MAIMON WORKING PAPERS No. 13 MAY 2022

MEASUREMENT AND MULTIPLE SCLEROSIS: REJECTING IMAGINARY PATIENT REPORTED OUTCOME VALUE CLAIMS

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

Abstract

Value claims for existing and new products in multiple sclerosis must meet the standards of normal science and the axioms of fundamental evidence. All claims made for product efficacy or effectiveness must be empirically evaluable and replicable, supported by evaluation protocols. In the case of multiple sclerosis (MS) as in all other disease states, value claims that are not defined for single, unidimensional attributes must be rejected if the standards of normal science are not met. This may appear a harsh judgement, but if there is a commitment to a research program to monitor effectiveness and discover new facts for therapy impact then assumption driven simulations that invent non-evaluable claims must be abandoned. They are unacceptable as they make clear the failure to meet the demarcation criteria between science and non-science or pseudoscience. Unfortunately, this belief system supporting imaginary value claims is still firmly in pace as demonstrated by the recent CHEERS 22 guidance. The purpose of this commentary is to build on a recent one in F1000 Research that pointed to illegitimate nature of the current belief system in health technology assessment, with CHEERS 22 as the reference, with the recommendation that it be overturned to accommodate a new paradigm, the New Start, to support health system decision making. The focus here is on the proposed evaluation by the Institute for Clinical and Economic Review (ICER) of the effectiveness and value of treatments for relapsing MS. It will be shown that this, again, is an irrelevant and misguided undertaking which fails normal science; with New Start offering a form basis for value claims, their evaluation and application in decision making. Applying the criteria of modern measurement theory or Rasch measurement theory we come to the, possibly to support claims for quality of life and symptoms in MS unpalatable conclusion, that there are only two patient reported outcome measures in MS that meet the required measurement standards.

INTRODUCTION

A recent commentary on the CHEERS 22 guidance for constructing assumption driven imaginary cost-effectiveness claims, demonstrated that the current health technology assessment (HTA) belief system or meme for creating approximate information failed to meet the standards of

normal science and fundamental measurement [1] [2]. The commentary pointed to manifest deficiencies, many of them in their own terms fatal, which ensured the overall failure of these modeled claims to meet the demarcation standard distinguishing science from non-science; or metaphysics and pseudoscience [3].

The F1000 commentary proposed abandoning the approximate information belief system in HTA with a NEW START in technology appraisal to support formulary and health system decisions for new and existing pharmaceutical products and devices. It was argued that the NEW START represents the adoption of a paradigm in health decision making that met single attribute value claims, supported by evaluation protocols, were an essential input to health care decision making.

The purpose of this follow-up commentary is to illustrate the failure of the current HTA belief system with a proposed multiple sclerosis (MS) value claim therapy assessments as a case study. The case study is the recently released proposal by the Institute for Clinical and Economic Review (ICER) in the US, to undertake a modeled assessment of the effectiveness and value for a range of treatments in relapsing forms of MS, including monoclonal antibody interventions and oral therapies [4]. The case presented is that the proposed ICER assessment framework is irrelevant to the needs of those concerned with comparative HTA value claims in MS; with particular reference to the misapplication of the notion of health related quality of life (HRQoL) supported by the focus on the mathematically impossible QALY to drive simulation models.

THE NEW START VALUE FRAMEWORK

The proposed NEW START paradigm value claim framework rejects approximate information assumption driven models that extend into an unknown future with no hope of ever meeting required evidentiary standards for empirical evaluation

The NEW START paradigm for value claims embraces two premises or principles:

- **All value claims must refer to single attributes and meet the demarcation standards for normal science**
- **All value claims must be consistent with the standards for fundamental measurement, meeting ratio or interval properties**

Modeling value claims must meet the standards of normal science. If a model is proposed to support a value claim, with associated assumptions, then the claims must be credible, evaluable and replicable, subject to being accepted by a formulary committee for reporting results in a meaningful timeframe. All value claims must be supported by a protocol detailing how they are to be evaluated. If, in the case of ICER, all value claims, typically those involving QALYs and costs, are not empirically evaluable, then these standards are absent together without any notion of protocols to support value claim assessment.

