MAIMON WORKING PAPERS No. 21 OCTOBER 2022

POST-MENOPAUSAL QALITY OF LIFE CLAIMS: OVERLOOKING THE REQUIREMENTS OF FUNDAMENAL MEASUREMENT IN ICER'S COST-EFFECTIVENESS ASSESSMENT OF FEZOLINETANT FOR MODERATE TO SEVERE VASOMOTOR SYMPTOMS

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

ABSTRACT

One of the signal failures in health technology assessment is the absence of consideration given, not only to the standards of normal science, but to those of fundamental measurement. A recent draft evidence report by the Institute for Clinical and Economic Review (ICER) is emblematic of this failure. Based on a simple linear regression model that translates aggregate scores from the ordinal Menopause-specific Quality of Life Questionnaire (MENQOL) to the ordinal EuroQol EQ-5D-5L, ICER has applied these scores to an assumption driven model simulation to produce preferences, QALYs and incremental cost-per-QALY claims for fezolinetant for moderate to severe symptoms associated with menopause. Unfortunately, the attempt to crosswalk ordinal scores is mathematically absurd. The 'created' EQ-5D-5L preferences are meaningless. The result is that the ICER modelled claims for cost-effectiveness are impossible. Although it might be possible to assess certain domains of the MENQOL for their approximation to an interval score with the application of the Rasch Rating Scale Model, this will not support quality of life claims. A preferred approach would be to consider an alternative latent trait for quality of life in menopause, applying Rasch Measurement Theory, to develop a polytomous instrument that has the required measurement properties.

Keywords: MENQOL, Fezolinetant, ICER imaginary claims, failed EQ-5D-5L crosswalking

INTRODUCTION

The recent publication of the draft report by the Institute for Clinical and Economic Review (ICER) to assess the value and effectiveness of fezolinetant (Astellas Pharma) for moderate to severe symptoms associated with menopause, while the drug is under FDA review, raises a number of serious concerns [1] . Principally, that the modelled claims for quality of life (QALYs) are mathematically impossible; the mapping function used to translate to EQ-5D-5L (EuroQoL) preferences from the Menopause-Specific Quality of Life questionnaire (MENQOL) fails to meet the standards of Rasch or modern measurement theory that have been recognized and accepted for over 50 years. This is not the first time that ICER and its consultants who are responsible for construction of its assumption driven simulation models have failed to recognize fundamental measurement standards; indeed, despite these standards, ICER continues to maintain that the EQ-

5D-3L/5L preference scores are actually ratio scores [2 3]. This is patently untrue [4 5]. The fact is that the EQ-5D-5L and other multiattribute generic preference instruments fail to meet Rasch measurement standards for reliability, invariance, additivity of the latent trait, unidimensionality and order [6]. To this failure we must add the fact that the ICER models fail the standards of normal science; judged by the demarcation criterion, they are non-science in not providing credible, evaluable and replicable claims for product value [7].

The purpose of the present brief commentary is not to go over arguments that have been presented on a number of previous occasions, primarily in this *Journa*l, for required measurement standards but to focus on the MENQOL instrument. To demonstrate that its failure, both as a measure of quality of life as is claimed, but also its irrelevance as the basis for a crosswalk or mapping algorithm to support assumption driven cost-per-QALY simulations for imaginary cost-effectiveness claims.

STANDARDS FOR VALUE CLAIMS

The needs of patients, physicians and health system decision makers are not met if there is a failure to recognize the evidence requirements for therapy claims post-menopause. As noted in previous publications there are three requirements for any therapy impact claim: (i) the claim must refer to a single attribute that is credible, evaluable and replicable; (ii) the claim must meet ratio or interval measurement standards; and (iii) the claim must be accompanied by an evaluation and reporting protocol. These standards are in marked contrast to those that support the current standard in HTA for assumption driven modelled imaginary claims [8 9] .

