MAIMON WORKING PAPERS No. 3 March 2023

POST-MENOPAUSAL QALITY OF LIFE CLAIMS: OVERLOOKING THE REQUIREMENTS OF NORMAL SCIENCE AND 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 multidimensional or multiattribute ordinal scores is mathematically impossible. The 'created' EQ-5D-5L preferences are, as a result, of no interest. The overall result is that the ICER modelled claims for cost-effectiveness fail the required standards for normal science and fundamental measurement. fundamental are impossible. This is unfortunate, 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 (RMT), to develop a polytomous instrument that has the required measurement properties. The purpose of this commentary is to point out, as a number of previous commentaries have done, that this framework for creating assumption driven simulated modelled claims has no role in decisions for product assessment, access to formulary and pricing. This commentary expands upon these previous commentaries in placing RMT is the context of a needed paradigm shift to support the evolution of objective knowledge. This is critical if we are to understand, from the individual's perspective, not only an accurate assessment of the burden of menopause but to see this as part of an on-going research program that has to rely on fundamental measurement.

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

INTRODUCTION

The recent publication by the Institute for Clinical and Economic Review (ICER) of the evidence report to assess the value and effectiveness of fezolinetant (Astellas Pharma) for moderate to severe symptoms associated with menopause raises a number of concerns [1]. These stem from both the required standards of normal science and fundamental measurement but, possibly most significantly, the contribution, if any, that assumption driven lifetime simulation models contribute to what Popper describes as the evolution of objective knowledge. If we are to commit to long-term research programs, rather than one-off modelling exercises as in the present case if the ICER fezolinetant report, we must recognize that while measurement is a necessary condition for scientific investigation, it will never be sufficient without a substantive theoretical orientation. This brings us to Rasch theorizing, which is not just about measurement per se but with the theorizing that must precede the application of Rasch Measurement Theory (RMT) to assess the credibility of a latent construct, preliminary item or instrument development to capture the manifestation of interest and the application of Rasch rules to guide establishing single attribute, unidimensional, linear, interval, additive and invariant instruments to support claims; a measure that can claim that we have an acceptable application for an approximate interval measure. This is not speculation; we have had the tools for the past 60 years to achieve this goal with widespread application in education, to a less extent psychology, and economics; but not in health technology assessment (HTA). if we are to commit to a long-tern, evolutionary research program, we must envisage a paradigm shift in HTA to support a commitment to the standards of normal science and fundamental measurement; this is task that has hardly begun.

The purpose of this commentary is to consider the ICER fezolinetant report as evidence for the lack of appreciation of the lack of relevance, at least in the quest for objective knowledge, of assumption driven, simulated modelled claims for lifetime cost-effectiveness recommendations. We will not be revisiting arguments that have been presented on a number of previous occasions, primarily in this *Journa*l, for required Rasch measurement standards, most recently in a commentary on value claims in hemophilia [2] review 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 ordinal cost-per-QALY simulations for imaginary cost-effectiveness claims.

OBJECTIVE KNOWLEDGE

For Popper objective or evolutionary knowledge *is an objective evolutionary process which involves the creation and promulgation of new problem-solving theories, which are then subjected to the challenge of criticism, modification, elimination and replacement* [3]. For Popper, theory starts with problems not with observations, hence his distinction between theories that are couched in terms of confirmation (e.g., psychoanalysis) and must be rejected in favor of those with testable implications which, if false, would in turn, have falsified the theory (e.g., general relativity). It should be noted that Popper was not a dogmatic positivist. Popper does not deny that nonscientific

theories may be enlightening or that purely mythogenic explanations may expedite a deeper understanding of the nature of reality; but these are only steps in the evolutionary process for the creation of provisional knowledge. It is, of course, of interest as to whether or not we would describe ICER HTA modelling for non-evaluable cost-effectiveness claims as mythogenic; that is, capable of producing myths, not objectively true stories, supporting a deeper insight into cost-effectiveness claims.

Taking falsification as the demarcation criteria separating science from non-science, value claims that support falsification must be compatible with empirical measurement which means that for value claims to be recognized they must meet standards for fundamental measurement: single attributes measured in interval or ratio terms. Measurement is critical to our acceptance of value claims for therapy impact claims, but more importantly measurement is a fundamental input to the evolution of objective knowledge. This point is made clear, in the application of RMT to patient reported outcomes: observations, countable events, are always ordinal while measurement must be based on the arithmetic properties of interval scales [4].