Rather, with new start and the application of modern measurement theory (MMT) or Rasch Measurement Theory (RMT), it is possible to capture the patient voice in a single attribute quality of life (QoL) measure and the further application to the experience of MS patients in respect of MS symptoms, again as a single attribute measure. This puts assumption driven simulation models and claims for cost-effectiveness to one side in favor of disease specific value claims focused on clinical endpoints, MS specific instruments and claims for drug and resource utilization; all driven by evaluation protocols

THE PROPOSED ICER ASSESSMENT

The ICER proposed assessment is in two parts. First, an evidence review, primarily a clinical evidence review, using the PICO(TS) framework, extending to randomized clinical trials, high-quality systematic reviews, high quality comparative cohort studies and input from patient advocates, regulatory documents, information submitted by manufacturers and the 'grey' literature. Second, there will be a modeled comparative value analysis, driven by assumptions to support non-empirically evaluable imaginary claims. It is the second part that is the focus here because it is the basis of claims to support pricing and access recommendations; and is the aspect of the evidence report that receives the most media attention.

Even so, it is important to note that in the proposed ICER evidence review no account is taken of the measurement standards required to support claims (and assumptions) from clinical reports, systematic reviews and comparative cohort studies. This absence, unfortunately, characterizes claims for response to therapy and is a fundamental feature of ICER evidence reports. The question that must be addressed is whether or not these reviews, including indirect clinical evaluations, have undertaken a prior assessment of the measurement properties of outcome claims to determine whether they meet ratio or interval standards. In the case of the PICOT (PICO) process in evidence-based practice whether applied to support systematic literature reviews or the framing and answer to clinical questions, the question of measurement standards in outcomes is not addressed. As a result, systematic reviews and study designs are endorsed that fail to meet standards of normal science for credible, evaluable and replicable claims and the standards of fundamental measurement.

The same criticism applies to the Grading Quality of Evidence and Strength of Recommendations (GRADE) process for systematic reviews covering randomized clinical trials (RCTs), non-randomized trials and observational studies; none of criteria for grading (upgrading/downgrading) make any recognition of the limitations of fundamental measurement on value claims and recommendations [5]. GRADE is not alone, we should also add the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) study design checklist for patient reported outcome (PRO) instruments if these are to produce other than ordinal scores as a value claim for response to therapy, where fundamental measurement is ignored [6].

PERSEVERANCE WITTH IMAGINARY CLAIMS

It seems pointless to undertake systematic reviews if there is not a prior measurement filter attached. If a PRO instrument is to create other than ordinal scores; studies should be excluded if the choice of PRO or other outcome fails to meet the required measurement standards. Inadmissible studies are those that fail to meet single attribute, unidimensional ratio or interval standards. While we might expect clinical measures (e.g., diagnostic tools) to meet ratio or interval standards, the same cannot be said of PRO outcomes, both generic and disease specific. The overwhelming majority fail to meet the required standards; they are, at best, ordinal scores, which are inadmissible to support value claims [7] .

Yet, despite criticisms that have pointed to ICERs failure to meet MMT/RMT standards the belief system holds fast; It has now been conclusively demonstrated that all multiattribute preference scores, both direct and indirect, have only ordinal properties. As such they are not able to capture and support claims for response to therapy. As ordinal measures they cannot support claims for quality adjusted life years (QALYs) as ordinal scores cannot support multiplication. Unfortunately, QALYS are a centerpiece for ICER models. The previous ICER MS model and the current proposed model both rely (or will apparently rely) on the QALY. Despite criticism, ICER is quite clear that the modelling structure will be based on prior relapsing MS models including those developed for prior MS related reviews by ICER itself [8]. The base case model will take a health sector perspective with only modeled direct medical care costs. Separate modeled analyses to include productivity and other indirect costs will be accommodate by a separate modeled analysis. The 'end product' will be an incremental cost-per-QALY assessment, using the mathematically impossible QALY, to gage incremental changes greater than 20% of $200,000 per QALY and /or when the result crosses the threshold of $100,000 - $150,000 per QALY gained. This is the standards ICER simulation, driven by assumptions with no concept of empirical evaluation of so-called outcome claims.