In the present case, this means that value claims for fezolinetant must be expressed, not as a non-empirically evaluable blanket claim for cost-effectiveness, but in terms of attributes that meet these standards, whether they are for clinical outcomes, patient reported outcomes and drug and resource utilization. None of the ICER modelled outcomes meet these standards. In addition, the latest systematic review of cost-effectiveness studies in menopausal hormone therapy puts to one side recognition of the standards of normal science and fundamental measurement in its assessment of five studies, all of which used a cost-per-QALY assumption driven model following the CHEERS reporting guidelines [10]. It is worth noting that the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) recommended rejection of traditional evaluation of claims with hypothesis testing in favor of approximate information to support economic evaluations in formulary decisions [11] .

The draft ICER report, in following the traditional HTA model, creates assumption driven outcome claims that are not designed to be empirically evaluated; perhaps they are intended to be helpful (or not) for decision makers. If so, it is difficult to see attention being given when any number of assumption-driven competing modelled claims are trivially easy to produce. As it stands, the ICER base case imaginary results, credit fezolinetant with an annual placeholder price of $6,000 and a total discounted cost of $200,000, with discounted QALYs of 16.43, compared to 16.33 QALYs for non-pharmacologic treatment. The cost-per-QALY gained versus the no treatment comparator was $390,000 with only a 14% likelihood of being cost-effective after modelling with probabilistic sensitivity analysis. The claims fail the demarcation test; they are non-evaluable and assumption driven and must be categorized as non-science.

THE ORDINAL MENQOL

The MENQOL was introduced in 1996 as a tool to assess health-related quality of life in the immediate post-menopausal period [12]. The MENQOL is a multi-domain instrument. Rather than consider latent traits or attributes that may be relevant to the response of post-menopausal patients to therapy interventions, including the question of whether the needs of these patients are being met, the MENQOL proposes to assess the quality of life in terms of 29 items in a Likert-format capturing patient-reported symptoms experienced in the preceding month: vasomotor (items 1–3), psychosocial (items 4–10), physical (items 11–26), and sexual (items 27–29). Items pertaining to a specific symptom are rated as present or not present. If the symptom is present it is scored on a zero (not bothersome) to six (extremely bothersome) scale. Non-endorsement of an item is score I; endorsement a 2. Each domain is scored separately, with subject responses converted to a composite mean range 1 to 8 (endorsement score plus Likert integer value). The overall questionnaire score is a mean of the domain items.

Failure to appreciate the limitation of fundamental measurement means that these scores are meaningless. Apart from the adding in of the bothersome/non-bothersome score, the Likert integers in their traditional scale data summation is based on two a priori assumptions: all Likert items are of equal difficulty and that the thresholds between steps are of equal distance or equal value [13]. In other words, the MENQOL scoring fails what has been described as the Rasch or modern measurement quality control test. Judged by the Likert items separately scored, the scores on those items are, in the absence of application of the Rasch Rating Scale Model for polytomous Likert-type data to assess measurement properties, the score has to be regarded as ordinal and not interval or ratio data [14 15] The addition of the bothersome/non-bothersome score creates additional confusion, possibly best described as an ‘adjusted’ ordinal scale. It is worth noting that one recent study of the MENQOL claimed it had, through Rasch analysis, acceptable psychometric properties with factor analysis indicating six domains [16]. Unfortunately, Rasch analysis was misapplied to the overall MEMQOL ordinal score, rather than considering the Likert-properties of the instrument and applying the Rasch Rating Scale Model and the criteria for approximation to a single attribute interval scale, capturing the unidimensionality of the quality of life latent trait.

Following the standards for fundamental measurement first proposed by Stevens in 1946, an ordinal scale cannot support standard arithmetic operations; this is achieved for addition and subtraction for interval scales (invariance of comparison and an arbitrary zero) and ratio scores (all operations and a true zero) [17]. Ordinal scores only support non-parametric statistics (which as an ‘adjusted’ ordinal score the MENQOL cannot support) to create median and modal values; averages are disallows together with any measures of dispersion. What is doubly unfortunate is that if the authors of this instrument had been aware of the standards for Rasch measurement, known since the 1960s with the application of conjoint simultaneous measurement in instrument development, together with the limitation of fundamental measurement, an acceptable instrument for the quality of life in menopause might have emerged. As it is the instrument is used widely by those, presumably, who are unaware of the measurement requirements of polytomous scales. This is in contrast to instruments focus on needs in quality of life have been developed to meet Rasch standards [18].

