of the impact of fundamental measurement standards on the I-QALY construct; it should never have been introduced. To base this number of publications over 30 or more years on a metric which is mathematically impossible staggers belief. It also casts serious doubt on the attempt by groups such as ICER to support their business case by unproven claims that multiattribute utility scales haves ratio properties, which they certainly do not. Unfortunately, an entrenched leadership in health technology assessment, holding to a dogmatic belief system, is not easily overturned. In addition, groups such as ICER will still face the substantive issue that they fail the demarcation test between science and pseudoscience; the failure to generate claims for products and devices that are credible, empirically evaluable and replicable.

Conflict of Interest: Dr. Langley is an Advisory Board member and consultant to the Patient Access and Affordability Project, a program of Patients Rising.

Academic Freedom and Responsibility Statement: Dr. Langley is writing on matters of public interest and not speaking for any institution.

References

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

[2] Neumann P, Willke R, Garrison L. A health economics approach to US value assessment frameworks – Introduction: An ISPOR Special Task Force Report (1). *Value Health*. 2018;21:119-123

[3] Langley PC. Sunlit uplands: the genius of the NICE reference case. *Inov Pharm*. 2016;7(2): No.12. https://pubs.lib.umn.edu/index.php/innovations/article/view/435/430

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

[5] Langley PC. Nonsense on Stilts – Part 1: The ICER 2020-20234 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348

[6] Drummond M, Sculpher M, Caxton K et al. Methods for the Economic Evaluation of Health Care Programmes (4th Ed.). New York: Oxford University Press, 2015

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

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

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

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

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

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

[13] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? *J Rehabil Med*. 2012;44:97-98

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

[15] 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

[16] 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

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

[18] 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

[19] McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes 1: The search for the Holy Grail. *J Med Econ*. 1019;22(6):516-522

[20] McKenna S, Heaney A, Wilburn J. Measurement of patient reported outcomes 2: Are current measures failing us? *J Med Econ*. 2019;22(6):S23-30

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

[22] 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.

[23] 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

[24] 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