This is not the first time ICER has evaluated competing therapies in relapsing MS applying the approximate information assumption driven model [9]. Indeed, the manifest deficiencies in this first model incarnation look set to be repeated in the currently proposed evaluation, despite a critique published in response to the earlier ICER report which noted: *Rather than focusing on modeled claims that are credible, evaluable and replicable in a timeframe that provides rapid feedback to formulary committees, the focus has been on the development of modeled claims that have no chance of ever being evaluated [10].* Non-evaluable outcome claims for therapy impact are a unique feature that defines assumption driven, approximate information modeled simulations. These have been the ICER stock-in-trade for the past decade. The stakeholder draft background and scope documentation for the many modeled simulations makes no mention of the role of empirically evaluable outcome claims; they are of no interest when invented non-evaluable claims are the focus.

It is important to emphasize that ICER occupies an imaginary multiverse of potentially competing model claims; no one model can be claimed to be 'superior' to another or, more to the point, based on more 'realistic' assumptions. Every model's incremental cost-per-QALY

claims and thresholds are of equal merit (or demerit). This, as detailed in the previous commentary, is due to a failure to recognize the problem of induction; claims from the past cannot justify claims on the future. It is a question of simple logic. ICER cannot claim that its modeled choice of assumptions is more 'realistic' or 'believable' than any other choice of assumptions garnered from clinical assessments of product performance or wider systematic reviews; belief, as always, is in the mind of the observer or analyst. The issue was resolved in the 1920s by rejecting a symmetry between proof and disproof by confirmation of claims in favor of an asymmetry: we cannot prove a claim, but we can disprove it. This, of course, leads to the role of empirical assessment with the process of discovery of new facts through the process of conjecture and refutation; or, more simply, the scientific method. This effectively demolishes assumption driven ICER-type imaginary simulations for MS or any other disease state. If PRO instruments are proposed in MS or other chronic disease state then the objective must be for these instruments to support progress in discovering new, yet provisional, facts in MS.

Value claims must be focused on the discovery of new facts; they must subscribe to the contributions of information to discovery that has been the basis for science since the 17th century [11] . This means that value claims must not be seen in isolation, as one-off claims. Rather, they should be justified, not just to fill evidence gaps but as part of a research strategy supporting ongoing disease area and therapeutic class reviews. Certainly, evidence is often limited at product launch, particularly in rare diseases, but the response is to propose a strategy for meeting gaps and hopefully supporting the product over its life cycle not a one-off simulated approximate imaginary information model with non-evaluable conjectures. This may entertain a matinee audience in academe but is no basis for effective health care delivery.

New Start proposes a categorization of value claims to match with the required measurement standards: (i) clinical value claims supported by instruments or algorithms that have ratio or interval properties; (ii) PRO claims that have ratio or interval properties; and (iii) drug and resource utilization claims that have ratio properties. No outcome claim can be accepted unless it meets one of these standards supported by an assessment protocol.

PREFERENCES AND QALYS IN ICER MULTIPLE SCLEROSIS MODELS

The 2017 ICER evidence report's modeling of disease modifying MS therapies relies for its utilities (or preferences) on a UK study that reported EQ-5D-3L values for each MS disease stage [12]. The authors were apparently unaware of the required standards for fundamental evidence and the impossibility of applying ordinal EQ-5D-3L preference scores to create QALYs; an oversight that was replicated in a follow-up study [13] .

The same oversight can be directed towards successor multiattribute utilities from the EQ-5D-5L instrument. There have been a number of MS studies and a recent systematic review of the application of EQ-5D-5L utilities in MS [14]. Once again, however, there is no apparent recognition of the fact that the instrument generates only ordinal scores, with a high proportion of health

states yielding negative scores. This, as in the case of the EQ-5D-3L, points conclusively to the absence of a true zero which means that the QALY based on either of these utility instruments in mathematically impossible. The assumption driven model is, therefore, an impossible construct. Yet, ICER perseveres; even going so far to claim that health economists have confidence that these utility scores actually have ratio measurement properties; no proof is provided for this belief.

Once the standards of fundamental measurement are applied the problem faced by ICER and its academic consultants is that their proposed approximate information HTA simulation models fall at the first hurdle. It is wasted effort to continue to construct incremental cost-per-QALY claims and thresholds on an assumption that the EQ-5D-5L preferences can support QALYs because they have ratio properties; an assumption (or belief) that is patently false.

Modelling MS interventions through the assumption driven ICER simulation is, to be blunt, a waste of time. This reflects the HTA belief system that has been the mainstay of the past 30 years. If coherent and credible value claims in MS are required we need a firm basis in disease specific PRO value claims, but this raises further issues of the relevance in MS of widely used and misapplied MS instruments.

