PATIENTS RISING

PATIENT ACCESS AND AFFORDABILITY PROJECT

THE NEW START FORMULARY SUBMISSION AND EVALUATION GUIDELINES FOR VALUE CLAIMS WITH PHARMACEUTICAL PRODUCTS AND DEVICES

Version 2.0 May 2022

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

Executive Director: Terry Wilcox

700 12th St. NW, Suite 700
Washington, DC 20005
202-750-1186 | www.info@accessandaffordability.org

OVERVIEW

These NEW START formulary submission guidelines represent a new and significant development is proposing the evidence base required to support the evaluation of formulary submissions and value claims for pharmaceutical products and devices. These guidelines have been developed under the direction of Terry Wilcox, Executive Director, Patients Rising as part of the Patient Access and Affordability Project by Dr Paul C Langley, Adjunct Professor, College of Pharmacy, University of Minnesota. The focus is on claims for products that not only address the needs of patients, caregivers and providers but meet the standards for normal science, including fundamental measurement. They include pivotal clinical claims that are supported by Phase 2 and 3 clinical trials, but with the requirement that all claims are consistent with the axioms of fundamental measurement. As well, the guidelines recommemd the exclusion of any patient report outcome claim (PRO) that that not meet fundamental measurement standards. The guidelines reject the construction of imaginary modeled simulation value claims that have to date been the standard for formulary submissions, in favor of value claims that are credible evaluable and replicable in a meaningful time frame [i]. As well as supporting submissions for new products and devices, the standards set out are designed to support ongoing disease area and therapeutic reviews.

Simulated claims for cost-effectiveness with ersatz assumption driven recommendations for 'social' pricing and access by groups such as the Institute for Clinical and Economic Review (ICER) [ii] [iii] are an analytical dead end; imaginary non-evaluable claimns do not support a program to discover new therapy facts. The NEW START submission has to be in terms of protocol supported claims that meet the two basic premises of the NEW START package:

- **All value claims for a product or therapeutic intervention must refer to a single attribute that meets the demarcation standards for normal science: all value claims must be credible, evaluable and replicable**
- **All value claims must be consistent with the limitations imposed by the axioms of fundamental measurement: they must be unidimensional and meet interval or ratio measurement standards**

These premises apply to value claims that are disease or target patient population specific, where every claim is supported by a reporting and assessment protocol.

These claims may be for individual clinical attributes, for quality of life and for drug and resource utilization. The formulary committee is in the box seat to determine the relevance of claims for a target patient population and the process for factoring these into pricing and access recommendations. The key point is that claims assessment is an ongoing process where each claim is judged by its credibility, ability to be empirically evaluated and replicated across different treating environments.

Developing these guidelines and the foundation premises given above are also a response to what we may call the 'measurement crisis' in health technology assessment and value claims for response to therapy; the majority of scales and scores that have been developed and applied over the past 40 years, whether disease specific or generic, fail the axioms of fundamental measurement [iv]. The required instrument measurement properties for evaluating response to therapy must be dimensionally homogeneous, unidimensional and with adequate construct validity. The value claim must apply to a single attribute of therapy response defined by a coherent construct theory with demonstrated interval or ratio measurement properties.

Formulary decisions for pharmaceuticals and devices should not be based upon flights of fancy. Assumption driven lifetime modeled simulations, which defy scientific standards let alone elementary rules of logic, are no basis for pricing and access decisions which impact the quality of life of patients and caregivers. Groups such as ICER and the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) might see these simulations as acceptable shortcuts to create approximate information; their application of mathematically impossible QALYs is nothing more than sleight of hand. Health care decisions deserve better than imaginary and non-verifiable claims [v].

These NEW START guidelines are applicable both to formulary committees and other health care decision makers who may, for specific target patient population require disease specific value claims other than those based on pivotal clinical trials, as well as to manufacturers proposing evaluable value claims to distinguish their product from comparators in target patient populations. The formulary committee would only propose a value claim or claims appropriate to a target population; it would be up to the manufacturer to draft the protocol. The protocol structure proposed for value claim submissions can, of course, be a subject for negotiation between the parties.

An important caveat is that the manufacturer should not wait until the last minute to assemble a submission package. Pkanning for a NEW START submission should start no later than the OPhase 2 period of product developmet. This does no exlude preliminary discussion wirg heatrh systems, patient advocacy groups and professional groups over the appropriate value claims, anticipated evidence gaps, a prioiule f the target population and a commitment to PRO instruiments to capture the patient voice and needs fulfillment quality of life.

The NEW START guidelines envisage a two stage process in engaging with a manufacturer: a process which is applicable to new products approved by the FDA for a target patient population and also for ongoing disease area and therapeutic reviews. In the first stage for new products the health system would invite the manufacturer to make a submission within a nominated timeframe; a copy of the guidelines would be included, together with required value claims. In the second stage the formulary committee reviews the submission, determines whether it meets required standards and, if appropriate request a revised submission. At both stages, all claims would be required to have a protocol for specific claims or attribute evaluation. Claims based on imaginary simulations would be rejected.

NEW START TRAINING PROGRAM

To support the introduction of the NEW START Guidelines, Patients Rising has prepared a training package which addresses three key areas in health technology assessment. The areas addressed as part of this 14 modukle pakage are:

- **The Standards of Normal Science**
- **The Failkure of Approximate Information**
- **A NEW START in Health Technology Assessment**

Each module comprises slides/audio presentation supported by notes with referenxces. An Introductory module may be accessed through the www.Maimonresearch.com website at https://maimonresearch.com/training-videos . The training package has been authored and is presented by Dr. Paul C. Langley, Adjunct Professor, College of Pharmacy, University of Minnesota. The modul;es are password protected.

In addition, Patients Rising can support workshops and working groups to respond to issues, including implantation by manufacturers.

TABLE OF CONTENTS

TABLES

1. INTRODUCTION: INVENTED OR REAL WORLD EVIDENCE

The potential role of protocols to support the evaluation of value claims by formulary committees and other health system decision makers has been recognized for some 20 years [6] [7]. Protocols to support product development and the pivotal claims for clinical response are well established; what is missing is the application of these same standards to value claims outside of Phase 2 and Phase 3 trials. Unfortunately, within the context of hypothesis testing and evidence based medicine, this prospective role has been sidelined in favor of the creation of simulated cost-per-QALY models which yield imaginary non-evaluable value claims [8]. Leading this imaginary information crusade in the US has been the Institute for Clinical and Economic Review (ICER) who have taken upon themselves the mantle of the International Society of Pharmacoeconomics and Outcomes Research (ISPOR) in their acceptance of the dominant role of creating 'approximate information'; non-evaluable claims to support formulary decisions [9]. Put simply, ICER and ISPOR have opted for non-science [metaphysics and pseudoscience], the creation of imaginary cost-outcome claims that fail the demarcation test for the standards of normal science. The standards must be that any claim is credible, evaluable and replicable.

