MAIMON WORKING PAPERS No. 5 March 2023

VALUE CLAIMS, FUNDAMENTAL MEASUREMENT AND THE PATIENT VOICE IN ONCOLOGY: A SYSTEMIC FAILURE

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

ABSTRACT

Value claims for pharmaceutical products and devices must meet recognized standards to include both the standards for credibility, empirical evaluation, and replication that characterize the belief in progress that supports the commitment to normal science as well as those for fundamental measurement which require value claims to be for single attributes with interval or ratio measurement. Value claims which fail these standards must be rejected. The implications of this insistence or imperative for these standards have profound implications for health technology assessment. For the past 30 years, HTA has been locked into a meme that denies these standards. Instead, there is a commitment to developing assumptions-driven modeled simulations to create approximate information and non-evaluable claims for cost-effectiveness. This methodology is applied across the board, supported by the mathematically impossible quality-adjusted life year (QALY), where value claims in oncology and other disease areas are essentially a waste of time. The purpose of this commentary is to make the case that we need a new start, not just in oncology, to establish a meaningful framework for evaluating therapy options, setting the stage for the evolution of objective knowledge through lifetime disease area and therapeutic class reviews, supported by effective real-world outcomes-based contracting. Oncology is a significant starting point because of the extent to which the failure to meet these standards is institutionalized with groups such as the European Society for Medical Oncology (EORTC) and the required adoption of patient-centric disease-specific oncology measures that meet the standards for fundamental measurement.

Keywords: EORTC, QLQ-C30, QLU-C10D, EQ-HWB, ASCO, Rasch measurement failure, unidimensional interval measures,

INTRODUCTION

Value claims for products and therapy interventions in medicine and health technology assessment (HTA) are only acceptable if they meet normal science standards and fundamental measurement standards. Accepting these standards means that HTA conforms to accepted standards in the physical sciences and the more mature social sciences such as economics and education [1]. The hallmark of the standards in normal science is that all value claims must be credible, empirically evaluable and replicable; the commitment to set the stage for hypotheses testing, the discovery of provisional new facts, and ongoing disease area and therapeutic class reviews. Meeting the standards for fundamental measurement is also critical. Whether claims are for clinical endpoints, patient-reported outcomes, drug utilization or other resource utilization, all must meet standards

as interval or ratio measures for a single attribute; bundling attributes to produce overall scores is unacceptable.

The purpose of this commentary is to make the case that, judged by the standards of normal science and modern measurement or Rasch Measurement Theory (RMT), the last 30 years have witnessed a profound systemic failure to create instruments to evaluate therapy response and value claims in oncology. The failure is deep-seated, in particular with the role of the European Society for Medical Oncology (EORTC) in its support for the generic EORTC Core Quality of Life Questionnaire (QLQ-C30) together with guidance for supplementary disease-specific oncology modules and instruments, and more recently general population normative data for 13 European countries, the US and Canada [2 3 4]. Attention must also be given to the proposal to use items from the QLQ-30 to create a multiattribute preference cancer instrument, the QLU-C10D as an analog for generic multiattribute instruments such as the EQ-5D-5L for application in cancer claims with preferences determined at national levels. At the same time over the past eight years, other value frameworks have been proposed, notably, the American Society for Clinical Oncology (ASCO) value framework which again fails the required standards for normal science and fundamental measurement [5]. Given the case presented here, we need to put the current standards for oncology outcome instruments to one side as an analytical dead end in favor of a new paradigm that sets the basis for an understanding of oncology outcome measures that meet long-established standards in both normal science and, for patient response to therapy; modern or Rasch measurement Theory (RMT) [6 7]. If we are concerned with progress and the discovery of new yet provisional facts in oncology therapy response to capture the patient voice, then we must start with measurement where the Rasch model is the ideal. Oncology is not, it should be emphasized, unique in its lack of awareness of these standards; it is a basic belief in what has been described as the HTA meme with 30 years of advocacy for measures and modeled claims for cost-effectiveness that are an analytical dead-end [8 1].

REQUIRED VALUE CLAIM STANDARDS: NON-PHYSICAL ATTRIBUTES

Judged by the standards for fundamental measurement, in particular RMT, the development of PRO oncology instruments, both the generic QLQ-C30, the QLU-C10D offshoot and the disease-specific modules to capture non-physical attributes, are an unprecedented failure in measurement. Clinicians and others have developed modules to capture symptoms, functioning, and other specific problems with no thought given to the measurement process: first, to make the case that a proposed latent construct or trait is quantifiable and, second, that it is necessary to construct a measure of this trait that can be subject to statistical analysis. Allocating numbers to events is not measurement; fundamental measurement is not just allocating numbers to events based on the classification of nominal, ordinal, interval, and ratio scales; we have to ensure that the data we are analyzing adhere to the measurement principles of the physical sciences. The essential claim is that RMT instantiates the principles of probabilistic conjoint measurement in detecting measurement structures in non-physical attributes. Certainly, in RMT we look to the creation of a single attribute, unidimensional linear interval scale with additive and invariance properties by applying Rasch rules to subjective ordinal observations. The term transformation is often used as shorthand for the RMT process, but the rules require a process by which the conjoint items (in the Rasch model patient ability and item difficulty) are evaluated to produce probabilistic responses. That is, the probability of successfully responding to (or meeting) an item is a function of the

difference between the ability of the person and the difficulty of an item. The importance of RMT is that with the application of conjoint simultaneous counts (of item difficulty and patient ability), we have the basis for modeling measures based on probabilistic relations Rasch rules provide the necessary and sufficient means to transform these ordinal counts into linear interval single attribute scales, *where the measured behaviors are expressions of the underlying construct* [7].