PATIENT REPORTED OUTCOMES IN MULTIPLE SCLEROSIS

PROs in health technology assessment are, not to put too fine a point on it, an unmitigated disaster; MS is no exception. Judged by MMT/RMT standards for fundamental measurement both the generic and disease specific PROs fail. None of the generic direct or indirect multiattribute HRQoL preference claims are acceptable; they only produce ordinal scales. This means that, unequivocally, the QALY is an impossible mathematical construct. Not only do these multiattribute HRQoL instruments or algorithms create negative preference values but they lack dimensional homogeneity and hence construct validity, by bundling together symptoms and response levels which should be assessed individually as single attribute value claims (if this is possible).

The extent of the failure to recognize the limitations of fundamental evidence extends to disease specific PROs, including those applied extensively in MS studies and models. In MS and other disease state the most common PRO is polytomous; that is, it rests upon the application of Likert scales and their integer responses to create scores to capture response to therapy. This is achieved by adding up the integer values over a set of Likert item responses with rescaling, if necessary, to create a score out of 100 (or unity). We need to be quite clear on what the scoring assumes. Consider the Likert scale as the obvious example of polytomous response where each response category, separated by thresholds, is assigned an integer value (e.g., 0 = no problem to 5 = extreme problems) with 4 thresholds: 1-2, 2-3, 3-4, 4-5). Progression in a Likert scale is defined in terms of the probability of moving from one category to another. If 0 = no problem and 5=extreme problem then moving to a lower category (and eventually zero) indicates an 'improvement'. The traditional (and still most common) approach to scoring a set of Likert items is to just add up the integer values (0, 1, 2, 4, 5). This traditional approach

to statistical summation, as Bond and Cox emphasize, is based on the a priori assumptions that all items are of equal difficulty and that the thresholds between steps are of equal distance or equal value [15]. The Rasch model, the foundation for MMT/RMT, makes no such assumptions.

Needless to say, these assumptions are not met in Likert scale-based instruments. Aggregating integer values, even if there is a claimed unidimensional structure where a single latent construct or trait is measured, does not solve the problem. The problem is, however, solved by application of RMT, the Rasch Rating Scale Model (RSM). Unlike the dichotomous or binary Rasch model where there are single item estimates, the RSM not only has a difficulty estimate but threshold estimates that apply to all items on the scale. Thresholds are not imposed; RSM detects (locates) the appropriate threshold structure which in turn defines the relative difficulty in terms of crossing thresholds. This emphasizes the point that the Likert scales are ordinal; RSM transforms these to interval scales based on patient response with items selected that meet or fit Rasch model requirements. Even so, the selection of items may yield a unidimensional structure of increasing item difficulty, but does nothing to address the question of thresholds and their scoring unless RMT is applied through, for example, the partial credit Rasch model.

THE NEURO-QOL INSTRUMENT IN MULTIPLE SCLEROSIS

The NEURO-QoL instrument, which has been used widely in both clinical trials and observational studies in MS, is a prime example of the *add em'up* school of instrument scoring [16]. The Neuro-Qol is a polytomous multiattribute instrument intended for use by adults and children. It is composed of item banks and scales that evaluate symptoms, concerns, and issues that are relevant across disorders (generic measures) along with instruments that assess areas most relevant for specific patient populations (targeted). Scales and short forms are presented with 5 – 10 items in each. There are 17 adult domains and 11 pediatric domains. Each item is presented as a Likert scale with five response options for all items (e.g., 1=Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=Very much). The Neuro-QoL is scored by summation of the integer values for each item. A domain with 8 items has a score range from 8 to 40. The aggregate item count, the raw score, is then translated to an IRT-based T-score for each participant.

Despite the effort put into developing the Neuro-QoL instrument and its widespread acceptance in MS and other neurological conditions, it fails to meet MMT/RMT measurement standards. A current example is the application in the TREAT-MS trial (NCT03500328) where the secondary endpoints include 11 subscales from the Neuro-QoL, none of which meet required MMT/RMT standards for evaluating response to early aggressive therapy versus standard therapy in MS. The primary endpoints, to include assessment of relapse and application of MRI therapy, meet the required standards. Another example is the DELIVER-MS trial (NCT03535298) to assess early intensive intervention versus escalation approaches in MS where, again, the secondary outcomes involve the Neuro-QoL and the various subscales, again failing to meet required standards for evaluating alternative treatment standards.