1.1 Value Claims and Real World Evidence

The foolishness of this decision to reject the standards of normal science, driven by an unwillingness to admit that with limited data for clinical end points and direct medical costs at product launch claims have to be speculative, has unfortunate consequences. The attraction of this easy option, rejecting the option of a program to meet evidence gaps in favor of the immediate satisfaction of imaginary assumption drive claims, is obvious.. Any imaginary simulation model is just one of possibly hundreds of models each producing by choice of assumption, and probably engineered to produce, a claim of your choice. To this should be added the logical problem of inductive inference: assumptions or observations taken from past cannot support claims for the future. We cannot justify, other than by belief, a psychological defense of 'the realism of assumptions' that a claim from the past can be justified to support a claim on the future. Hence the importance of replicating pivotal claims as all too few appear capable of replication, even with identical protocols. Building approximate information claims from past observations to support unobserved future claims through simulation modeling is nonsensical; even more so if the future imaginary claims are not, by intent, empirically evaluable.

Central to the belief in the creation of approximate information is the application of cost-utility or incremental cost-per-QALY modelling. Unfortunately, those such as ICER advocating the construction of lifetime cost-per-QALY models, estimates of future costs-per-QALY and then the application of cost-per-QALY thresholds overlook the standards required of fundamental measurement. The case is quite straightforward. To construct a QALY . a utility or preferene score is required which meets the standards of ratio measurement: a scale with a range of 0 – 1 where 0 = death and 1 = perfect health. The scale bounded at unity must have a lower bound of a true zero; there must be no score possible that takes negative values. Preference scores, such as the EQ-5D-5L do not have this property; they are just ordinal scales that can only support nonparametric statistics. A recent attempt to generate values for the EQ-5D-5L in the US found that 20% of the possible 3125 health states, combining symptoms and response levels, took negative values or health states worse than death. Giventhe ordinal scoring this means that the imaginary or I-QALY is a mathematically impossible construct. No cost-per-QALY claim has any validity; let alone serving as support for pricing and access recommendations. Claims from ICER simulation modelling should not play a role in formulary decision making for pricing and access.

The impossibility, indeed futility of the approximate imaginary information belief system puts formulary committees and other health decision makers in a quandary. Either they accept the ICER framework, admitting it is an analytical absurdity and dead end, or they abandon it and seek instead a framework for formulary decision making which rests on a program of individual claim evaluations to support and maintain formulary listing. If the latter is accepted, then decision makers need to establish a set of standards for proposing credible and evaluable claims; this is best achieved by ensuring manufacturers meet required evidence standards, including protocols for claims assessment.

Proposing a successor formulary submission package is the role of the NEW START Guidelines. The commitment must be to the standards of normal science and the axioms of fundamental measurement to support disease specific value claims. This means that claims for product cost-effectiveness are redundant.

1.2 Response to Therapy

The value assessment of any pharmaceutical product or device rests ultimately on accurate measurement; the response to therapy from baseline for the target patient population. If the instrument employed, whether a clinical measure or a patient reported outcome (PRO) measure fails to meet the axioms of fundamental measurement, then any claims that rest on that instrument are invalid. Following the formalization by Stevens and others in the 1930s and 1940s, scales used in statistical analyses are classified as nominal, ordinal, interval or ratio [7]. Each scale has one or more of the following properties: (i) identity where each value has a unique meaning (nominal scale); (ii) magnitude where values on the scale have an ordered relationship with each other but the distance between each is unknown (ordinal scale); (iii) invariance of comparison where scale units are equal in an ordered relationship with an arbitrary zero (interval scale) and (iv) a true zero (or a universal constant) where no value on the scale can take negative scores (ratio scale) and the scale has interval properties.

Nominal and ordinal scales do not support any arithmetic operations; only nonparametric statistics. Interval scales can support addition and subtraction while ratio scales support the additional operations of multiplication and division as they have a true zero. This zero point characteristic means it is meaningful to say the one object is twice as long as another. To measure any attribute on a ratio scale it has to be demonstrated that all criteria for an interval scale have been met with a true zero.

The focus of the NEW START guideline is on real world evidence; not on the construction of incremental cost-per-I-QALY or similar imaginary worlds which is an analytical dead end**.** If there is one overarching theme that drives these guidelines it is a belief in the need to recognize and accept the axioms of fundamental measurement; a standard that should be applied across the board to RCT based claims in product development and FDA approval as well as prospective RCTs or observational studies following market entry. Rejecting fundamental measurement, the unfortunate hallmark of the current belief system in health technology assessment, means that value assessment claims for pharmaceutical products and devices, including many protocols accepted by the FDA, are untenable. We have wasted 30 years in health technology assessment in chasing the imaginary claim will o'the wisp.

1.3 Measurement and Attributes

If we accept the undeniable fact that current standards, as proposed by ISPOR/ICER are untenable, then we need to set minimum evidentiary standard for value claims; standards that are consistent with those of the physical sciences. Value claims made by manufacturers must meet minimum standards, notably in

respect of the axioms of fundamental measurement in respect of both interval and ratio scales; a standard that is recognized and applied in the physical sciences and mature social sciences. The required standards represent a new and long overdue paradigm in technology assessment to replace the current belief system or meme (it is not a paradigm). NEW START represents, not an extension of the ISPOR/ICER imaginary world belief system with its commitment to approximate invested information, the creation of evidence from incremental cost-per-I-QALY lifetime or reference case modeled claims, but a complete rejection of that meme. 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 which respect the concerns of patients and caregivers.

Accepting the measurement standards of the physical science also means that any claim for therapy response must relate to single attributes. An attribute is a characteristic, quality or feature of response to a therapy intervention, often comparative, that can refer to a clinical end point, associated endpoints such as the incidence of specific side-effects, a patient reported outcome as endpoint, a quality of life claim or drug and resource utilization claim that has either interval or, preferably, ratio measurement properties. Value claims for products should be reported as single attributes with dimensional homogeneity, unidimensionality and construct validity, together with accepted psychometric properties. Combining attributes (e. g., symptom responses in a multivariate preference scale) fails to meet these requirements unless each of the attributes has ratio measurement properties.

Accepting claims for therapy response couched in terms of single attributes with ratio or interval measurement properties represents a paradigm that is 30 years overdue. The required minimum standards that a formulary committee should insist on are detailed in Table 1 below.