Despite its undoubted popularity and increasing use over the past 25 or more years, the fact that as a polytomous multi-domain instrument, no one questioned its measurement properties (or their absence) including authors of systematic reviews [19] . Psychometric evaluations of instruments are accepted but only after the measurement properties of the instrument have been evaluated; the instrument must have a demonstrated interval or ratio score. Claims, therefore, for the psychometric properties of the MENQOL are premature and irrelevant.

THE FAILURE OF CROSSWALKING ORDINAL SCORES

If the intent is to crosswalk or map from one scale to another, as the basis for establishing claims for one scale when it is absent but there are responses to the other scale, then the crosswalking algorithm should meet two essential properties. It should be created from two instruments administered to the same target patient population where the two instruments are designed to capture the same latent construct as a single attribute and the instruments have interval or ratio measurement properties. The easiest example is the crosswalking algorithm for translating temperature measured in the fahrenheit scale to the equivalent centigrade scale, together with the algorithm from translating centigrade to the Kelvin scale. In the former case both unit temperatures are on an interval scale, with an arbitrary zero and can take both negative and positive values. In the latter case the Kelvin scale captures absolute temperatures with a true zero (absolute zero Kelvin -273.15). Crosswalking can be between interval scales and ratio scales and between interval and ratio scales. To do this we have to demonstrate the scales meet the required standards of fundamental measurement. As example, the translation from fahrenheit to centigrade is simply $^0C = (^0F - 32) * 0.56$ while for the Kelvin scale translating to centigrade is $^0C = K – 273.15$. Note that temperature is a latent construct and, over some 3 centuries, we successively developed instruments to capture a measure of the latent construct [20] . A true zero for a ratio scale means the absence of the latent construct; a ratio scale cannot take negative values. Exactly the same argument applies in patient reported outcomes where we may have an attitudinal latent construct but require a technique for a measure of that latent construct; hence the place of Rasch Measurement Theory (RMT). The issue to address is whether we can, given RMT, translate a latent to a ratio measure, because to create QALYs (if it is important) the preference score must be in the form of a bounded ratio scale [21].

In the ICER report, the crosswalk to translate MENQOL scores to create the EQ-5D-5L score is:

$$\text{EQ-5D-5L} = 0.992 – 0.042 * \text{MENQOL}$$

The fundamental error associated with this ordinary-least squares regression model, although it should be noted that the fit is poor with a reported $R^2 = 0.347$ and root mean squared error of 0.093, is the fact that both the EQ-5D-5L and MENMQOL are ordinal scores; the both fail to meet Rasch measurement standards [22]. This means that crosswalking using a regression model is disallowed; no attempt was made to demonstrate that the scores were interval or ratio, just the assumption, which is incorrect, that the MENQOL score is a continuous variable; in fact, it has neither ratio not interval properties. The MENQOL is just a summation of scores which have no discernible properties to support mean values by domain and average of domain means,

Once the inadvisability of believing that crosswalking between ordinal scores is allowed, is admitted, the entire ICER modelling exercise collapses. The contrived EQ-5D-5L preferences, and consequent creation of mathematically impossible QALYS and assumption driven imaginary comparative cost-per-QALY claims is impossible. There is, presumably, a limit on how far-fetched assumptions can be to support a cost-per-QALY model that fails accepted standards. Unfortunately, given ICER's commitment to cost-per-QALY assumption driven simulations, there has to be a source, however whimsical, of multiattribute generic preferences (preferably the EQ-5D-3L/5L) to justify the model claims. In this case the choice was unwise, irrespective of the model framework not meeting the standards of normal science.