It is worth noting that both TREAT-MS and DELIVER-MS also use the Multiple Sclerosis Impact Scale (MSIS-29) [17]. Again, this is a Likert item scale (5 response levels: 1 = not at all; 5 = extremely) to assess impact of MS on everyday life in the two-weeks prior to administration. The scale comprises a physical subscale (20 items) and a psychological subscale (9 items). The scales are scored by adding the integer responses on the Likert items to yield overall a score in the range 29 to 145, which is then rescales to a range 0 – 100. As this instrument fails to meet MMT/RMT standards, it products only ordinal scores, it cannot be used to assess any impact measure, in either aggregate terms or for the two subscales.

If we require a raw score for the Likert responses in MS evaluations two assumptions are, therefore, required: first, in each section or domain each item (from an item bank) must be equally difficult and, second, the integer values for each Likert scale the thresholds between steps must be of equal distance or value; the distance between the integer values must be the same. This clearly is not the case; indeed, no proof is provided. The items do not carry the same relative value and the Likert scale is ordinal. It can rank responses, but there is no measure of the difference; it can support only non-parametric statistics. The aggregate score to support the T-scores is ordinal, which means that T-scores are impossible. The item number tables for translating to T-scores are worthless.

Given the emphasis in the Neuro-QoL on item banks as the source for item selection for Likert scoring, it is important to note the distinction between IRT and MMT/RMT: IRT fits the model to the data to establish probabilistic response functions where items are a mathematical function of a person's ability with additional data items to ensure goodness of fit. MMT/RMT applies the principles of conjoint measurement to justify the scale's use as a measurement instrument with invariant, interval properties. IRT in its various levels focuses on accommodating all of the data; MMT/RMT is parsimonious in only utilizing those data elements that fit the model.

THE MULTIPLE SCLEROSIS QUALITY OF LIFE (MSQoL-54) INSTRUMENT

The Multiple Sclerosis Quality of Life (MSQoL-54) Instrument is one of the most widely used instruments in MS in its application over the last 20 years or more and continues to be used in clinical trials; for example, an ongoing German study with Zeposia (ozanimod) NCT05335031 with other recent examples for Zeposia including RADIANCE (NCT02047734) and SUNBEAM (NCT02294058) clinical trials. Overall, according to www.clinicaltrials.gov, 94 MS studies report MSQoL-54 as a secondary endpoint (search 4 May, 2022). Although not apparently as widely used as the MSQoL-54, the Neuro-QoL (28 citations) is also applied with Zeposia, with a new observational study for fatigue in MS currently posted but not recruiting (NCT05319093). There are 200 citations for MSQoL-54 on PubMed (4 May, 2022).

Unfortunately, judged by MMT/RMT the MSQoL-54 exhibits all of the failings that have been attributed to both multiattribute and disease specific PRO instruments. Developed in the mid-1990s, the instrument shows no evidence of the authors being aware of the need to meet the standards of fundamental measurement. The MSQoL-54 is described as a multidimensional

HRQoL instrument to be applied in evaluating the quality of care and the MS therapy effectiveness studies, combining generic as well as disease specific approaches. The generic component is based on the SF-36 instrument (which fails MMT/RMT) and 18 MS specific items. Overall, there are 52 items matched to 12 subscales. Interestingly, MS patients were not apparently involved in its development which means the patient voice and any notion of patient ability and item difficulty, the cornerstone of MMT/TMT are absent. A total HRQoL score can be calculated as well as two SF-36 based scores the Mental Health Composite (MHC) and Physical Health Composite (PHC). A number of classical statistical assessments have made the case for its application [18]. Unfortunately, these are beside the point; they are descriptive and exploratory; the data from the MSQoL-54 have primacy and the intent is to ensure that all the data are accounted for in a multiattribute assembly of disparate symptom and response levels that lack dimensional homogeneity, construct validity and the needs of the patient. This is the direct opposite of MMT/RMT where, for a single attribute or latent PRO construct the data must fit the model; items are selected from patient interviews that support the model; the model is confirmatory and predictive. This is the essence of the premises for the New Start value claims. The MSQoL-54 fails to create single attribute, unidimensional, empirically evaluable value claims for the QoL of MS patients; it should be put to one side.