TABLE 1

REQUIRED MINIMUM STANDARDS FOR VALUE CLAIMS

- All value claims should meet the standards of normal science for credibility, evaluation and replication
- All value claims should meet standards set by the 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 each claim is to be evaluated and reported

---

1.4 Rejecting Cost-Effectiveness Claims

It is worth emphasizing that it is not the purpose to produce a single overarching imaginary ‘value’ claim for a therapy intervention; a composite claim that represents the clustering of single attributes. Rather, all claims must refer to single attributes for the target population within the disease area. There can be no composite overall ‘cost-per-incremental QALY’ claim or a claim that, in a comparative simulation, a

product is ‘cost-effective’ at a nominated unit price; this is absurd as it fails the standards of normal science for single attribute measurement. Claims for product ‘cost-effectiveness’ are redundant. Rather, proposed single attribute claims would be submitted to a formulary committee, where each is empirically evaluable. It is up to the formulary committee to factor these claims into a provisional agreement on pricing and access to pharmaceuticals.

A blanket cost-effectiveness claim for a proposed timeframe is, therefore, impossible. There are too many elements as single attributes that comprise the claim for it to be empirically evaluable. Certainly, we can put together empirically meaningful claims for various response to therapy metrics, from ratio scale clinical measures to bounded ratio scale claims for need. There is no universal metric that is acceptable; indeed, there could not even be agreement on the impossible or I-QALY measure, without appreciating it is a mathematically impossible construct with preference scores on ordinal scales. Nor is there any prospect of agreement on what is meant by ‘effectiveness’ at a generic, let alone a disease specific level for a target patient group. Claiming effectiveness in terms of cost-per-I-QALY thresholds is a non-starter. At best, a formulary committee may require disease specific metrics on either a ratio or interval scale, with a designated minimum clinical difference or ‘improvement’ as a guide. Even so, minimum clinical difference is an ambiguous concept and a metric that may fall short of what a formulary committee or a practicing physician might consider appropriate. The question becomes more tricky if a range of attributes for effectiveness are proposed to include both clinical endpoints (e.g., primary pivotal trial endpoints) and endpoints that are proposed to capture need-fulfillment quality of life. In all cases the metric chosen must have the required measurement properties. The formulary committee is open to requiring a range of possible metrics, none of which may be part of pivotal protocols. This applies also to under-powered secondary end-points from pivotal trials.

1.5 Disease Area and Therapeutic Class Reviews

The NEW START Guidelines are intended to meet the formulary requirements for new products as well as ongoing disease area and therapeutic class reviews. To this extent value claims should be considered provisional; subject to potential re-assessment for pricing and access as new data are presented. This is the purpose of requiring not only value claims to meet the standards of normal science but claims that are supported by protocols. These, in turn, are to be reviewed and agreed by the formulary committee and the manufacturer. This is in contrast to the imaginary evidence soft option where new data may lead (rarely) to simulation models being modified to create yet more imaginary claims. This seems a singularly fruitless exercise.

2. TARGET PATIENT POPULATION

Given the focus on credible and evaluable claims, the first step in a submission should be to define the target patient group in terms of the indication approved for the product. Value claims for new and competing products should be seen in the context of an unmet medical or social need, defined by attributes, in the target patient population. Claims made in a vacuum of information are meaningless. This points to the standard, common in the physical sciences, that any attribute claim must be credible, evaluable and replicable. All too often claims are made for clinical endpoints in pivotal trials that prove to be impossible to replicate. Indeed, we also find that two pivotal trials with the same protocol produce conflicting results or the data have to put through a sieve to find a subgroup for whom the claim can be justified, even if the claim is not powered for the subgroup.

If a specific diagnosis is the basis for identification of this group 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.

At the same time, the manufacturer has to address the issue of measurement and claims for response to therapy for the target patient population. This means detailing the characteristics of instruments used in clinical trials and observational studies for the proposed claims in the patient population. This will, undoubtedly, result in a substantial culling of instrument that not only fail to report on single attributes but fail to meet interval or ratio measurement standards. Ordinal measures such as multiattribute preferences should be rejected. Claims must be specific to single attributes.

2.1 Epidemiological and Social Profile

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 detailed in Table 2:

TABLE 2

PROFILE OF THE TARGET PATIENT GROUP

- **Data sources**: detail the data sources, codes and possible algorithms that are considered necessary to identify the target population, including appropriate references for the data elements below
- **Systematic Reviews**: detail the search terms for all systematic reviews together with the objecte(s) of ech review
- **Population Estimates**: provide estimated target population counts for the last 5 years detailing the data sources and potential sources of error

- **Incidence**: given population or 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 are 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

---

2.2 Unmet Medical and Social Need

The extent to which a new therapy contributes to meeting an 'unmet' medical need(s), defined by attributes, in a target population is a critical aspect of NEW START formulary assessment. As part of the formulary submission package, manufacturers should present a clear assessment of unmet needs in the target population group, which should include caregivers, family members as well as patients. This should be based on a demonstrated systematic literature review supported by a statement as to what the company perceives to be the unmet need(s) that their product or device is intended to meet. Clearly, other products may be promoted as meeting the same 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 and their apparent 'superiority'.

A persistent criticism in the creation of imaginary simulation models by groups such as ICER is that, in relying upon HRQoL or clinical symptom instruments, they overlook critical dimensions of patient and caregiver need in evaluating response to therapy within target patient groups. Apart from the fact that the accepted measures of response to therapy fail to meet the required measurement standards, there is a need to consider more appropriate instruments and recognize that the term 'need' can be open to a number of interpretations. Of particular interest here is the concept of need-fulfillment; as patients and caregivers are the ultimate beneficiaries of therapy intervention a fundamental question that should be addressed is the extent to which need is met in the target patient population and the contribution of the new product to further meeting the need. This brings in the question of the appropriate instruments for

evaluating need and unmet need, defined as a quality of life attribute, and the standards determined by Rasch Measurement Theory (RMT) or Modern Measurement Theory (MMT) to determining the question of what PRO instruments should be considered after rejecting those that fail to meet the required standards [10].

Other dimensions of need such as access to therapy, health insurance cover, should also be considered as complem,ents to the evaluation of need-fi;lfillmentr. Systematic literature reviews, including possible liaison with patient advocacy groups should be reported. .

2.3 Core Value Claims

Where a product is identified for an indication specific to a target patient group in a disease state, formulary committees and other health system decision makers should consider what they recognize as the core value claims for that patient group. These should be communicated to manufacturers not only ahead of any submission for a new product but also for existing products that may be scheduled for disease area and therapeutic class reviews. Establishing core value claims by disease state given existing therapeutic options and their impact. Effectively eliminates value claims that are non-evaluable as epitomized by the CHEERS 22 guidance. Journals may publish non-evaluable claims, as intended by CHEERS 22, but these will be of no interest to formulary committees. Of course, as detailed in this NEW START Guideline, as well as credible, evaluable and replicable claims, the value claims should be for single attributes, dimensionally homogeneous with construct validity and with the properties for interval or ratio measures. If a ratio measure is proposed then, for claims for quality of life as an example, it should have bounded ratio properties, capped at unity and with a true zero.

