MAIMON WORKING PAPERS www.maimonresearch.net

MAIMON WORKING PAPER No. 2 FEBRUARY 2021

THE POVERTY OF MULTIATTRIBUTE UTILITY SCALES

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

Abstract

Presumably the purpose of a formulary submission is to argue for the clinical comparative efficacy of a new product or device against a standard of care. The options are: (i) generate the required evidence as part of a research program; or (ii) invent or manufacture evidence to support comparative claims. We know, over 30 years, that the first option has been rejected by leaders in the health technology field; possible agreements with a manufacturer for a research program to meet evidence gaps and support amendments to a provisional pricing agreement have been rejected. At best, there are attempts to support claims as part of value contracting. The choice has been to manufacture or invent claims for cost-effectiveness, pricing and access through imaginary simulation models. Unfortunately, a key element in this modeling has been the application of utility scores from multiattribute instruments (e.g., EQ-5D-3L/5L). These are a curious amalgam of symptoms and response levels which fail to meet the requirements of fundamental measurement. The individual symptoms or attributes they claim to include are just ordinal scores. Adding these together as part of a composite scoring algorithm confounds the problem because these attributes are each dimensionally unique. The result is a dimensionally heterogeneous amalgam that lacks construct validity. The utility scales are, from the perspective of fundamental measurement, a disaster.

INTRODUCTION

To those who recognize the constraints imposed by the axioms of fundamental measurement, the extent to which multiattribute utility measures has been unquestionably accepted by those developing claims for cost-effectiveness is quite disturbing. Alone among the social sciences, health technology assessment has embraced evidence of comparative product claims which is created not discovered [1]. Rather than the discovery of new provisional facts based on a coherent research program, technology assessment has, with the support of leaders in the field at world renowned academic institutions, opted to create through simulation modeling claims which are imaginary and which, not surprisingly, fail the standards of normal science [2].

This failure is clear cut: rather than underwriting a research program to discover new, yet provisional facts regarding response to therapy options, the leaders in health technology assessment opted for a creation of the evidence required to convince formulary decision makers. The leaders failed to appreciate the demarcation test to distinguish science from pseudoscience: claims must be credible, empirically evaluable and replicable [3]. This is no novel insight; it has been recognized since the scientific revolution of the 16th century.

Creating the evidence to support claims for competing therapies achieved the good housekeeping seal of approval in the commitment to approximate information through the construction of incremental cost-per-quality adjusted life year (QALY) simulation models. These clearly violated the standards of normal science; yet they had a willing audience. Central to these modeling exercises was the application of utility scores from multiattribute instruments such as the EQ-5D-3L/5L, the HUI Mk 3 and the SF-36 family of measures. Unfortunately, none of these utility measures meet the standards of fundamental measurement required if they were to be anything other than ordinal and dimensionally heterogeneous constructs that were not intended to measure response to therapy.

RECOGNIZED STANDARDS OF MEASUREMENT

It is surprising to those with knowledge of measurement standards why there has been ready acceptance to put these to one side in favor of imaginary claims driven by multiattribute utility scores that fail to meet fundamental measurement standards. In the physical sciences where accurate measurement is the sine qua non of empirical evaluations and reporting, an understanding of the axioms of fundamental measurement is accepted without question. More to the point, the measures found in the physical sciences are single attribute; they measure one trait at a time and are dimensionally homogeneous (unidimensional).

The situation is different in the social science where, for many participants, the notion of fundamental measurement is quite foreign. Little thought, if any, is given to the distinction between scores and measures. Scores are simply observational counts while measures are designed to conform to fundamental measurement axioms. The result is that a preponderance of scales has no intrinsic meaning; they are simply a product of ad hoc 'adding-up' with no thought given, particularly in assessing response to therapy with an explicit statement of the theory for making and recoding observations.

Following the formalization by Stevens and others in the 1930s and 1940s, the axioms of fundamental measurement were well understood [4]. The measurement scales are nominal, ordinal, interval and ratio. Each scale of measurement has one or more of the following properties: (i) identity where each value has a unique meaning; (ii) magnitude where values on the scale have an ordered relationship with each other but the distance between is unknown; (iii) invariance of comparison where scale units are equal to each other in an ordered relationship and known distance; and (iv) a true zero where no value on the scale can take negative scores. The implications for the ability to utilize a scale to support *arithmetic operations (and parametric statistical analysis) are clear cut. A nominal scale is just a set of unique meanings but nothing else (e.g., gender). An ordinal scale has identity and magnitude in an ordered relationship but we do not know the distance between the values (i.e., it cannot support arithmetic operations, only non-parametric statistical evaluations, modes and medians). An interval scale has known differences but no true zero and can support only addition and subtraction (i.e., it can change the point on an integer line but only relative to other points). A ratio scale can support the additional operations of multiplication and division because it has a true zero (i.e., it can change the point on an interval line relative to zero).