The MSQoL-54 is an impossible measure. It is a classic example of the failure of what has been described as the *add em'up* school of instrument development. It comprises a variety of Likert scales (from the SF-36) together with 5 level Likert scales for health in general, health distress, cognitive function, sexual function, bowel and bladder function, pain and self-assessed (single Likert 10 item scale) quality of life. Scoring the MSQoL-54 is by simple integer addition over the various scales and rescaling to create a composite total and separate MHC and PHC integer scores. In developing the instrument no though was given to the assumptions needed for item difficulty and invariance of thresholds. The result is that these overall scores are only ordinal scores, as are all of the subscales. This instrument cannot, in MMT/RMT terms provide any confidence as a measure (or measures) of response to therapy; it is irrelevant. This points to the need to filter proposed PRO measures against a checklist for appropriate measurement standards; this should be standard practice for all TCTs and observational studies, not only for MS.

The same objections apply to the Functional Assessment of Multiple Sclerosis (FAMS) instrument. FAMS is a 58 item, 5-point Likert scale self-report, first developed in the mid 1990s [19]. Once again, although readily available, it fails the standards for fundamental evidence with a scoring system that creates only ordinal measures. As such, it cannot report response to therapy interventions or provide an acceptable measure of functional status. There are no citations for FAMS on www.clinialtrials.gov.

The result of this failure to build in the appropriate measurement properties for PRO instruments means that we have only a few that can pass muster in terms of measurement. More specifically, to meet MMT/RMT standards, disease specific PRO value claims are limited by access to disease specific instruments that meet interval or ratio properties (less than 30).

Unless there is a commitment to creating such instruments, then we face a PRO vacuum in technology assessment and PRO value claims. Fortunately, MS can offer a partial respite.

THE PRIMUS CONTRIBUTION IN MULTIPLE SCLEROSIS

The scarcity, or indeed absence on the majority of disease states gives no option but to focus on evaluable clinical and drug and resource utilization value claims. However, although unrecognized by ICER and the majority, if not all, commentators addressing MS PRO instrumentation, there is an acceptable instrument in MS to support empirically evaluable PRO claims; this is the Patient Reported Indices of Multiple Sclerosis (PRIMUS) questionnaire [20] [21]. Not surprisingly, PRIMUS was overlooked by ICER in the 2017 evidence report and will no doubt, in the proposed report.

The PRIMUS instrument was developed some 15 years ago and published in 2009. The genesis was what the authors perceived as the failings of existing PRO measures in MS, specifically the need for a holistic measure to gage the impact of MS to go beyond impairment and activity. The decision was made to create a needs-based measure of QoL. At the same time the opportunity was taken to create scales of symptoms (impairment) and activity limitations that could be used as measures for application in clinical trials; or as value claims in the New Start terminology.

The PRIMUS instrument comprises three scales: MS QoL, MS symptoms and MS activity limitations. The conceptual basis for the PRIMUS classification rests, for the symptom and activity limitation scales of the respective World Health Organization (WHO) classifications for impairment (physiological and anatomical) and activity limitations (capacity and performance) respectively. The PRIMUS QoL scale is based on the needs-fulfillment conceptual model. All three measures take the patient voice as the relevant latent construct or attribute.

The item content for all three scales was derived from intensive patient interviews designed to explore how MS impacted their lives; the result was a selection of item pools for each scale with a final item pool selected for each scale, where the item was selected to fit the Rasch model while maintaining face validity. PRIMUS also supports claims for construct validity given that the instrument is based on a model of the construct assessed and good reliability. The Rasch model captures both the difficulty of the item, expressed in the patient's own words, and the ability of patients to respond to that item as assessed by item responses. This yields a ranking of items with scores representing the extent to which QoL as needs fulfillment is met and the severity of symptoms experienced by the respondent. Response to therapy is captured, therefore, in terms of the possible improvement in QoL and the alleviation of symptoms as well as the experience, from the patient's perspective of the course of MS defined in subjective terms.

The QoL scale comprises 22 binary response items (True/Not True); examples include:

- **I'm neglecting my appearance**

- I feel as if I have nothing to offer anyone
- I avoid physical intimacy
- My self-confidence is affected
- I don't like staying away from home

The symptoms scale comprises 22 binary response items (Yes/No); examples include:

- Have you had any muscle spasms?
- Have you had any numbness?
- Have you been tired all the time?
- Have you found it hard to concentrate?
- Have you had constant pain?