In order to assist formulary committees in identifying core value claims and presenting the list to a prospective submission by a manufacturer, the claims can be considered in terms of (i) purely clinical claims; (ii) patient reported outcome (PRO) claims; (iii) claims for drug utilization following formulary listing; and (iv) claims for resource utilization. In the case, for example, of clinical claims, the formulary committee may be interested in comparative (not placebo controlled) claims for therapy impact, rejecting indirect or modelled comparisons which fail to meet measurement standards for empirical evaluation.

PROs raise a number of critical issues, not least the fact that the overwhelming majority of both generic and disease specific PRO value claims fail to fundamental evidence standards. They are all ordinal scales. These are unacceptable as they fail to capture response to therapy due to the lack of invariance of comparisons and a true zero. A further issue is the mathematically impossible quality adjusted life year (QALY). Claims expressed as QALYs are only acceptable if they are created from instruments that allow ratio measurement properties and are empirically evaluable.

Even so, claims for quality of life as a measurable latent attribute ahould always be considered a core value claim. Expressed in needs-fulfillment terms the technques are available to develop disease or target patient population specific interval and, in ceetain circulstance, bounded ratio measures. Ifneed fulfillment is envisaged as a core value claim then this needs to be accommodated at no later that Phase 2 of product development as a unique instrument will have to be developed.

Finally, resource utilization is the preferred metric as it can be captured through existing database as ratio measures; attempting to introduce costs is of no interest to health systems as they all employ their own resource pricing algorithms and budget impact assessments.

3. CLINICAL EVIDENCE AND MEASUREMENT STANDARDS

If an endpoint, in both trials and observational studies, is to be judged an appropriate measure of response to therapy then the instrument must conform, as noted above, to the axioms of fundamental measurement. If an instrument does not meet these standards then the formulary committee may reasonably reject any claims based on this instrument. As detailed above there are only two measures that have the required properties for reporting therapy response for single attributes: interval measures (with an arbitrary zero) and ratio measures (with a true zero). Few patient reported outcomes (PRO) instruments reach these standards; this not surprising as few were designed to have these properties [11]. Given this, it is important that formulary committees are in the position of being able to reject instruments and claims that fail these standards; basing response claims on such instruments is clearly a waste of time. In some cases, if the primary powered endpoint of a pivotal trial is based on an instrument that fails ratio or interval measurement standards, then the trial should be rejected as providing potentially misleading claims information. This rejection would apply to any modeled claims that rest on those data.

3.1 Rejecting Claims for Clinical Response

Increasingly, over the past 20 years, concern has been expressed over the use in clinical trials of multiattribute composite scales to capture a mathematically impossible constructhealth related quality of life (HRQoL) comprising a bundle of symptoms and response levels. These include both surrogate and intermediate endpoints. All too often these include a mix of endpoints that may include both measures of safety and efficacy in the same composite scale. The term multiattribute encompasses both PRO instruments which ask respondents to value a 'basket' of attributes or description of a health state (e.g., standard gamble [SG], time trade-off [TTO]) as well as those instruments which, more formally, capture HRQoL through preference scoring a list of symptoms with ordinal response levels (e.g., EQ-5D-3L/5L). These fail the axioms of fundamental measurement, producing only ordinal scores. These cannot support claims for response to therapy. The rule should be that composite measures are impossible measures. If reported as either primary or secondary endpounts in clinical trials they should be ignored. This injunction applies to all generic instruments and disease specific instruments that are only capable of creating ordinal scores.

As well as reecting generic preference score, QALY claims based on randomized clinical treials (RCT) should also be discarded. The commonly used SF-36 patient reported outcome is an outstanding example. The confusion arises, in part, because of a failure to recognize the difference between a profile and an index. The SF-36 comprises eight health domains where seven of these domains are scored as a multidimensional composite scale. Generating profiles for these domains is implausible, creating an overall index by aggregating over the standardized scores for each domain is even less plausible. Composite summaries are used for physical (PCS) and mental (MCS) component summary scores; these are popular but problematic as they are not independent and produce mirror image scores with a high PCS correlating with a low MCS. Despite these concerns, there are any number of papers that have either reported on the PCS and MCS as independent score together with those that gave combined all health domain scores into a single value. None appreciate that these are ordinal measures.

Formulary committees, as presumably agents of patients and providers, should reject these composite multiattribute scales; they are not adequate measures of response to therapy. The only acceptable measures are those for single attributes which have demonstrated fundamental measurement properties. As a short cut to a single value claim, a 'composite' ordinal value measure is a failure. Invalid claims can lead to adverse intervention decisions.

3.2 Accepting Clinical Claims

The NEW START criteria for accepting a clinical outcome claim is quite straightforward: it needs to be demonstrated by the manufacturer submitting the claim that the outcome measure has dimensional homogeneity. It is only variables or attributes that have the same dimensionality that can, if required, be combined through either addition or multiplication; they must all have ratio properties. If not, they lack unidimensionality and construct validity. If this standard is not met, then the manufacturer will have to revisit pivotal trial claims and demonstrate that all outcome measures, as single attributes, meet the axioms of fundamental measurement; this points to the importance of replicating claims when pivotal protocols admit clinical outcome measures that lack dimensional homogeneity. There should be no assumption that an interval scale can be applied as if it had ratio properties or, more egregiously, the assumption held by groups such as ICER that the ordinal preferences of multiattribute instruments are actually ratio measures in disguise.

Each manufacturer should be asked to detail all clinical and PRO endpoints utilized in clinical trials (typically Phase 3) in terms of the criteria detailed in Table 3. If the manufacturer cannot demonstrate that the response in a trial protocol meets standards for ratio or interval dimensional homogeneity properties, then claims should be rejected. This applies to comparator claims for the product as well as attempts to pool claims across clinical trials. Pooling should only be recognized if each trial included meets the required evidence standards. 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. Clinical value claims, whether they are based on pivotal phase 3 clinical or on meta-analyses, must meet the required axioms of fundamental measurement: all clinical claims must be formulated to meet either interval or ratio measurement standards. Value claims that are based on instruments that meet only ordinal properties should be rejected. This standard is seldom applied in clinical reviews, notably for patient reported outcome (PRO) measures, whether these are generic or disease specific.