STANDARDS FOR VALUE CLAIMS

The standards for normal science where value claims are credible, evaluable and replicable can only be applied to single attribute linear unidimensional measures where RMT provides the necessary and sufficient means to create interval measures from ordinal counts for patient reported outcome capturing patient response ability and item difficulty. Interval measurement is the requirement of fundamental measurement and one that, in the physical sciences has supported the pursuit of objective knowledge. The problem to be faced is that in HTA patient centric value claims are based, either directly or indirectly on ordinal scales; this is the hurdle we have to overcome. A hurdle that is made the more difficult by the applications of assumptions in lifetime simulation modelled 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 belief of the central role of assumption driven modelled imaginary claims [5 6] .

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 or 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 [7]. It is worth noting that the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has recommended rejection of traditional evaluation of claims with hypothesis testing in favor of approximate information to support economic evaluations in

formulary decisions [8] . In the HTA context, the term approximate information is undefined; even if the sub-text is assumption driven information. There is no reference point for either notion.

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. These outcome claims are entirely imaginary and fail the demarcation test; they are non-evaluable and assumption driven and must be categorized as non-science.

SIMULATIONS, ORDINAL SCORES AND MODEL ASSUMPTIONS

The weaknesses of the ICER framework for evaluating therapy options with the application of assumption driven simulated model claims are well established and have been extensively reported. In essence, they come down to the key requirement: any commitment to a long term, substantive research program must recognize the requirements and limitations of fundamental measurement. The ICER simulation do not and cannot support a substantive research program for the evaluation of new therapies in any disease area. Once this is accepted, we can go no further, despite possible inferences from mythogenic claims. The point emphasized by RMT is that interval measurement precedes statistical analysis. Unless we can demonstrate that the various input measures have interval or ratio properties as single attributes, the resulting outcomes will remain ordinal scales with no possibility of assessment irrespective of the time line involved. Basing modelled claims ion ordinal inputs is to invite failure. This does not deny modelling; only that assumption driven models must meet measurement standards and provide meaningful empirical value claims; this a requirement and not an option in modeling value claims.

Assumption driven simulations, as has been extensively reported, cannot be defended on the grounds that the assumptions are ‘realistic’. This denies to problem of induction and confirmation of past claims as claims on the future. There is a simple logical argument: the fact that past futures have resembled past pasts does not mean that future futures will resemble future pasts. Simulation models, built on dozens of assumptions dredged from the literature and even expert opinion cannot be seen as justification for non-evaluable claims on the future; as evidenced by all ICER evidence modelled reports.

But the failure is more-deep seated in the reliance on ordinal scores. In the past 30 years there have been in excess of some 20,000 papers indexed by PubMed that report the application of QALYs to create incremental cost-per-QALY claims for cost-effectiveness. The error is fundamental: the preference scores created to support QALYS from generic instruments such as the EQ-5D-5L are ordinal; there was no intent to create interval single attribute preference measures let alone the application of RMT. The ICER simulation collapses. Needless to say, by design, the simulation was not, and could not, produce evaluable value claims for a lifetime assumed behavior of a

hypothetical population. Even so, despite continued and ongoing criticism of the misapplied role of assumption driven simulations, ICER persists; after all it is their business model.

THE ORDINAL MENQOL

The MENQOL was introduced in 1996 with the objective of determining quality of life differences between menopausal women and to measure changes in their quality of life [9 10]. Currently, there four versions available for 1 week and 1 month recall periods and intervention versions where adverse events could negatively impact quality of life. The MENQOL was developed from an initial 106 symptom item assessment of women 2-7 years post-menopausal with a uterus and not on hormone replacement therapy. Items were reduced to 30 using the importance or propensity score method [11]. The MENQOL is a multi-domain instrument. Rather than considering 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.

The MENQOL falls at the first hurdle: it fails to recognize that if we are to calibrate response to any therapy intervention, the instrument that is applied must be capable of generating a single attribute, linear, unidimensional interval measure. The MENQOL views quality of life in multidimensional symptom terms; it lacks dimensionality and dimensional homogeneity. Calibrating the relative importance of symptoms to create a multi-domain instrument with a separate quality of life item fails the standards for RMT. The basic flaw is the failure to present a credible latent construct or trait to capture quality of life as an abstract or hypothetical entity. The latent construct or trait is not measured directly, but indirectly through its manifestation of the property of interest. The investigator should identify a latent construct or trail that is of interest and use that construct as a guide to operationalizing those manifestation of the latent trait or construct. If the trait focuses on quality of life defined in terms of post-menopausal patient needs and their fulfillment, we required an assessment procedure, a set of observations or items, to manifest that property. In terms of the construct validity of an instrument the question to be addressed is whether measured behaviors are expressions of that construct. The need to meet the requirements of fundamental measurement is overlooked; the quality control of RMT is missing.