Unfortunately, from the material presented for the activity limitations scale it is impossible to consider a single score to assess the impact of therapy interventions. This is because the responses are represented as 22 5-level Likert scales (0 = never to 4 = All the time) where the results are reported, incorrectly, as just sums over the integer values for each Likert item. The activity limitation scale, therefore, is an ordinal scale. The other two scales have interval properties, given the objectives of Rasch modeling, where each can be transformed to a bounded ratio scale: the N-QoL (MS) for QoL and the (newly designated) R-SYM (MS) scale for symptoms [22].

The same objections apply to the Unidimensional Fatigue Impact Scale (U-FIS) [23]. While subject to Rasch analysis with patient interviews and additional items added to the earlier FIS scale, it is based on 22 Likert scales (5 integer responses: "never" to all the "time") [24]. In the absence of the application of Rasch rating scale model to capture polytomous responses, the scale is scored by integer addition. This is unacceptable unless it can be demonstrated, as noted above, that the Likert items are equally difficult and the response levels have interval properties. It is perhaps unsurprising that even recent systematic reviews of fatigue scales seem oblivious to these standards for evaluating fatigue and the impact of therapy interventions on fatigue in MS patients.

A MULTIPLE SCLEROSIS VALUE CLAIMS STRATEGY

In MS we have all of the elements of a comprehensive value assessment strategy that embraces not only clinical, drug and resource utilization endpoints but also, and most importantly PRO measures that capture the patient voice in MS; PRO measures for QoL and symptoms in MS that have been available, but neglected, for some ten years or more. This presents manufacturers and health systems with the opportunity to address the comparative performance of MS therapies from PRO value claims represented by the patient voice: QoL and symptoms as single attribute, interval scored unidimensional constructs, with the possible transformation to bounded ratio scores; a true zero capped at unity.

The strategy is straightforward; with the first essential step one of rejecting the ICER simulation modeling assumption driven evidence base; making quite clear the reasons for rejection. This will counter systematic reviews that have attempted to review how MS modeled economic evaluations have evolved. In fact, evolution is the wrong term: with ICER leading in the US; they are not only an analytical dead end, but could never meet the required standards for assessing therapy response. It is clear, in the most recent review, that the authors had no idea of the required standards of normal science and measurement theory, continuing to believe in the role of Markov cohort models and multiattribute utility scores. Rasch analysis seems to be an unknown in instrument development and critiques of the current range of instruments [25] .

We have to focus on single value claims, agreed by the manufacturer, health systems and patient representatives. These set the stage for a research program in MS, not just to meet evidence gaps but to discover new facts in therapy response. Protocols are the key link between the value claim and the evidence base; if these are agreed for clinical, PRO and drug and resource utilization value claims, then we have a basis for tracking MS patients and supporting ongoing disease area and therapeutic class reviews. Modeling is an unnecessary and unwanted distraction.

CONCLUSIONS

MS in common with other chronic disease states has few options if PRO value claims are proposed to support formulary submissions. The fact that all generic PRO claims and the overwhelming majority of MS specific PRO claims fail to meet the standards of normal science and fundamental evidence should give pause for thought as a salutary lesson in a willingness to accept PRO measures that fail criteria that should have been applied before committing them to RCTs and observational studies. Both clinicians developing these irrelevant PRO measures and manufacturers accepting them at face value are to blame; not to forget the HTA belief system promoting imaginary constructs.

As it stands there are only two PRO measures in MS that meet the required measurement standards. These are: (i) the PRIMUS need-fulfillment quality of life instrument; and (ii) the PRIMUS symptoms instrument. In both cases, as the item responses are binary, the Rasch interval scale can be transformed to a bounded interval ratio scale to create an N-QoL and a R-SYM scale respectively. There is no PRO instrument that captures fatigue and meets the required measurement standards. The PRIMUS activity limitations instrument should not be utilized if an overall score is required as it is invalid as the sum of item integer values. The same prohibition applies to the U-FIS scale. Unless transformed to meet Rasch standards the PROMIS activity and U-FIS scales should only be presented as integer distributions that lack interval properties.