Central to RMT is order; while subjective responses are qualitative, they have order which implies less or more of the property being assessed. Where item responses in a questionnaire are dichotomous (1,0), a positive response (=1) indicates more of the trait than a zero response; there is direction to the response. Similarly, for polytomous responses (scored 0,1,2,3,4), a response of 4 indicates more of the trait than 3. However, while assigning an integer to indicate order is a key step towards quantification, it is not measurement. The seminal contribution of RMT is to demonstrate how these assessments, or subjective responses, can be transformed to approximate single attribute or unidimensional linear, interval measurement.

With the application of Rasch rules, we fit an assessment onto a line, a linear continuum, that is divided into equal units. To manifest or measure the property of a trait or latent construct that is of interest, the object has to be matched to the measuring instrument. In other words, measurement in the social sciences is no different from measurement in the physical sciences where the property of an object (e.g., temperature) is to be measured or manifested. This is the step that has been overlooked, not just in developing response claims in oncology, but generally in health technology assessment in that measures must apply only to single attributes that manifest a latent construct or trait. If a value claim for response to therapy is to have any validity, it must be based on an interval (or ratio) measure. RMT is unique in being the necessary and sufficient mathematical framework for processing assessments or subjective responses to a single attribute, linear interval scale, or a measure consistent with the standards of the physical sciences [9].

The application of Rasch analysis provides a formal test of an outcome scale against a measurement model, operationalizing the formal axioms that underpin measurement: these axioms of additive conjoint measurement are the only rules and will determine whether interval or ratio scales have been constructed [10]. The Rasch model takes precedence with the items selected determined by the model. Application of the axioms of conjoint measurement is achieved by an iterative process to generate an approximation to an interval scale (Rasch continuum); the Rasch criteria support the approximation of the application of Rasch rule to move from ordinal counts to single attribute, additive linear interval measures for items and persons that are invariant across intended applications [7]. The criteria are:

- Overall instrument and item functioning (reliability, individual item fit statistics, global model fit)
- Unidimensionality of underlying construct
- Local independence of items
- Categories and thresholds ordering (polytomous instruments)
- Differential item functioning
- Person and item alignment

The judgment is holistic; which means it is important that a full range of statistical assessments for each of the criteria are presented and reasons for acceptance detailed. Presenting these assessment criteria is important because it presents third parties with the option of agreeing or disagreeing with the holistic or overall assessment that the hypothesis is reasonable in claiming approximation to an interval scale; there is no magic transformation but a maximum likelihood estimation taking us from scores on items to locations on the Rasch continuum. The focus is on the measurement qualities of the data we have collected [7]. To be clear, Rasch rules assign numbers to objects (items) that preserves the relations between the objects. Empirical relations come first. The resulting logit scales are not units of something real but just real numbers generated from data. There is no preassigned standard measure. It is only under stringent conditions of conjoint measurement imposed on the observed data that the logit scale of the Rasch model can be shown, if this is the case, to be an interval scale. The Rasch model instantiates the principles of probabilistic conjoint measurement to produce invariant interval-scale measures in which the principles of concatenation apply. We approximate the interval-level scales of the physical sciences in the human sciences via probabilistic conjoint measurement. This precedes statistical measurement. Rasch attempts the task of developing and calibrating data collection instruments.

RMT does not support the creation of composite instruments. These are disallowed because of the need to ensure dimensionality and dimensional homogeneity [11]. Proposing composite measures such as the QLQ-30 and its reduced item offshoot the QLU-C10D are illusory; chasing measures that are nothing more than a will o'the wisp. Ensuring the unidimensionality in an instrument that is designed to manifest a latent trait is mandatory. All the items in an instrument must support a single construct. Attempting to add together different latent constructs, such as bundled health state descriptions, will deny unidimensionality. Of course, once your composite score has been developed, there is the appeal of attempting to claim unidimensionality, or factors with unidimensional attributes. This is disallowed by the fact that in ignoring Rasch rules, you have only an ordinal score which supports only non-parametric statistics; *factor analysis is faulted by mistaking ordinally labeled stochastic observations for linear measures and failing to construct linear measures* [12] . If we are to judge the merits of a measured manifestation of a latent construct then, as with Rasch measurement, we require a coherent construct theory that orders observations and a specification equation. This allows scores can be predicted on a linear interval scale from responses to items. Add to these requirements the role of dimensional homogeneity: we can compare variables only if they have the same dimension and can be converted to each other (e.g., centigrade and fahrenheit). If there are different dimensions, as there are by definition in composite health related quality of life (HRQoL) bundles (e.g., EQ-5D-5L symptom dimensions) then they all break the rules for dimensional homogeneity and hence construct validity.