TABLE 3

MEASUREMENT STANDARDS FOR CLINICAL TRIAL ENDPOINTS

For each clinical or PRO response endpoint in submitted clinical trial data detail:

- Instrument title, response parameter, source reference, trial NDC code
- Status as primary or secondary end point in the clinical trial
- Deconstruction of the instrument(s) in terms of: attribute(s), dimensional homogeneity, unidimensionality and construct validity
- Acceptance or rejection of claims based on the trial
- Whether the manufacturer has submitted a protocol for replication of an outcome claim that meets required measurement standards

---

3.3 Internal and External Validity of Trial Based Value Claims

An oft voiced criticism of claims created from pivotal clinical trials is that the application of inclusion and exclusion criteria (e.g., age, gender, comorbidities) effectively limits any claim to the subset of the target patient population that meets these criteria. This creates issues where a formulary committee may ask for the claims to be evaluated in the target population in what may be described as a real world treating environment.

These standards for the validity of a trial protocol apply equally to meta-analyses that attempt to make 'weighted' comparative claims for product response. All trials that are included in the meta-analysis must meet these standards; there cannot be combinations of trials that include ones that fail these standards. This means that if a value claim is presented in comparative clinical terms the formulary submission must be quite clear that the comparative claim is based only upon trials that meet interval or ratio response properties that are acceptable to the formulary committee.

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. Standards for accepting RCT claims for response to therapy are detailed in Table 4.

TABLE 4

ACCEPTANCE OF RANDOMIZED CLINICAL TRIAL CLAIMS

- 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
  - Inclusion and exclusion criteria
  - Status (completed, ongoing, proposed)

- Rejection of Claims
  - Reject all RCTs where the primary outcome fails to demonstrate either interval or ratio properties
  - Flag all RCTs which meet primary outcome requirements but report secondary outcomes that fail these standards
  - Reject all comparative product claims where any trial included in the meta-analysis fails to meet required measurement standards for the primary outcome

- Replication

- Assess the likelihood that each of the accepted 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

- External Validity
  Match the inclusion/exclusion criteria for subject selection in each accepted RCT to the
  - demographics (age, gender, ethnicity, race)
  - socioeconomic status
  - stage of disease
  - comorbidity profile

  of the target patient population (as described in Section 2)

- Estimate the proportion of the target patient population at (i) the national and (ii) the health system level that the individual trial protocols have captured in terms of inclusion/exclusion criteria for each RCT listed

---

3.4 Clinical Claims Hierarchy

Claims for the efficacy or effectiveness of value 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 [12].

All that is missing ub GRADE is an appreciation of fundamental measurement; this is a major oversight which, as noted above, provides grounds for the rejection of both individual studies as well as meta-analyses (including any GRADE claims). This is a key qualification to any attempt to argue for an evidence hierarchy. A well conducted observational study, for example, which met the required measurement standards would clearly outrank an RCT which failed to meet these standards and, as a result, was rejected. It is of interest to note that ICER in its assessment of the 'quality' of evidence from pivotal trials completely overlooks the requirement to meet interval or ratio measurement standards in assessing response to therapy from RCTs.

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. To this should be added fundamental measurement standards. 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 [13].

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 (AHRQ) *Methods Guide for Comparative Effectiveness Research* to support the Evidence-base Practice Center (EPC) Program [14] . 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 (but not the axioms of fundamental evidence). 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 EPC framework must be qualified by the application of the axioms of fundamental measurement. PRO measures that fail to meet these standards should be rejected. While effect size may be claimed, the absence of response defined by interval or ratio measures invalidates such value claims.

3.5 Value Claims Evidence Base

In the first instance, it is at the discretion of the manufacturer to suggest the recommended parameters for clinical claims assessment protocols for the target patient population and the required data set or evidence platform to support value claims assessment. This could be achieved through an observational study tracking patients over an agreed timeframe where the therapy response is evaluated through econometric modelling. The evidence platform for claims assessment would be, in effect, a patient registry.

For a manufacturer to support a registry as the preferred evidence base has a number of advantages. It avoids high cost one-off RCTs, it has a built in flexibility to accommodate changing endpoints and to ensure that measurement standards are met. Most importantly, a registry should allow a detailed assessment of the characteristics of the target patient population that might qualify pivotal RCT value claims, including disallowed claims based upon generic multiattribute instruments. If there are observed well defined sub-populations within the target population, then comparative clinical value claims could be proposed for more targeted interventions and clinical guideline development.

3.6 Posting Protocols

With due regard to commercially confidential information, the manufacturer should be invited to post their agreed clinical value claims protocols on the US National Library of Medicine's https:// clinicaltrials.gov website. These should provide sufficient detail for the protocols to be replicated across target patient groups in health care systems.

3.7 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 required measurement standards. Future value claims rejection should be anticipated.

4 PATIENT NEED AND QUALITY OF LIFE

Technology assessment in healthcare has long promoted quality of life (QoL) as the gold standard in value assessment; as the critical value dominating health care and health care decision making [15]. To the World Health Organisation (WHO), QoL is defined as *an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns*. This definition is somewhat removed from the concept of HRQoL found in the health technology assessment literature. Taking a societal rather than an individual perspective, QoL is transformed to health related quality of life (HRQoL) defined in terms of a multiattribute 'easy to use' list of symptoms and response levels, valued by a sample of the general population, which are intended to generate preference on a 0 to 1 scale where 0 = death and 1 = perfect health (which given the limited symptoms is a misnomer). The voice of the patient (or caregiver) is absent; subsumed under clinical parameters that may have little relevance to the disease area and the needs of its members.

4.1 Quality of Life, Preferences and Eugenics

Unfortunately, attempts to create the required preference scale for HRQoL claims (which is, by poor design, ordinal but assumed to have ratio properties) failed as all direct (standard gamble, time trade-off) and indirect preference instruments (EQ-5D-3l/5L, HUI Mk2/3, SF-6D) generate negative preferences (or states worse than death). These societal valuations of health states, with unfortunate connotations to the eugenics literature, lead arguments for the potential denial of therapy access for those individuals deemed to be an unreasonable burden on the health care system, defined by membership of a pre-determined health state,. Apart from the failure of the QALY to meet fundamental measurement requirements due to preference scores having only ordinal properties there is then a darker side with preference scores valuing the 'worth' of health states; the worth of individuals occupying those health states [16]. The EQ-5D-5L, for example, in its application for US preference scores, yields 3,125 health states; of these 624 (20%) are negative scores or 'states worse than death' [17]. The fact that both direct and indirect multiattribute preference algorithms produce negative scores has been noted since the 1980s but put conveniently to one side when creating 'average' preference scores for the patient population. This can hardly be avoided in the case of the EQ-5D-5L as inputs to simulated ICER imaginary value claims.

If health care resources are considered fixed, then authorities may consider states worse than death, and the associated negative preferences and hence negative QALYs as justification for the denial of, or limiting access to, patients or caregivers in unworthy health states. This societal preference is expressed in purely clinical terms which ignores the fact that, for the individual patient, caregiver and family, clinical performance as defined by a treating physician may have no relation to life itself and the need of these respective players; in other words life quality where the need of the patient is a critical input.