AFTER THE MENQOL

The MENQOL is not the only instrument that has been developed to assess the symptom burden and quality of life in perimenopausal and post-menopausal patients (the so-called climacteric syndrome) [23]. It has not been the intent here to review these instruments, although the assessment of the MENQOL has made clear the assessment standards that should apply. Among the other instruments that have been developed are: (i) the Menopause Symptoms Treatment Satisfaction Question (MS-TSQ); (ii) the Kupperman Index (KI) [24]; (iii) the Menopause Rating Scale (MRS) [25] ; and (iv) the Greene Climacteric Scale [26] . All are polytomous Likert-based instruments with multiple integer scored response options. None have been assessed for Rasch measurement properties (e.g., Rasch Rating Scale Model) for an approximation to an interval score, with the various authors and commentators assuming that the ordinal integer-based summation scale has properties to support classical statistical analysis, which is incorrect as shown by the Tao et al study [27]. If the objective is to measure therapy response, then the MENQOL should not be included as a criterion in clinical trials, although it is used, for example, in the REPLENISH study (NCT 01942668) for the evaluation of estrogen plus progesterone oral capsule (TX-001HR) [28].

Given the popularity of the MENQOL, including a commitment to a range of language versions, the pertinent question is whether the MENQOL has a future, if it is to escape the Likert ordinal summation problem. One avenue would be to apply the Rasch Rating Scale Model for polytomous data to specific domains of the MENQOL. This could provide the basis for evaluating the extent to which the MENQOL has properties that approximate to an interval scale. This would not apply to all domains captured by the MENQOL, which leaves the MENQOL as an inappropriate instrument. There are a number of readily available software packages which could be applied o evaluate responses to the MENQOL domains (but not the add-on bothered/not bothered conversion scores). The result is that the MENQOL would have to be put to one side; the current version is not sustainable as the basis for assessing response to therapy, let alone quality of life. It fails the essential requirements for RMT.

If quality of life is considered a required outcome for evaluating therapy response, the preferred route would be not to continue to prop-up the MENQOL but to follow the RMT framework for polytomous instruments and consider the appropriate latent construct. This could involve the needs-fulfillment holistic trait with conjoint simultaneous measurement applied to instrument development. This would create, if a measure of the latent construct was achieved, an interval measure that met the required properties; but it would not be a bounded ratio scale. It could not support QALYs or ICER type assumption driven imaginary claims.

CONCLUSIONS

The commitment over some 25 years to the MENQOL points to a failure, across the board, to understand the standards and limitations imposed by fundamental measurement. The MENQOL is a poor choice, and one that should not be made, as a vehicle for assessing quality of life for post-menopausal patients. The failure of the instrument is such that there seems no basis for its acceptance for ongoing assessments of therapy impact; aside from the failure of crosswalking and the promotion of ersatz EQ-5D-5L ordinal preference scores to support assumption driven simulations for imaginary claims by ICER for cost-effectiveness.

But the MENQOL is not alone; all other instruments designed to capture menopausal symptoms suffer from the same weakness. None meet, or have been assessed for, fundamental measurement or Rasch properties. They are, by default, ordinal scores which, lacking invariance, cannot capture response to therapy. This is in marked contrast to the area of rehabilitation medicine where there has been a long-standing commitment to Rasch measurement and, most recently, guidelines proposed for Rasch reporting (RULER) [29] [30].

REFERENCES

[1] Beaudoin FL, McQueen RB, Wright A et al, Fezolinetant for Moderate to Severe Vasomotor Symptoms Associated with Menopause: Effectiveness and Value; Draft Evidence Report. Institute for Clinical and Economic Review, October 11, 2022.