In the early 1960s a new type of fundamental measurement was introduced: probabilistic conjoint simultaneous measurement [5]. This new measure subsumed the existing fundamental measurement categories, providing a framework for identifying measurement structures in non-physical attributes. This provided the basis of going from ordinal to interval scales, becoming known as the Rasch model or Rasch Measurement Theory (RMT) where two attributes such as difficulty of a question and the ability of the respondent can be jointly evaluated to determine whether or not an interval scale measure might exist to capture a latent trait or attribute such as needs fulfillment quality of life as a measure of therapy response. Latent traits are not directly observed; only their outcomes which provides a basis for inferring the presence and amount of the latent trait. To achieve appropriate measures requires a deliberative process, not just allocating numbers to events. RMT does not replace statistical analysis, it precedes it. It is possible to apply Rasch assessment to existing disease specific instruments to assess the overall and item fit of the instrument to assess its appropriateness for measuring response, with possible item and scoring modifications using readily available software packages [6] [7].

MULTIATTRIBUTE SCALES

It is a mystery as to why, for over 40 years we have embraced multiattribute utility scales, such as the EQ-5D-3L when these measures have demonstrably failed to meet the required measurement standards of normal science. It is, as if, having developed these respective multiattribute instruments, the supporters have disappeared into a rabbit hole of denial and disbelief when faced with criticism.

A multiattribute scale is, from the perspective of the standards of fundamental measurement, nonsense. It is impossible to combine in a single score the various symptoms that comprise an instrument such as the EQ-5D-3L into a single score that makes any sense. Each symptom or attribute that has been selected by physicians, with no appreciation of the standards of fundamental measurement, to comprise the overall score (typically responded to on an ordinal scale) is an attribute with its own dimension(s). The result is a 'score' that lacks dimensional homogeneity (or unidimensionality); let alone the possible interactions between the listed symptoms. Clearly, if the focus is on deciding which symptom collection should be combined to create a quality of life score (not measure) then it is open season, together with choice of ordinal response levels.

At the same time, when the development of multiattribute scales is reviewed it becomes clear that no serious attention was given to the measurement properties of the various multiattribute scales, noting there is no agreed gold standard for the symptoms captured by the various scales. There is no evidence whatsoever that there was any appreciation of measurement theory and, most importantly, no Appreciation of the role of conjoint simultaneous measurement in the construction of instruments to measure latent traits, such as patient quality of life, which are common claims in health care outcomes.

Ignorance was no a barrier; the barrier was not even recognized. Defining quality of life in terms of a collection of clinical symptoms and response levels was a non-started from day one. It violated the

rules of fundamental measurement where different attributes with their own dimensionality were lumped together. The resulting observational scores, summarized by utility algorithms that attempted to combine these scores to fit to a 0 – 1 scale lacked not only construct validity but also yielded negative values for states worse than death. Respondents were asked to respond to symptoms of depression (undefined for the EQ-5D-3L questionnaire) on an ordinal scale: no problem some problems extreme problems with no attempt to consider what life experience and clinical manifestation defined the dimensions of depression experience.

The term 'multiattribute' is self-defeating. If the focus is on response to therapy then the relevant attributes should be identified. Some will involve agreed measures that meet the required measurement standards; other will be measures of latent traits. In the latter category the focus is on the patient where, as RMT emphasizes we should focus on the needs of the patient. Applying conjoint simultaneous measurement we can assess the likelihood of a patient responding successfully to an intervention by combing the difficulty of responding positively to an item with the ability of the patient to respond successfully. The likelihood of a high jump competitor achieving success combines the height of the bar and the inherent ability of the participant.