A NEW START IN HEALTH TECHNOLOGY ASSESSMENT

Briefly, the proposed new start in HTA focus on the creation of a profile of single attribute value claims to support formulary submissions, ongoing disease area, and therapeutic class reviews, and, if required, outcomes-based contracting. The application of RMT is only one aspect of the required evidence and value claim standards. The three premises for the new start are:

- All value claims must refer to single attributes that meet the demarcation standards for normal science: 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 meet ratio or interval properties
- All value claims must be supported by a protocol detailing how the claims are to be assessed and reported, in a meaningful time frame, to the health system

Value claims are to be considered as (i) clinical claims with instrumentation that meets the standards for measurement instruments in the physical sciences; (ii) patient-reported outcomes claims that refer to a latent construct and application of RMT to create single attribute, linear interval measurement; (ii) drug utilization claims for market entry therapy impact, drug switching, and compliance behavior; and (iv) non-drug resource utilization impacts. Both (iii) and (iv) are claims presented as units, not costs. This allows tracking of the claim in real-world treating environments from established online databases with Federally mandated classification systems. If there is a requirement to translate these to costs, then this is the responsibility of the health system, which may provide a unit pricing schedule.

The application of RMT for non-physical attributes is essential if we are to meet standards for fundamental measurement; this sets it apart from the mainstream focus in HTA with its acceptance of ordinal scales as the basis for evaluating therapy response. The application of the Rasch model to construct fundamental measures sets it apart from classical test theory (CTT) and item response theory (IRT). The data have primacy in CTT and IRT where the objectives are exploratory and descriptive in attempting to account for all the data. The Rasch model is confirmatory, where data items are identified to fit the model and predictive of success in item response [7]. The Rasch model asks two questions: how well does the empirical data fit the measurement model requirements and does the instrument that has been developed yield single attribute or unidimensional interval measures on a linear scale? If we can ensure the item fits to the Rasch model, then we can claim that the requirements of probabilistic conjoint measurement have been realized sufficiently to make the claim that we have the required measure [7].

It should not be thought that RMT has only just been proposed to meet requirements for measuring therapy response. Rasch's seminal contribution was developed in the 1950s with widespread application since then in education and psychology and, to a lesser extent since the early 1990s with needs fulfillment instruments to assess therapy response in clinical trials [13]. Not only are there Rasch conferences, websites, textbooks, and transactions but also low-cost software packages to apply Rasch rules to develop linear integer measures for dichotomous and polytomous questionnaires. There is no reason why these could not have been applied to EORTC supported oncology instruments as the software has been available for over 40 years, as well as to instrument development in the wider HTA context.

MEASUREMENT FAILURE IN ONCOLOGY: THE EORTC QLQ-C30 (Version 3.0)

It's instructive to consider that at the time the EORTC QLQ-C30 was first developed in the 1980s there were ample red flag warnings that pointed to the need to focus on fundamental measurement: manifesting single attribute linear interval measure of a latent trait or construct [14]. Yet, the focus on the development of a generic instrument that bundled together disparate HRQoL dimensions deem relevant across the board for cancer patients, took priority. In terms of Rasch measurement, the effort failed at this first hurdle. This decision put to one side a more thoughtful question: if

Rasch measurement disallows composite multidimension (or multiattribute) ordinal scores why not focus on a single attribute that is common across cancer states: A coherent manifestation of the patient voice in therapy interventions as the patient is the ultimate judge and possible beneficiary of therapy interventions.

The starting place must be a latent construct that is considered credible for patients and caregivers in disease states, and which can be manifested by the application of Rasch rules. Since the early 1990s the proposed latent construct is needs-fulfillment; the framework for a latent construct or trait that the quality of life of a patient (or caregiver) is determined by the extent to which the needs that are identified from extensive subject interviews are met. In oncology and other chronic disease states, while health interventions may be expected to be the principal factors that impact needs, the needs of patients may not correlate with HRQoL clinical parameter considerations. Whether a therapy facilitates needs being fulfilled, the judgement belongs to the patient. It is an empirical question which requires a linear, interval single attribute measure, with items defined in either dichotomous or polytomous terms, subject to Rasch rules to include the Rasch Rating Scale Model with a rating scale structure common to all items, or the Partial Credit Rasch Model which incorporates having different numbers of response levels for items in the same instrument [6 7] . If RMT had been recognized there was the opportunity, given agreement on the needs-fulfillment attribute of a common latent construct to be assessed and transformed into a linear, interval measure.

Judged by the standard for fundamental measurement the QLQ-C30 is a failure. All that has been produced are transformed raw scores from the averaging of Likert items. There appears to be no concept of the need for a single attribute linear interval scale to capture measurement for specific cancer disease states. At best, we have a collection of functional items and symptoms which even by the standards for aggregating Likert responses fail because if we want to add Likert items we require an *a priori* assumption that all items are of equal difficulty and that the thresholds between the steps are of equal distance. The Rasch model for polytomous responses makes no such assumptions. Indeed, we know, by application of IRT analysis that the QLQ-C30 items vary in terms of their difficulty [15] ; the problem of thresholds for different items has also been addressed in terms of thresholds for clinical symptoms [16]. While the QLQ-C30 has been categorized as an instrument to capture quality of life, it is better seen as including a one-item question asking about quality of life as simple integer values (c.f., Likert pain scales) with symptoms and functional status tagged on. As it stands, we have a composite or multiattribute instrument that lacks dimensionality and dimensional homogeneity and which fails the standards for simple aggregation of integer values to generate a raw score (where standardization is also disallowed). While we might categorize these raw scores as ordinal, they are no better than the multiattribute preference scores or utilities generated by the EQ-5D-3L/5L instruments. In terms of the standards for fundamental measure, the QLQ-C30 is not a generic quality-of-life instrument with acceptable measurement properties. A focus on specific items raises the question of why bother when there are many disease-specific instruments that could report, possibly more comprehensively, on the symptoms and functions identified in the QLQ-C30. Unfortunately, the failure to meet Rasch standards applies equally to these as they are typically Likert-based polytomous instruments with the same lack of appreciation of prior assumptions for item difficulty and threshold distance to create numbers.