None of the modeled outcome claims proposed by ICER in MS will, by intent and design, be empirically evaluable. Apart from the fact that the adoption of an assumption driven lifetime simulation model, irrespective of its choice of possible structural features, is an irrelevant non-

starter, the assumptions built into the model, also guarantee its irrelevance as one of many possible Imaginary. If formulary committees and health care systems are interested in comparative value claims for the products ICER is proposing to evaluate for MS interventions, there is only one analytical framework that ensures all value claims meet the standards of normal science and fundamental measurement; a framework that has been described as the New Start in formulary decision making. The New Start framework is distinguished by its rejection of assumption driven simulation models and its acceptance of the role of single attribute, unidimensional vale claims which, supported by evaluation protocols for each value claim; providing empirically evaluable value claims specific to clinical, PRO, drug and resource utilization end points. Unfortunately, given the commitment to rejecting the standards of normal science and fundamental measurement in HTA, as witnessed by the ICER belief system, the acceptance of these New Start standards is by no means assured, in MS or other chronic disease.

REFERENCES

[1] 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 (https://doi.org/10.12688/f1000research.109389.1)

[2] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 22) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth*. 2022;25(1):3-9

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

[4] Institute for Clinical and Economic Review. Treatments for Multiple Sclerosis: Effectiveness and Value. Draft Background and Scope. April 21, 2022

[5] Brozek J, Akl E, Alonso-Coello P, Lang D et a; GRADE Working Group. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy*. 2009; 64(5): 669-77

[6] McKenna SP, Heaney A. Setting and maintaining standards for patient-reported outcome measures: Can we rely on the COSMIN checklists? *J Med Econ*. 2021;24(1):502–511

[7] 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

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

[9] Institute for Clinical and Economic Review. Disease-Modifying Therapies for Relapsing-Remitting and Primary-Progressive Multiple Sclerosis: Effectiveness and Value. Final Evidence Report. March 6, 2017

[10] Langley PC. Multiple Sclerosis and the Comparative Value Disease Modifying Therapy Report of the Institute for Clinical and Economic Review (ICER). *InovPharm.* 2017;8(1): No. 12

[11] Wootton D. The Invention of Science: A New History of the Scientific Revolution. New York: Harper Collins, 2015

[12] National Institute for Health and Care Excellence. Dimethyl fumerate for treating relpsing-remitting multiple sclerosis. Final Appraisal Determination. 2014

[13] Mauskopf J, Fay M, Iyer R et al. Cost-effectiveness of delayed-release dimethyl fumerate for the treatment of relapsing forms of multiple sclerosis in the United States. J Med Econ. 2016;19(4):432-42

[14] Wiyani A, Badgujar L, Khurana V et al. How have economic evaluations in Relapsing Multiple Sclerosis evolved over time? A systematic literature review. *Neurol Ther*. 2021; 10:557-83

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

[16] National Institute of Neurological Disorders and Stroke (NINDS). User Manual for the Quality of Life in Neurological Disorders (Neuro-QoL) Measures, Version 2.0, March 2015

[17] Hobart J, Lamping D, Fitzpatrick R et al. The Multiple Sclerosis Impact Scale (MSIS-29). *Brain*. 2001;124(5):962-973

[18] Giordano A, Testa S, Bassi M et al. Viability of a MSQoL-54 general health-related quality of life score using bifactor model. *Health Qual Life Outcomes*. 2021;19:224

[19] Cella D, Dineen K, Arnason B et al. Validation of the functional assessment of multiple sclerosis quality of life instrument. *Neurology* 1996;47(1):129-39

[20] Doward L, McKenna S, Meads D et al. The development of patient-reported outcome indices for multiple sclerosis (PRIMUS). *Mult Scler*. 2009;15(9):1092-102

[21] McKenna S, Doward L, Twiss J et al. International development of the Patient-Reported Outcome Indices for Multiple Sclerosis (PRIMUS). *ValueHealth*. 2010;13(8):946-51

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

[23] Meads D, Doward L, McKenna S et al. The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS). *Mult Scler*. 2009;15(10);1228-38

[24] Twiss J, Doward L, McKenna S et al. Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS9). *Health Qual Life Outcomes*.2010; 8:17

[25] **Machado M, Kang N-Y, Tai F et al. Measuring fatigue: a meta review. *Int J Derm*. 2021;60(9): 1953-69.**