Community preferences (or those of a political party) for worthy versus unworthy health states has a long and disturbing tradition in the advocacy of eugenic criteria; the denial of health care (or worse) for those considered to be unworthy; the taint of eugenics hangs over the application of community preference to create QALYs to judge the 'worth' of interventions, all too often expressed in cost-per-QALY terms; the introduction of eugenic principles by the back door. Irrespective of the technical limitations on the QALY construct, the fact that it is mathematically impossible, the fact that health state preferences rank the 'worth' of health states are sufficient grounds for its rejection. If we wish to make value claims for quality of life, then societal preferences must not enter the calculus.

4.2 Abandoning Ordinal Scores

Over the past 40 years the focus of health technology assessment has been on the construction of a gold standard single score to drive resource allocation in health care, with particular reference to pricing of pharmaceutical products and devices. Fortunately, the quest for the 'holy grail' has proved to be an analytical dead end. In the US, ICER has built its business case, if not its reputation, on simulation, assumption driven imaginary modeling of incremental cost-per-QALY claims matched to cost-per-QALY thresholds to support pricing and access recommendation. This is a barren exercise with the promotion of the imaginary analytical framework neglecting the standards for normal science, including fundamental measurement. The quality adjusted life year (QALY) suffers from a fatal defect: it is a mathematically impossible construct. Where preferences are ordinal these cannot be used to multiply time spent in a disease state (as modelled by ICER) to create a QALY. The exercise is a waste of time.

Formulary committees face two challenges: first, justifying their rejection of QALY based claims and, second, endorsing an alternative measure of quality of life that meets the standards of fundamental measurement yet reflects the need of the patient, the caregiver (if relevant) and the family. The first challenge is easily disposed of as the societal preference values or utilities, with their taint of eugenics, that are applied to create QALY measures are only generic ordinal scales. The second is more difficult as it requires an understanding of modern measurement theory (MMT) or Rasch Measurement Theory (RMT). As MMT/RMT are the required standard, then we have to put aside all generic multiattribute instruments and the overwhelming majority of disease specific instruments; they only produce ordinal scores. We have, in fact, a vacuum in PROs to evaluate health status and response to therapy; an unenviable position for HTA to find itself in.

4.3 The Imaginary QALY (I-QALY)

To defend the construction of a QALY, the advocate has to make the case that the preference score, whether in terms of values or utilities, on a bounded scale from 0 = death to 1 = perfect health, has ratio measurement properties; a bounded ratio scale with a true zero and capped at unity (supporting an interval scale). This is an impossible proof because all preference instruments can produce negative scores; states worse than death. There is no true zero; the choice of zero is arbitrary with the only possible claim (which is again impossible) that the preference score has interval properties. In fact, as noted, all preference instruments produce only ordinal scores. This follows not only from their lack of dimensional homogeneity, unidimensionality and construct validity as multiattribute instruments but the apparent lack of concern by their developers in the 1970s and 1980s of the standards of fundamental measurement and the need for a QALY to be created from a bounded ratio scale. The false assumptions by generations of analysts, numbering in the thousands, is that the pereference score has (i) intvariance of comparisons and (ii) a bounded ratio scale.

If the manufacturer persists in the claim that the preference score has a true zero (as ICER has attempted over the years) there must be no circumstance under which the preference algorithm can create negative values (e.g., for different target patient populations in all disease groups). To this would be added a further proof that while the instrument, a generic multiattribute creation, covers a number of dimensionally and nominally independent attributes (pain, depression, mobility, etc.) the result is not a dimensionally homogeneous bounded scale, which is unidimensional and with accepted construct validity. Unfortunately, even if this were achievable, there is still the problem of non-evaluable ICER assumption driven incremental cost-per-QALY claims and the application of I-QALY thresholds. There is no evidence that even these individual symptoms or attributes have the required measurement standards.

If persistence requires proof, then the questions a manufacturer would have to address to satisfy the formulary committee are detailed in Table 5.

TABLE 5

DEFENDING A 'CLAIMED' RATIO SCALE PREFERENCE SCORE

- Can the manufacturer provide a proof that, under all circumstances, the proposed preference score is a bounded ratio scale in range 0 to 1 with interval properties?
- Can the manufacturer provide an explanation as to why the selected instrument has created negative scores or preferences for states worse than death?
- Can the manufacturer explain why this particular instrument was selected to create a claimed ratio scale QALY?
- Can the manufacturer provide a systematic review of the preference scores (including the range from negative scores to unity) created by the different preference instruments for the target patient group?
- Can the manufacturer identify any proofs that the other preference instruments that have been applied to the target patient group yield bounded ratio scales?

4.4 Evaluating a PRO Claim

In terms of the general acceptability of a PRO onstrument and value claims it produce, there are a number of general issues that should be addressed, as questions to the manufacturer, as a first pass, with the emphasis in the first instance on the patient voice and its acceptability to patients in disease areas. These are listed in Table 6

TABLE 6

INSTRUMENT ACCEPTABILITY (GENERAL QUESTIONS)

General Questions	Response (YES/NO)	Comments
Is the instrument acceptable to patients?		
Is the instrument unidimensional, reporting on only one attribute?		
Is the instrument, in measurement terms, reliable?		
Is the instrument responsive to real change?		

Is there evidence that the instrument has construct validity? • Is the instrument measuring what is intended? • Is the instruments conceptual model appropriate?		

It is important, when a PRO instrument is presented to ask the manufacturer to respond to more in-depth questions on why and how the instrument was developed. As will be noted from the questions below, the development is dictated by Rasch standards; this is expected as it is only RMT that can meet fundamental measurement standards.. These are listed in Table 7

TABLE 7

INSTRUMENT DEVELOPMENT FILTER

Development Questions	Response
What is this PRO instrument intended to measure?	
What does it measure?	
Is it patient or clinician or patient centric?	
How were the items in the instrument selected?	
Are the items specific to the disease or are they generic?	
Has the instrument been tested with relevant respondents?	
Has the instruments reproducibility been tested?	
How strong is the evidence for construct validity?	
Does the Instrument measure at the ordinal, interval or ratio level?	
Were modern measurement techniques applied, preferably Modern Measurement Theory (MMT) /Rasch Measurement Theory (RMT)	
Was evidence for internal validity reported?	

Was the effectiveness of the response format reported?	
Did the authors report item fit?	
Was local item dependency reported?	
Was differential item functioning reported?	
Did the authors report overall fit to the Rasch model?	
What assessment of responsiveness was reported?	

Few disease specific PRO instruments could pass muster on the criteria set out above. This means rejecting literally hundreds of instruments. Although, in the case of the patient voice in need fulfillment, fewer instruments for each disease state and target patient population (to include separate instruments for patients and caregivers) but all meeting the required MMT/RMT standards. The commitment to the EQ-HWB is an aberration..