Certainly, given the focus here on the patient voice in needs fulfillment, patient interviews are the first requirement; but they should not be about issues; we must be more specific. The items selected as the first stage in instrument development must focus on the assessment of the latent trait or construct that we are hoping to manifest quantitatively as a single attribute, linear interval measure. It must always be remembered that focusing on symptoms and functions may suggest prospective instrument items that are of little interest to patients as elements in their quality of life; we might infer that there is an impact but we need a direct measure. This is where the needs of patients and caregivers become pivotal [17 18 19]. If quality of life takes its cue from the ability of patients to meet their needs, where chronic disease may have a major role, then we have to identify those needs; not a collection of Likert scores for categories of symptoms and functional status. That is, by items selected from interviews that are intended to assess an underlying trait or latent construct. If the latent construct is needs fulfillment, then we have a firm basis for developing cancer disease-specific instruments aiming for an accurate repose for perceived direct patent benefit. This is an imperative: a measure with over 60 years of experience in its application.

MEASUREMENT FAILURE IN ONCOLOGY: THE EORTC QLQ-C10D

Over the past 10 years considerable effort has gone into creating a multiattribute utility instrument from the QLQ-C30, the EORTC QLU-C10D. Supported by the Multi-attribute Utility in Cancer (MAUCa) Consortium the QLU-C10D is intended to support the use of HRQoL data in cost utility analysis with country specific value sets for the QlQ-C30, where the QLQ-C30 is a multiattribute polytomous instrument that yields only ordinal integer summation scores. The failure of integer scores to meet the standards of Rasch or modern measurement theory is, as noted here, well established. Although there has been an attempt to apply Rasch rules to improve reliability, the effort has been effectively ignored [15]. The QLU-C10D preference scoring attempt is based on 10 of the QLQ-C30's 30 items, combined into 10 dimensions, with valuation of a small sample of these health state polytomous descriptions (4^{10} = 1,048, 576) with discrete choice comparisons to estimate the utility model parameters consistent with quality adjusted life year (QALY) model restrictions. That is, with 1 = perfect health (or no problems for each of the health states), decrements were estimated that interacted with time so that the worse possible health state had a utility of 0 = death. The model actually yielded, as with the EQ-5D-3L/5L, algorithms, states worse than death where each response level for each symptom dimension was assigned a negative score as the decrement. In effect, we now have to proposed generic measures of quality of life in cancer: the QLQ-C30 30 item polytomous instrument that yields integer ordinal scores and the QLU-C10D preference instrument that, as detailed below, yields only ordinal scales for a subset of items. The two instruments yield incompatible ordinal scores; neither of which can capture response to therapy. They can only support non-parametric statistics. Measures must be interval (or ratio) and the creation of these interval linear scales is a prerequisite to classical statistical analysis, including factor analysis [13] .

There are a number of issues that need to be resolved in order to judge whether the QLU-C10D has any useful application on health technology assessment; questions which are common to all multiattribute utility scores. The first issue is the impossibility of constructing multiattribute scores; they have no value as measures as they lack any basis in a single latent construct or trait. Bundling symptoms and response levels into a dimensionless hybrid claim as a meaningful starting point is a non-starter. The QLU-C10D fails at the first hurdle for precisely the same reasons as the

QLQ-C30 and the popular generic multiattribute instruments such as the EQ-5D-5L and the recently 'launched' EQ-Health and Wellbeing (EQ-HWB) instrument [20]. The problem is that multiattribute scores (which applies to both instruments) fail to recognize the limitations of measurement theory together with the absence of a coherent theoretical model that justifies the addition of the individual dimensions or indicators to create a composite instrument. The instrument is based on different dimensions of health and functioning and therefore lacks dimensional homogeneity. They cannot be just added together to give an overall score (and differential weighting merely confuses the issue). Given the axioms of fundamental measurement unless there is coherent latent construct we should go no further

The second issue is the community valuation of health states. These are just subjective weights; even with discrete choice as opposed to time trade off (TTO) or standard gamble (SG) preference weights, and as such yield only ordinal scores [15]. Again, this reflects a failure to understand Rasch measurement theory (RMT) and the role of RMT as *the necessary and sufficient means* to transform ordinal counts to linear, interval measures for single attributes [10].