[2] Langley P. Concerns with Patient Reported Outcome Measurement and Value Claims for Therapy Response: The Case of Mavacamten and Symptomatic Hypertrophic Cardiomyopathy (SHCM). *InovPharm*. 2022;13(2): No. 16

[3] Langley P. Peter Rabbit is a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review in Health Technology Assessment. *InovPharm*. 2021; 12(2): No.20

[4] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved] *F1000Research* 2020, 9:1048

[5] McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting Patient Reported Outcome Instruments and Interpreting the Data they Produce. *InovPharm*. 2021; 12(2): No. 17

[6] Combrinck C. Is this a useful instrument? An introduction to Rasch measurement models, in Kramer S et al (eds.) Online Readings in Research Methods. Psychological Society of South Africa. Johannesburg, 2020

[7] 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

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

[9] Langley P. Facilitating bias in cost-effectiveness analysis: CHEERS 2022 and the creation of assumption-driven imaginary value claims in health technology assessment [version 1; peer review: 2 approved]. *F1000Research* 2022, 11:993

[10] Velentis L, Salagame U, Canfell K. Menopausal hormone therapy: a systematic review of cost-effectiveness evaluations. *BMC Health Ser Res*. 2017;17:326

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

[12] Hilditch J, Lewis J, Peter A et al. A menopause-specific quality of life questionnaire: development and psychometric properties. *Maturitas*. 1996;24(3):161-75

[13] Bond T, Yan Z, Heene M. Applying the Rasch model: Fundamental Measurement in the Human Science (4th Ed.) New York: Routledge, 2021

[14] Andrich D, Application of a rating model to ordered categories which are scored with successive integers. *App Psych Measure*. 1978;12(4):581-94

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

[16] Gazibara T, Kovacevic N, Nurkovic S et al. Menopause-specific Quality of Life Questionnaire: Factor and Rasch analytic approach. *Climateric* 2019;22(1):90-96

[17] Stevens S. On the Theory of Scales of Measurement. *Science*, New Series. 103: No. 2684 (Jun. 7, 1946:677-680

[18] Doward L, Wilburn J, McKenna S. Development and validation of the Bowel Cleansing Impact Review (BOCLIR). *Frontline Gastroenterology*. 2013;0:1-8

[19] Sydora B, Fast H, Campbell et al. Use of the Menopause-Specific Quality of Life (MENQOL) questionnaire in research and clinical practice: a comprehensive scoping review. *Menopause*. 2016; 23(9):1038-51

[20] Chang H. Inventing Temperature: Measurement and Scientific Progress. New York: Oxford University Press, 2004

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

[22] Coon C, Bushmakin A, Tatlock S et al. Evaluation of a crosswalk between the European Quality of Life Five Dimension Five Level and the Menopause-Specific Quality of Life questionnaire. *Climateric*. 2018;21(6):566-73

[23] Sourouni M, Zangger M, Honermann L et al. Assessment of the climacteric syndrome: A narrative review. *Ann Gynecol Obstet*. 2021;304(4):855-62

[24] Kupperman H, Blatt M, Wiesbader H et al. Comparative clinical evaluation of estrogenic preparations by the menopausal and amenorrheal indices. *J Clin Endocrinol Metab.* 1953;13:688–703

[25] Hauser G, Huber I, Keller P et al. Evaluation of climacteric symptoms (Menopause Rating Scale) *Zentralbl Gynakol.* 1994;116:16–23 [German]

[26] Greene J. Constructing a standard climateric scale. *Maturitas*. 2006;61(1-2):78-84

[27] Tao M, Shao H, Li C et al. Correlation between the modified Kupperman Index and the Menopause Rating Scale in Chinese women. *Pat Pref Adher*. 2013;7:223-29

[28] Constantine G, Revicki D, Kagan R et al. Evaluation of clinical meaningfulness of estrogen plus progesterone oral capsule (TX001HR) on moderate to severe vasomotor symptoms. *J North American Menopause Soc.* 2018; 26(5):513-19

[29] Mallinson T, Kozlowski A, Johnston M et al. Rasch Reporting Guideline for Rehabilitation Research (RULER): the RULER Statement. *Arch Physical Med Rehab*. 2022; 103:1477-86

[30] Van de Winckel A, Kozlowski A, Johnston M et al. Reporting Guideline for RULER: Rasch Reporting Guideline for Rehabilitation Research: Explanation and Elaboration. *Arch Physical Med Rehab*. 2022;103:1487-98