4.5 The Bounded Ratio Needs Fulfillment Measure (N-QOL)

For those devoted to a therapy response defined in terms of comparative preference or QoL gains, there is one avenue open to creating a measure that meets the standards of normal science and fundamental measurement. This is the recently developed disease specific measure, called the Need or N-QOL scale, a disease or target patient population specific bounded ratio scale from 0 = no needs are met to 1 = all needs are met (for all questionnaire need items are responded to affirmatively for binary responses). This measure stands in contrast to societal preference measures: it is a measure of the need of patients (and caregivers) not of the burden of disease, although this is implicit in the items identified to construct the disease specific instrument.

Constructing the N-QOL requires two steps: first, the construction of a disease specific scale applying the techniques of Rasch Measurement Theory (RMT) to create a count of responses for items ranked by their difficulty of meeting patient (and caregiver needs) and the ability of respondents to meet those needs. Second, the item counts are transformed to a an interval scale and then a bounded ratio scale which meets required measurement standards [18] . The critical role of RMT is that it is focused on the patient (and caregiver) in analyzing categorical data where the likelihood of a positive response is a function of the trade-off between item (need) difficulty and the respondent's abilities or proficiency in meeting that need as locations on a continuous latent variable. The objective of RMT is to develop a probabilistic index of response from ordinal scales that has interval level properties. In certain instances (as is the present case) this can be translated to a bounded ratio scale.

In traditional test theory (TST) and item response theory (IRT) the observed data have primacy; results are exploratory and descriptive of those data. Rasch models are, on the other hand, confirmatory and predictive; a confirmatory model requires the data to fit the model where, following the principles of conjoint measurement, they are sufficiently realized to claim the results are a measurement scale with interval measurement properties detecting measurement structures in non-physical attributes.

The N-QOL yields an estimate of the extent to which patient need (or in a separate questionnaire caregiver need) is met. A therapy response is defined as the improvement over baseline of the proportion of needs being met. As this is a ratio scale, the bounded N-QOL score can be applied to time spent in a disease state to create N-QALYs. This is not a preference scale, so should not be compared to value or utility scales. It is a measure from the patent perspective in a target disease state of the extent to which need as defined by the target patient group, is met. N-QAL is interpreted, therefore, as the equivalent time spent in a disease state or stage of disease when all needs, as defined by the disease specific N-QOL instrument are hypothetically met. The N-QOL measures is to be preferred for evaluating need and response to therapy.

The N-QOL captures a single latent attribute: patient or caregiver need. It avoids the multiattribute approach by avoiding clinical criteria, which may be of more interest to the clinician than to the patient or caregiver. It is patient-centric within disease states, focusing on the patient as the ultimate beneficiary of therapy interventions. This is not new; some 30 instruments have been developed over the last 25 years, covering the major disease states. These instruments are detailed at www.Galen-Research.com and include the following: ***pulmonary hypertension, Alzheimer's disease spousal caregivers, atopic dermatitis, psoriasis (patients and parental need), growth hormone deficiency, Crohn's disease, plexiform neurofibromas, herpes, migraine, multiple sclerosis, depression, asthma, COPD, ankylosing spondylitis, psoriatic arthritis, rheumatoid arthritis, systemic lupus erythematosus and incontinence and urogenital atrophy.***

4.6 Disease Specific Likert Instruments

Rasch analysis is not restricted to dichotomous item responses; it can also be applied to polytomous data (Likert scales). Again, this has been a feature of RMT since the 1970s, with a number of software packages designed to transform these data into measurement scales that meet required interval standards..

The application of Rasch analysis is self-evident; to avoid the adding up problem in Likert scales where the integer value responses are just summed to give an aggregate score. This is incorrect because to 'add' these scores you require (i) an interval scale for each Likert scale which is identical across all Likert scales in the instrument and (ii) that each Likert question is equally 'difficult'. While users try to hedge around (i), they fail typically to appreciate (ii).

Rasch analysis can be applied to Likert instruments to yield a single rating structure common to all items on the scale: The Rasch Rating Scale Model. This model provides person estimates for ability and a difficulty threshold for each Likert item but provides a set of rating scale thresholds common to all of the items. Hence, for 5 integers there will be four thresholds (not those assigned as integers). As an example: with four non-equidistant thresholds (A, B, C, D) and (say) 5 responses UU, VV, XX, YY, ZZ: UU to XX over thresholds A and B is not the same as moving from VV to YY over thresholds B and C. Rasch will assign respondents to a threshold defined category where there are interval properties.

On the presumption that, as far as measurement is concerned, HTA will never change, mention should be made of the ongoing EQ-HWB (Health and Wellbeing) program. This is a Likert based instrument package which replicates all the faults we have noted for Likert scales; it is presumed that the authors not only believe the scales have interval properties but allow the adding-up of integer scores. Once finalized, it not recommeded and will, in fact, fail to meet required measurement standards.

4.7 Limited Choice of Patient Reported Outcomes

PROs as the basis for claims for response to therapy are restricted to those that meet RMT/MMT standards. This narrows the field significantly with the majority, if not all in some disease area or for target patient populations. The mandated standard for single attribute unidimensional instruments, linked to a defined latent construct, is strict; but there is no alternative. This does not eliminate Likert scales, but these have to meet Rasch rating scale model measurement standards; to date no PROs have been developed. This leaves Rasch models with bibary responses the only contender. The techniques are known and the latent trait or attribute of the patient or caregiver voice expressed in needs-fulfillment quality of life terms is the obvious way forward.

5 DRUG AND RESOURCE UTILIZATION

Alone among the various disciplines comprising the physical and social sciences, health technology assessment has, as noted above, the dubious distinction of creating imaginary claims for pricing and product access. These rest in part on notional claims for drug utilization, resource utilization and what are effectively guesses as to unit costs and budget impact for target patient populations. Attempting to make claims in terms of anticipated cost and budget is a waste of time, not least because of the question of formulating and monitoring cost and budget impact claims. This is best left to the health system. From the formulary committee's perspective, a more useful approach is to propose evaluable claims for (i) drug utilization and (ii) resource utilization. Expressed in unit rather than cost terms these claims are easily monitored from existing data bases and reported on a regular basis in real time. Claims for drug and resource utilization should be consistent. If, for example, there is an unanticipated increase in switching to the new therapy then resource utilization estimates should be adjusted accordingly, with claims for drug and resource utilization use adjusted accordingly. All claims should be for the health system and, if off-label, not just for the target patient population but for the target population and off label uses separately identified.