The third issue is the required characteristics of a preference score to create the even more popular quality adjusted life years (QALYs). To create a QALY time (a ratio measure) must be multiplied by a ratio preference measure; not an ordinal scale. This is not the case with the QLY-C10D which yields only a disallowed multiattribute ordinal scale. There seems no intent on the part of those developing either the QLQ-C30 or the QLY-C10D that at the end of the day they require a bounded ratio scale that meets the standards of RMT, as demonstrated in a recent proposal for an algorithm that creates such a scale from linear interval data.[21] .

The fourth issue goes to the heart of RMT: to focus on latent constructs that yield an assessment instrument that capture, as the contribution of conjoint simultaneous measurement, requires capturing the interaction between the patient and the difficulty of an item. This takes us to a latent construct in quality of life that has been recognized for over 30 years: needs-fulfillment. If life takes its meaning from needs being articulated and met, with health a potential dominant factor, we need to develop single attribute instruments to measure or manifest those needs. The focus must be on the patient voice as the ultimate beneficiary of therapy interventions. In the QLU-C10D the patient is absent; there is no attempt to consider needs, only community valuations of health states which, deliberately, fail to mention cancer as bundles of multiattribute Likert dimensions which are subjectively valued. Presumably, the case could be made that the QLU-C10D could equally be applied in other chronic disease which then sets it up as a competitor the EQ-Health and Wellbeing (EQ-HWB) instrument which is seen as both a complement and a successor to the EQ-5D-5L [23]. Indeed, as the principal authors of the EQ-HWB are the same as those for the QLU-C10D, this failure in measurement is compounded. A reasonable question is why we need the two instruments as the latter was valued as health states without mention of cancer. As it is, they are both irrelevant as measures.

The fifth and final point is the proposed application of the ordinal QLU-C10D score. If it is proposed to contribute to a 'cancer' QALY (which the ordinal EQ-HWB could equally well serve) to populate assumption driven simulation lifetime cost-effectiveness models then we reach an analytical dead end; these models fail both the standards for normal science and fundamental measurement. In short, there is no place for the various national CLU-C10D country specific

ordinal utility weights or preferences in HTA [22] [23] [24] [25] [26]. If the QLU-C10D methodology is seen as a necessary successor to mapping from the QLQ-C30 to a preference instrument such as the EQ-5D-5L, then it has not succeeded [27] . We must never confuse counts with measures; they must be analyzed to discover whether a linear, interval measure can be constructed. Both the QLQ-C30 and QLU-C10D fall at the first hurdle in failing to construct a single attribute, unidimensional, linear, interval, additive and invariant measures to support subsequent analysis.

MEASUREMENT FAILURE IN ONCOLOGY: THE EORTC CANCER SPECIFIC MODULES

There appears to be no concept by EORTC of focus on a credible latent construct nor how the preliminary item collection (the instrument) is to be assessed, other than by evaluating content validity, but not construct validity, and then introducing the item collection as the instrument. This holds for all the cancer modules endorsed by EORTC. At no stage is there discussion of the imperative of an instrument that is created by an application of Rasch rules for conjoint assessment of initial assessment to create a single attribute, linear interval scale. The argument here is that if we accept, as we must to maintain credibility, that claims for cancer therapies must be defined, across the board, as single attributes subject to Rasch rules for an interval measure, then the most obvious latent construct to capture as a common measures latent construct or trait quality of life, is needs fulfillment; a framework that has been utilized for almost 30 other disease state instruments over the past 25 years.

The failure of the QLQ-C30 to meet fundamental measurement standards is not an isolated occurrence as it applies across the board to the various disease-specific modules (and the module item library) promoted by EORTC. This is seen in the EORTC guidelines for module development, both for stand-alone modules and those developed as complements to the QLQ-C30 [2]. Four process steps are required by EORTC: (i) item identification from patient interviews to identify relevant quality of life ‘issues’ of concern; step (ii) s to convert issues to items with mandatory use of the EORTC QLQ item library (to avoid duplication of existing items) and construct a provisional instrument; (iii) pretesting of the provisional module and its psychometric properties.; and (iv) field testing to determine acceptability, reliability, validity, responsiveness, and cross-cultural applicability.

At no stage in this process is the question of measurement and response addressed. Any instrument must be designed to capture a meaningful measure of response for a latent construct or trait: a single attribute, linear interval measure. This is the only basis for response to therapy. There is no discussion of the latent construct or trait of interest apart from a belief that quality of life (undefined) can be proposed, manifested, and measured. What is overlooked is that in order to support a psychometric evaluation you need first a single attribute, linear interval measure; an ordinal scale only supports non-parametric statistics.

Certainly, the EORTC process is focused on issues, but this is insufficient.There is no concept of the difficulty of relevant ‘issues’ or the ability of the patient to respond to those issues. This effectively excludes any notion or recognition of the Rasch model for instrument development; RMT has never been of concern to EORTC to support instrument development. If patients have

needs that are identified in specific cancer states there is no interest in the application of fundamental measurement to manifest those needs in a coherent analytical framework.