Judging quality of life in terms of symptom experience says nothing about the needs of post-menopausal women and the extent to which their needs are met. Certainly, symptom assessments can be an input to quality of life, but they may be only peripherally related to needs. The Rasch model provides a framework which brings needs in terms of the ability of the respondent and the difficulty of the question into focus in the progression from counting observations to measurement.

Applying propensity scoring to calibrate the 'difficulty' of an item is only part of the instrument assessment process; following Rasch, we have to capture the interaction between item and the respondent. The importance for therapy response, or just the distribution of needs fulfilled, is that the Rasch equivalent of the MENQOL instrument is the invariance requirement where the measure retains its quantitative calibration irrespective of respondent or location and the importance of capturing the interaction between the respondent and the questionnaire item. Both are absent in the MENQOL which, in Rasch terms, fails to proceed beyond ordinal counts.

The failure to go beyond ordinal counts is exemplified in the scoring algorithms applied to the MENQOL. 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 the Likert items must be of equal difficulty and the thresholds between steps are of equal distance or equal value [12]. In other words, the MENQOL scoring fails what has been described as the Rasch or modern measurement quality control test. This is entirely expected; summation of integer Likert scale values is common in disease specific quality of life and other response scales, pricing only ordinal scores. Absent Rasch measurement we are left with observations or counts which are always ordinal; we must recognize that meaningful measurement is based on the arithmetical properties of interval scales · Unfortunately, although advised on many occasions through public comments on draft evidence reports, ICER rejects the Rasch claims for single attribute interval scores, insistent (with no evidence or proof) that generic instruments such as the EQ-5D-3L/5L preference scores are not ordinal scores but actually ratio scores [13 14]. This is patently untrue [15 16]. 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, let alone the composite nature of the use of health states [17 18].

Failure to appreciate the limitations of fundamental measurement means that the entire MENQOL construct and the attendant scores cannot be considered measures. The issue is straightforward: if there is a requirement for response to a polytomous instrument, the we have techniques for apply Rasch rules that create approximations to interval scales. These are the Rasch Rating Scale Model where all items have the same threshold structure, and the Partial Credit Rasch Model that relaxes this threshold requirement; they have been standard tools in Rasch modeling for the last 40 years to support interval measures.

Judged by Rasch measurement requirements, which combine respondent ability and item difficulty, the MENQOL scores are nothing more that ordinal scales; the score has to be regarded as ordinal and not interval or ratio data [19 20]. 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 [21]. While attempting *ex post facto* to apply Rasch criteria to an existing instrument is often attempted, the effort is largely wasted because the instrument selected was not developed following the strict application of Rasch rules for conjoint simultaneous measurement with the objective of creating a single attribute, linear unidimensional interval scale. In many case the Rasch 'test' is applied, as in the case of the MENQOL exercise, to a multiattribute scale where the separate domains are assessed without realizing that attempts to capture a single aggregate score invalidate the Rasch rules against multiattribute measures.

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 [22] . 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; it should be abandoned in favor of a new RMT standard instrument. We might, as a stop-gap attempt to assess the measurement standards for MENQOL domains. This may succeed, but it is not the basis for ongoing disease area and therapeutic class reviews for competing post-menopausal therapy options.

CROSSWALKING MENQOL SCORES

The failure of the ICER modelled simulation case for fezolinetant is compounded by the crosswalking (or mapping) of ordinal MENQOL scores to create equivalent utility or preference values. If the intent is to crosswalk or map from one patient reported outcome scale to another, as the basis for establishing claims for one scale when the other 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 with both the instruments having unidimensional, linear interval or ratio measurement properties. Recognizing this standard has supported a number of recent crosswalking or mapping assessments for measures of the activities of daily living to link similar instruments. A recent example assessed whether propensity scoring matching supported the unidimensionality assumption of the Rasch model; however, the analysis did not support the prospective role of propensity scoring leaving the Rasch requirement for a unidimensional measurement structure [23]

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 observations or raw counts; they both fail to meet Rasch measurement standards [24]. This means that crosswalking using a regression model is disallowed; no attempt was made to demonstrate that the scores were unidimensional, linear and interval or ratio, just the assumption, which is incorrect, that the MENQOL score is a continuous variable; it has neither ratio not interval properties, just a raw count. Once the inadvisability of believing that crosswalking, typically applying a regression equation, between ordinal scores is admitted, the entire ICER modelling exercise collapses.