5.1 Drug Utilization

Claims for drug utilization following approval and market entry involve a number of elements that can be tracked from data bases. It is the responsibility of the manufacturer to proposed how these elements can be identified (e.g., through drug and procedure codes) and reported, possibly on a quarterly basis. This type of exercise will have been undertaken, at least in part, by the manufacturer to support product development. In practical terms it is unreasonable and unlikely that formulary committees would require manufacturers to provide detailed estimates of direct medical costs. What would be required are claims from manufacturers as to their estimates of the rate of uptake of a new product, determined in large part by the effort they are proposing to put in for marketing and sales in the required timeframe, together with estimates of the rate at which comparator therapies are displaced. A key element in this assessment would be estimates of product discontinuation and consequent uptake of other therapies. The advantage of focusing in the first instance ondrug utilization and switching is that these claims are relatively easily monitored and the claims for these components of drug utilization evaluated.

It is proposed that the first claims would be monthly for the 12 months following product launch, with subsequent revisions and a six-month rolling forward claim profile.

TABLE 8

DRUG UTILIZATION CLAIMS

The elements to report as drug utilization claims are:

- Monthly uptake of the manufacturer's product (by formulation and dosage) for new patients
- Monthly discontinuation of the manufacturer's product (by formulation and dosage)
- Profile of time to product discontinuation by month (by formulation and dosage)
- Estimate of 'off-label' monthly utilization of manufacturer's product following product launch (by formulation and dosage)

- Comparator drug utilization following new product launch (by formulation and dosage)
- Estimate of discontinuation of comparator drugs and switching to new drug (by formulation and dosage)
- Overall utilization of new and comparator products by month following new product launch (by formulation and dosage)

Drug utilization projections could also serve as the basis for further claims for specific units of resource utilization if requested by the formulary committee. A formulary committee would be interested in drug utilization adjunct resource use requirements such as new equipment, regular scans and more frequent physician visits. If a new product is more resource use intensive than this has to evaluated by the committee in terms of its own cost structure; not a blanket guess in a submission to justify use of the new drug with an ersatz claim for cost-effectiveness. In short, it should be clear by now that, from the perspective of a formulary committee claims that a product is 'cost-effective' in any empirical sense are meaningless, not least because of the impossible direct medical cost component.

5.2 Resource Utilization

Once direct medical costs are considered the issue becomes more complex. Certainly, the measures that comprise the cost calculus would typically have ratio properties as the measures are in units with a true zero so the question of a metric does not arise. The devil is in the details with attributes for claims assessment to be defined at the unit level for each cost element. This involves making a claim for how the distribution of comparator therapies for the target patient population will change, for the timeframe selected, following the entry and uptake of a new therapy. This sets the stage for claims for resource redistribution, where resources used to support a new therapy, with due account for rate of uptake, compliance and discontinuation of the new therapy and comparator therapies. Separate estimates to support claims for drug utilization for all comparator therapies will be required. These would then be translated into resource utilization defined by a resource unit classification (current procedure (CPT) codes) with a claim specific to each CPT code. Relying on broad based classifications for numbers of physician visits (including specialist visits), urgent care visits, hospitalization days and drug utilization will not cut the mustard. It must be granular so that it can be monitored to establish the validity of a claim. It is only then that claims can be considered for direct medical costs by the formulary committee for the new therapy and comparator therapies in an agreed timeframe.

If a manufacturer wishes to establish a claim for the resource sparing possibilities of a new therapy compared to current therapies (fewer and shorter physician visits; fewer urgent care visits; fewer hospitalizations and hospital days; less use of supporting diagnostics and devices) then the claim(s) will have to presented in terms of the specific resource units together with the protocol identifying the required data sources to support the individual claims to track patients experience of these following product launch. It is not intended that resource utilization tracking will be required for all patients. As noted above the protocol may be best viewed as capturing a sample of the target population, although this may have to revised if, for example, off-label use is occurring.

As a standard for the resource utilization protocols, defined for each resource unit, it is proposed that the claims should extends for 2 years following product launch, reporting for the target patient group as the prodct's reach extends on a monthly basis. This can be contracted for a longer period with the results reported on a rolling forward basis.

TABLE 9

RESOURCE UTILIZATION CLAIMS

The elements to report as resource utilization claims are:

- Identification and agreement with formulary committee or health system on the specific resource units to track for individual value claims for product
- Identification and agreement with formulary committee or health system on the specific resource units to track for individual value claims for comparator products
- Agreement on algorithms (search terms) to identify and track resource units
- Monthly uptake of the specific resource units for the product
- Monthly discontinuation of specific resource units for comparator products
- Quarterly reports of resource utilization for product and comparators by resource unit

6 DISEASE AREA AND THERAPEUTIC CLASS REVIEWS

A commitment to an ongoing program of disease area and therapeutic class reviews should be an integral part of formulary management. This is not undertaken lightly given number of drugs typically captured in a formulary; so it should be targeted.

When a manufacturer makes a formulary submission, the submission should include a commitment to supporting disease area and therapeutic class reviews as part of the protocols submitted to support value claims. These claims are not one-off but should be seen as an ongoing commitment by the manufacturer to report on the results of an initial protocol value assessment but also to provide ongoing evaluations for value claims to maintain the place of the product on therapy.

These guidelines are designed to support ongoing value claims as well as those for new products being considered for formulary listing. Feedback from ongoing value assessments can support negotiations for improved pricing and product access as well as to judge the contribution of new products entering the formulary in competition with existing products that have made their own value claims, submitting the results to the formulary committee.

REFERENCES

[i] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: 2 approved]. *F1000Research* 2022, 11:248

[ii] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1):No. 12

[iii] Langley P. Peter Rabbit is a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(2): No 22

[iv] Langley P, McKenna S. Fundamental Measurement and Quality Adjusted Life Years. *ValueHealth*. 2021;24(3):461[letter]

[v] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7

[6] Schommer JC, Carlson AM, Rhee TG. Validating pharmaceutical product claims: questions a formulary committee should ask. *J Med Econ*. 2015;1-7

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

[8] 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

[9] Neumann PJ, Willke R, Garrison LP. A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *Value Health*. 2018;21:119-123

[10] Bond T, Fox C. App0lying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd Ed.).. New York: Routledge, 2015

[11] McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting patient reported outcome instruments and interpreting the data they produce. *InovPharm*. 2021; 12(2): No. 17

[12] 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

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

[14] 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.

[15] Koch T. Life quality vs the 'quality of life': assumptions underlying prospective quality of life instruments in health care planning. *Soc Sci Med*. 2000;51:419-27

[16] Langley P. Abandoning Eugenics and the QALY. *InovPharm*. 2021;12(3):No.20

[17] Pickard A, Law E, Jiang R et al. United States valuation of EQ-5D-5L health states using an international protocol. *ValueHealth*, 2019;22(8):93141

[18] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6