This position stands in contrast to considerations of the application of RMT to rheumatology that appeared some 15 years ago and, more recently the Rasch Reporting Guideline for Rehabilitation Research (RULER) for assessment and application in rehabilitation medicine. Indeed, the RULER guidelines are a model that EORTC could emulate [10] [28] [29]. Whether EORTC is in a position to adopt RMT as the focus for instrument development in oncology as part of a commitment to a new start in HTA is an open question; Unfortunately, the likelihood is low if for no other reason that too many people have too much to lose; including the implicit admission that value claims for comparative therapy impact, mapping to assumption driven simulation to model cost-effectiveness, that have appeared over 30 years fail standards for robust measurement [30].

THE MAPPING DEBACLE

Considerable time and effort have been expended over the past decade or more to the development of mapping functions to support modeled economic evaluations when the required generic multiattribute utility or preference scores are not available from disease-specific instruments with a recent example of the attempt to map the EORTC QLQ-30 and the QLQ-H&N35 to the EQ-5D [31] [32]. The primary driver has been the prospect of creating claims for increment cost-per-QALYs to support cost-effectiveness recommendations for pricing and access. Unfortunately, as has now been well established, assumption-driven modeled lifetime simulations to create overall benefit claims fail the standards of normal science [1]. Not only is the QALY a mathematically impossible construct but the extensive application of modeled assumptions invalidates any claim for one model to be more realistic of an unknown future than another [33] [33]. They fail Hume's problem of induction (known since the 18th century for denying confirmation of hypotheses) given that while past futures have resembled past pasts, there is no basis in logic for the assumption that future futures will resemble future pasts [34].

Mapping, if the intention is to populate one of any number of assumption-driven, non-evaluable claims on the future for cancer therapy, is not only supporting a modeling framework that is an analytical dead end but is made more irrelevant by the fact that the integer scores from the cancer instrument are ordinal, just as the multiattribute utility and preference scores are also ordinal (and composite rather than single attribute). While you can transform one interval measure to another (e.g., Fahrenheit to centigrade) you cannot transform by regression modeling or any other technique, one ordinal scale to another ordinal scale. It is a singularly fruitless exercise.

COMPOSITE VALUE CLAIMS FRAMEWORKS: ASCO NET HEALTH BENEFIT

The ASCO Value Framework is intended to assess the relative value of cancer therapies by calculating a net health benefit score (NHB) using measures of clinical benefit (survival estimates) and toxicities. This is a measure that is entirely clinician focused and we are asked to infer benefit to the patient from the benefits judged by physicians. This is an absurd position. Presumably, clinicians would agree that the patient is the ultimate beneficiary; not physician-determined end-points. The patient's voice is entirely absent. There is no measure of the direct benefit to the patient; there is no attempt to consider a latent concept or trait of 'benefit' and how this can be manifested

in a single attribute linear interval measure. Certainly, it is possible to compute the inputs to the NHB but this does not capture the relative benefits to the patient. This reflects a lack of attention in the protocols for cancer trials to introduce measures that follow Rasch rules and define the benefits in probabilistic terms where the probability of successful benefits with needs being met with a new therapy is a function of the difference between the difficulty of an item and the ability of the patient.

A further problem is the application of subjective weights for toxicities. These fail the standards for evaluating scores from Likert scales in terms of the toxicity impact, the assumption that all toxicities have equal value, and the thresholds or distance between the categories with arbitrary grading weights; with further subjective weights for bonus points.

The most significant objection to the ASCO value framework is that in attempting to create a single metric for net benefit it fails the standards for empirical evaluation and replication. It is a composite measure that bundles together, with subjective weights, attributes for survivorship and toxicity which lack construct validity. We cannot assume that a calculated NHB from prior clinical trials will hold in the future. The required standard for all value claims, as single attributes with the required linear, interval measurement properties, is that they are credible, evaluable, and replicable. Formulary committees may accept claims based on clinical endpoints from trials, but we cannot assume they will hold in the future; the fact that past futures have resembled past pasts does not mean that future futures will resemble future pasts. This is why all value claims for products are provisional. Value claims across the board must be empirically evaluable and replicable supported by an assessment protocol. This is the essence of the standards of normal science.

Value claims must be for single attributes whether these refer to clinical, instrument-defined endpoints, the non-physical patient-reported outcome, value claims for drug utilization and compliance, and claims for other resource utilization impacts. Formulary submission guidelines must make this clear. Certainly, propose value claims that might be inputs to a future estimate of the NHB worksheet; but be equally suspicious of the structure and weights proposed by the NHB and the justification for them. If a manufacturer is intent on an NHB benefit value claim it should be made quite clear what the numerical value of that claim is expected to be, supported by a protocol detailing how the NHB metric is to be created. The paradox is, of course, that in proposing a protocol for an 'enhanced' NHB metric, there is no basis for proposing what the metric will be to support provisional acceptance and not rejection by falsification. If it proves impossible to suggest the value claim target metric, the NHB should be disregarded as the basis for product assessment, pricing, and access.

Quite reasonably the formulary committee may consider the NHB not to be taken as the primary and only basis for claiming net benefit. The committee may insist on a measure of direct benefit to patients, not the clinical inference of benefit. This moves us from a focus on just clinical parameters as single attributes to the consideration of latent constructs, the role of needs fulfillment as the appropriate trait, and assessment of that trait in a Rasch modeled instrument.