Let's summarize the case against the multiattribute measures:

- **They are ordinal 'add em' up' scores that lack dimensional homogeneity and construct validity**
- **As ordinal scores they are incapable of providing any coherent measure of response to therapy**
- **As ordinal scores multiattribute scales cannot support the creation of quality adjusted life years (QALYs)**
- **Claims that they are interval measures in disguise are absurd as there was no effort to create an interval scale**
- **Even if these scores had interval properties they cannot support the creation of QALYs**
- **Further claims that they are ratio scales in disguise are also absurd as they fail to have a true zero with negative scores and no effort was made to create a ratio scale**
- **Reliance on ordinal scores effectively destroys simulation modeling to support incremental cost-per-QALY imaginary claims for pricing and product access**

If you truly believe that quality of life is measured in purely clinical terms with instruments designed to meet standards of fundamental measurement, then these physical attributes should be measured and reported separately; they may be combined later (e.g., body mass index) if they have the required measurement properties. If you believe there are non-physical latent attributes that also contribute to quality of life (e.g., psychological states) then these need to be measured – but this requires the tools of conjoint simultaneous measurement where RMT provides the closest approximation to an 'ideal' interval score. A ratio score is outside the scope.

A NEW PARADIGM

The manifest deficiencies of multiattribute utility scales means we should relegate them to the nearest dumpster. They contribute nothing to our assessment of comparative therapy response and the resolution of patient needs; nor are they meaningful measures at the population health level. It's a puzzle that they have survived so long. This is true in particular with the EQ-5D-3L/5L instruments with 5 symptoms and three/five response levels. They fail on all criteria noted above yet continue to be marketed, employed in clinical trials and used to support the I-QALY in simulation models. In the US, the institute for Clinical and Economic Review (ICER) continues to utilize these measures while admitting that they cannot provide proof that supports their belief that these measures have ratio properties [8]. The unfortunate consequence, at least from the patient and caregiver perspectives is that the use of these multiattribute measures to create estimates of incremental I-QALYS virtually guarantees minimum modeled I-QALY gains and hence substantial price discounts to achieve threshold values.

If the focus of technology assessment is on response to therapy within disease-specific patient groups, then the obvious answer is to develop patient-centric measures of selected response(s) rather than crudely constructed and non-responsive generic measures. In previous commentaries, as well as the recently released Version 3.0 of the Minnesota proposed formulary guidelines the case has been made that we should follow the recognized measurement standards in the physical science and commit to single attribute measures to support claims for product performance that are credible, empirically evaluable and replicable [9]. Failure to justify the measures proposed for therapy response leads to irrelevance.

If we believe that life gains its quality in terms of meeting clinical outcomes, then the requisite outcome(s) must be identified, justified and separately measured. If we believe that life gains its quality through meeting the needs of patients and caregivers, then this latent trait needs to be identified, justified and, if possible, measured.

Finally, we need to focus on claims for products within disease areas. Generic measures that attempt to provide some central planning focus for resource allocation within health care systems collapse because we have no coherent basis for believing in them. There is nothing beyond the I-QALY; it is an analytical dead end.

CONCLUSIONS

The cult, there is no better word, of technology assessment driven by imaginary and non-evaluable claims to demand pole position, through groups such as ISPOR and ICER, in technology assessment is simply ridiculous. For 30 years or more leaders in the field have promoted the superiority of approximate simulated information to support claims for cost-effectiveness over accepted scientific standards. Rejecting hypothesis testing they overlooked the absence of any scientific foundations for the application of ordinal utility scores to support the I-QALY. It is not, in fact, the question of overlooking the importance of fundamental measurement, but an ignorance of measurement requirements. To a wider audience, this lack of awareness cuts to the core of technology assessment

as a discipline. Why should formulary committees pay any attention to claims from imaginary simulation advocates such as ICER when past decisions involving ICER recommendations fail the standards of normal science?

REFERENCES

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

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

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

[4] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-680

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

[6] Christensen KS, Oernboel E, Nielsen MG et al. Diagnosing depression in primary care: A Rasch analysis of the major depression inventory. *Scand J Primary Health Care*. 2019;376(1):105-112

[7] Robinson M, Johnson AM, Walton D et al. A comparison of the polytomous Rasch analysis output of RUMM2030 and R(ltm/eRm/TAM/lordif). BMC Ed Res Methodology. 2019;19:36

[8]Langley PC. To Dream the Impossible Dream: The commitment by the Institute for Clinical and Economic Review to rewrite the axioms of fundamental measurement for Hemophilia A and Bladder Cancer value claims. *InovPharm*. 2020; 11(4); No. 22 https://pubs.lib.umn.edu/index.php/innovations/article/view/3542/2613

[9] Langley P. Value Assessment, Real World Evidence and Fundamental Measurement: Version 3.0 of the Minnesota Formulary Submission Guidelines. *InovPharm*. 2020;11(4): No 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/3542/2613