This failure to recognize the role of RMT to create patient-centric single attribute measures is seen in the endorsement by the Institute for Clinical and Economic Review (ISPOR) of mapping or crosswalking practice guidelines [25]. ICER is not alone in failing to appreciate the imperative of a single attribute linear interval or ratio measure to support crosswalking. ISPOR in its practice

guidelines to support mapping, driven in large part, by the need to create preference scores, create QALYS and population assumption driven simulations, never addresses the requirement for a unidimensional interval linear scale; in proposing standards for selection and application, the limitations imposed by fundamental measurement are not addressed.

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) [26]. 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) [27]; (iii) the Menopause Rating Scale (MRS) [28] ; and (iv) the Greene Climacteric Scale [29] . 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 [30]. 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) [31].

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 or Partial Credit 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 to 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 Rasch measurement applied to instrument development. This would create, if a measure of the required manifestation of the latent construct was achieved, an interval measure that met the required properties; but it would not be a bounded ratio scale. Fortunately, algorithms have been proposed to transform interval integer scores to a continuous bounded approximate ratio scale (range 0 – 1). This would be a unique tool to evaluate both the extent to which needs defined by the patient are met in the post-menopausal target population as well as value claims for therapy response that met the required Rasch measurement standards [32]. It is worth noting that this would not be the first time Rasch standards have been applied to create similar patient population instruments. There are the two related RMT instruments that are available to capture needs fulfillment in female quality of life: the Urogenital Quality of Life Instrument (UGAQoL) and the Incontinence Quality of Life Index (IQoLI) [33] [34].

CONCLUSIONS

The commitment to the MENQOL over some 25 years, points to a failure, across the board, to understand the standards and limitations imposed by fundamental measurement for patient reported outcome claims; standards that were widely accepted by measurement theorists even before the MENQOL emerged. The MENQOL is a poor choice, and one that should not have been 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) [35 36].

If there a commitment to creating a Rasch-based instrument to assess needs-fulfillment as a manifestation of the latent construct, quality of like, then we have the tools available. RMT would guarantee that the focus is on the benefits assessed by patients, taking into account patient ability to respond to therapy and the difficulty of subjectively assessed needs. This would put to one side attempts to infer benefit from multiattribute instruments that concatenate or bundle clinically determined endpoints and the weighting of patient responses to a limited symptom set to generate ordinal scores or raw counts of observations as inputs to assumption driven modelled simulation with claims for a non-evaluable cost-effectiveness metric to support ersatz recommendations for pricing and access. Instead, the focus would be on interval and bounded ratio measures that met the standards of normal science and fundamental measurement to assess status and response to therapy in post-menopausal populations.

Conflicts of Interest: None

Note: The opinions expressed in this paper are those of the author (PCL)

REFERENCES

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

[2] Langley P. Rasch measurement and patient reported value claims: A primer for hemophilia. *InovPharm.* 2022; 13(4): No.

[3] Thornton S. "Karl Popper", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Zalta E & Nodelman U (eds.), https://plato.stanford.edu/archives/win2022/entries/popper/

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

[5] 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

[6] 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

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

[8] 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

[9] 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

[10] Lewis J, Hilditch J, Wong C. Further psychometric property development of the Menopause-Specific Quality of Life questionnaire and development of a modified version, MENQOL-Intervention questionnaire. *Maturitas*. 2005; 50(3):209-21

[11] Thoemmes F, Kim E. A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Res*. 2011; 46:90-118

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

[13] 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

[14] 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

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

[16] 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

[17] 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

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

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

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

[21] 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

[22] 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

[23] Hong I, Hay C, Reistetter T. Feasibility study using propensity score matching methods for the pseudo-common person equating requirement. *OTJR (Thorofare NJ).* 2019;39(1):32-40

[24] 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

[25] Wailoo A, Hernandez-Alava M, Manca A et al. Mapping to estimate health-state utility from non-preference based outcome measures: An ISPOR Good Practices for Outcomes Research Task Force Report. *ValueHealth*. 2017;20:18-27

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

[27] 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

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

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

[30] 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

[31] 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

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

[33] McKenna S, Whalley D, Renck-Hooper U et al. The development of a quality of life instrument for use with post-menopausal women with urogenital atrophy in the UK and Sweden. *Qual Life Res.* 1999; 8(5): 393-8

[34] McKenna S, Williamson T, Renck-Hooper U, Whalley D. The Development of UK and Canadian English Versions of the Incontinence Quality of Life Index (IQoLI). *J Outcomes Res*, 1997;1:9-16

[35] 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

[36] 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