It is, perhaps, fortunate that the ASCO value framework has had little traction in therapy choice and formulary decision-making [35]. Concerns have been expressed that the frameworks make little

sense to not only payers but also manufacturers and patients represented by the various cancer-related associations. While it might be claimed that they have 'potential', the fact is that they are analytical dead-ends. If we are focused on the patient voice, on the assessment by patients (not clinicians and other experts) of the benefit of a new therapy, then we need to focus on Rasch measurement to create disease-specific cancer instruments that create single attribute, linear interval measures.

CONCLUSIONS: A NEW PARADIGM FOR ONCOLOGY VALUE CLAIMS

To meet the demarcation test for distinguishing science from non-science, to distinguish sense from nonsense, and to separate science from pseudoscience, value claims must stand the test of falsifiability. Value claims for competing cancer therapies must have the property of being capable of being wrong. Value claims must be credible, evaluable, and replicable; credibility must recognize the limits of fundamental measurement. This is the starting point. When a new therapy is introduced specific to a cancer population, we must be capable of proposing a profile of value claims, supported by protocols for assessment, that establish the framework for resolving claims and set the stage for ongoing disease area and therapeutic class reviews; the discovery of new yet provisional new knowledge, together with options for outcomes-based contracting.

Falsifiability is not a new concept; it does back to the notion of a paradigm and the accumulation of value claims assessments that give continuing support to initial value claims and, if puzzles accumulate, the emergence of a new paradigm supporting the evolution, in Popper's terms, of objective knowledge [36]. In health technology assessment, in this case for cancer therapies, we have not got to first base in our assessment of patient benefit. Certainly, there may be support for claims with improved clinical outcomes and less toxicity, but that is, at best, only an inference, not a direct measure, of patient benefit.

After over 30 years of EORTC instrument development in oncology, we are faced with the uncomfortable conclusion that we have little if anything to offer to capture the patient's voice. In the case of EORTC, the most egregious failure is to neglect completely RMT and the application of Rasch rules to create measurement instruments from robust latent constructs to support the assessment of patient-relevant single attributes. A failure that is compounded by the aggregating of subjective Likert scores to support value claims. The failure of EORTC also lies in the absence of consideration of the framework for submission to the formulary committee where patient-centric claims are only one element in a profile of single attributes, each supported by an evaluation protocol.

There must be a recognition that all value claims, whether clinical, patient-centric or for drug utilization and resource use are provisional. Pricing and access to pharmaceuticals and other interventions have to be subject to value claim assessment; prior outcomes are not the basis for claiming future outcomes. We have to set the stage for ongoing value claim disease area and therapeutic class reviews, together with the option for specific value claims to support outcomes-based contracting. The focus for any submission to a formulary committee must be the insistence on single-value claims with the required linear, interval properties. This means the application of Rasch modeling to create instruments to meet those required measurement standards in non-

physical attributes. This means we have to avoid composite or bundled claims that attempt to create single metrics to support formulary decisions such as multiattribute generic measures.

Value claims for cancer therapies that rely on assumption-driven simulations with non-evaluable 'claims' for cost-effectiveness must be abandoned. These models fail the required standards for normal science and fundamental measurement. In the US this means claims for pricing and access created by ICER and in Europe, reference-modeled claims submitted by manufacturers too, as examples, the National Institute for Health and Care Excellence (NICE) in the UK and the Dutch National Health Care Institute (Zorginstituut Nederland or ZIN). This means that we abandon assumption-driven non-evaluable cost-effectiveness claims as well as efforts to populate these lifetime incremental cost-per-QALY models by mapping from cancer instruments to create generic utilities or preferences; the exercise is a waste of time. This does not mean that all cancer impact models are put to one side; we can still accept models that make evaluable claims that can be reported in a meaningful timeframe.

Cancer is no different from other disease states in the standards required for value claims that support initial formulary negotiations for pricing and access and ongoing disease area and therapeutic class reviews. Given the importance attached to capturing the patient's voice in cancer treatment, there is a clear need for an audit of all disease-specific cancer instruments or modules. The audit is easily initiated as we have checklists to assess the extent, if any, a disease-specific instrument meets Rasch standards as a stopgap for the investment in a single attribute, linear interval measure [37]. At the same time, the specific EORTC Quality of Life Group guidelines need to be rewritten to support the creation of disease-specific instruments that meet Rasch standards.

Once the imperative of RMT is recognized we can abandon the generic EORTC QLQ-C30 instrument, the CLU-C10D instrument, the EORTC cancer specific modules and attempt to create modules linked to the EORTC QLQ-C30. The focus must be on disease-specific instruments which yield both interval and by transformation approximate ratio measures. We must also abandon the ASCO net health benefit framework. At the same time, with the proposed new start in health technology assessment, a needs fulfillment measure would be one element in a profile of value claims including clinical endpoints defined by instruments that meet the standards of the physical sciences, claims for drug utilization, including drug switching and compliance claims and value claims for other aspects of resource utilization.

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] Kaasa S, Bjordal K, Aaronson N et al. The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. *Eur J Cancer*. 1995;31A(13-14):2260-63

[3] Wheelwright S, Bjordal K, Bottomley A et al. EORTC Quality of Life Group guidelines for developing questionnaire modules. 5th. EORTC, 2021

[4] Nolte S, Liegl G, Petersen M et al. General population normative data for the EORTC QLQ-C30 health quality of life questionnaire based on 15,386 persons across 13 countries, Canada and the United States. *Eur J Cancer*. 2019;107:153-63

[5] Schnipper L, Davidson N, Wollins D et al. American Society for Clinical Oncology Statement: A conceptual framework to assess the value of cancer treatment options. *J Clin Oncol*. 2015; 33(23): 2563-2577

[6] Andrich D, Marais I.A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences. Singapore: Singer, 2019

[7] Bond T, Zi Y, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences ($th Ed). New York: Routledge, 2021

[8] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved]. *F1000Research* 2020, **9**:1048 (https://doi.org/10.12688/f1000research.25039.1)

[9] Wright B, Linacre J. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. 1989; 70(12):857-60

[10] Tennant A, Conaghan P. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied and what should one look for in a Rasch paper. *Arthritis & Rheumatism (Arthritis Care & Research)*. 2007;57(8):1358-62

[11] McKenna S, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality. *J Med Econ*. 2020;23)10):1196-1204

[12] Wright B. Comparing Rasch measurement and factor analysis, *Structural Equation Modeling: A Multidisciplinary Journal*, 1996; 3:1, 3-24

[13] Rasch G. Probabilistic Models for some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

[14] Merbitz C, Morris J, Grip J. Ordinal scales and the foundations of misinference. *Arch Phys Med Rehab*. 1989;70:308-32

[15] Shih C-L, Chen C-H, Sheu C-F et al. Validating and improving the reliability of the EORTC QLQ-C30 using a multidimensional Rasch model. *ValueHealth*. 2013;16:848-54

[16] Giesinger J, Kuijpers W, Young T et al. Thresholds for clinical importance for four key domains for the EORTC QLQ-C30: physical functioning, emotional functioning, fatigue and pain. *Health Qual Life Outcomes*. 2016; 14:87

[17] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21(5):474-80

[18] McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes. 1:The search for the Holy Grail. *J Med Econ* 2019;22(6):516-22

[19] McKenna S, Heaney A, Wilburn J. Measurement of patient-reported outcomes. 2: Are current measures failing us? J *Med Econ*. 2019;22(6):523-30

[20] Brazier J, Peasgood T, Mukuria C et al. The EQ-HWB: Overview of the development of a measure of health and wellbeing and key results. *ValueHealth*. 2022;25(4):482-491

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

[22] King M, Viney R, Pickard A et al. Australian utility weights for the EORTC QLU-C10D, a multiattribute utility instrument derived from the cancer-specific quality of life questionnaire, EORTC QLQ-C30. *PharmacoEconomics.* 2018;36:225-38

[23] Kemmler G, Gamper E, Nerich V, Norman R, Viney R, Holzner B, King M. German value sets for the EORTC QLU-C10D, a cancer-specific utility instrument based on the EORTC QLQ-C30. *Qual Life Res.* 2019; 28(12):3197-3211.

[24] Nerich V, Gamper EM, Norman R et al French Value-Set of the QLU-C10D, a cancer-specific utility measure derived from the QLQ-C30. *Appl Health Econ Health Policy*. 2020. doi: 10.1007/s40258-020-00598-1. PMID: 32537694

[25] Gamper E, King M, Norman R et al. EORTC QLU-C10D value sets for Austria Italy and Poland. *Qual Life Res*. 2020;29: 2485-95

[26] Finch AP et al. Estimation of an EORTC QLU-C10D value set for Spain using a discrete choice experiment. *PharmacoEconomics*. 2021; 39(9): 1085-109

[27] Rowen D, Brazier J, Young T et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *ValueHealth*. 2011;14:721-31

[28] Mallinson T, Kozlowski A, Johnston M et al. Rasch Reporting Guidelines for Rehabilitation Research (RULER): the Ruler Statement. *Arch Phys Med Rehabil*. 2022;103(7):1477-86

[29] de Winckel A, Kozlowski A, Johnston M et al. Reporting Guidelines for RULER: Rasch Reporting Guidelines for Rehabilitation Research: Explanation and Elaboration. *Arch Phys Med Rehabil*. 2022;103(7):1487-98

[30] Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. *J Appl Meas*. 2002;3(3):325-59

[31] Beck A-J, Kieffer J, Retél V et al. Mapping the EORTC QLQ-C30 and QLQ-H&N35 to the EQ-5D for head and neck cancer: Can disease-specific utilities be obtained? *PLoS One*. 2019;14(12):e0226077

[32] Langley P. Mapping Impossible Utilities: The ICER Report on Tezepelumab for Severe Asthma. *Inov Pharm*. 2022;13(2): No. 1

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

[34] Russell B. The Problems of Philosophy. 1912

[35] Slomiany M, Madhavan P, Kuehn M, et al. Value frameworks in oncology: Comparative analysis and Implications to the pharmaceutical industry. Am Health Drug Benefits. 2017;10(5):253-260.

[36] Popper K. Objective Knowledge: An Evolutionary Approach (Rev. Ed.) New York: Oxford University Press, 1979

[37] Combrinck C. Is this a useful instrument? An introduction to Rasch measurement models in Kramer S, Laher A, Fynn et al (Eds.) Online readings in Research Methods. Psychological Society of South Africa. Johannesburg 2020