MAIMON WORKING PAPERS No. 22 NOVEMBER 2021

MAPPING IMPOSSIBLE UTILITIES: THE ICER REPORT ON TEZEPELUMAB FOR SEVERE ASTHMA

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

Abstract

The release on 4 November 2021 of the final evidence report by the Institute for Clinical and Economic Review for tezepelumab in severe asthma illustrates once again the failure to recognize the standards of normal science, in particular the limitations imposed by the axioms of fundamental evidence for value claims. The weaknesses in the ICER approach to invent evidence by assumption driven simulations are well known and have been detailed in a number of recent commentaries in Innovations in Pharmacy. The focus of this commentary is on ICER's attempts (supported by its expert modelling group) to create EQ-5D-3L ordinal preferences from a disease specific asthma questionnaire, the Asthma Quality of Life Questionnaire (AQLQ). The question is whether it is possible from the perspective of fundamental measurement to create a simple linear algorithm to map AQLQ scores to EQ-5D-3L preferences. As detailed here, the answer is that it is mathematically impossible as the aggregate AQLQ score is ordinal, apart from the fact that the AQLQ is a multiattribute score that lacks construct validity and any pretense to having interval properties. The reason is straightforward: The AQLQ is constructed from Likert scales, all of which have only ordinal properties, with the result that the aggregate AQLQ score is ordinal. Attempting to attach an algorithm to map to EQ-5D-3L scores (which are themselves ordinal; if measured directly) is disallowed as an ordinal scale will not support anything other than non-parametric statistics. Ordinal scales merely rank observations; they cannot support the standard arithmetic operations of addition, subtraction, multiplication and division. Disallowing the mapped utilities means that the expert group ICER modelling collapses; the model cannot be sustained, let alone for its other manifest failures. ICER and the expert modelling group were advised of this as part of the public comment period for the draft tezepelumab report, with the recommendation that the model be withdrawn. ICER failed to respond and went ahead. To do otherwise would impact their business model. The purpose of this commentary is to detail this attempt to create impossible utilities and the case for rejecting the ICER recommendations entirely.

Keywords: ICER, tezepelumab, pseudoscience modeling, nonsense pricing, AQLQ impossible mapping

INTRODUCTION

The release on 4 November 2021 of the final evidence report by the institute for Clinical and Economic Review ICER) for tezepelumab (Tezspire; Amgen and Astra Zeneca) in severe asthma is yet one more example of the manifest failings of ICER's many attempts, supported by groups of consultant academic model builders, to create imaginary evidence for cost-effectiveness and the consequent inconsequential recommendations for a social Health Benefit Price Benchmark (HBPB) [1]. As detailed in previous commentaries and in submissions to ICER on previous evidence reports, the ICER reference case requirements for modelled comparative claims fail the standards for normal science [2] [3] [4] . Assumption driven simulation modelling by ICER and the expert academic consulting groups

contracted to ICER produce just one of a potential multitude of tezepelumab models; none can claim superiority over the other in choice of assumption because claims from the past cannot support claims on an unknown future with the ICER modelling extending over decades. Add to this the fact that the application of generic multiattribute scores to create a quality adjusted life year (QALY) is mathematically impossible and you are left with value claims that lack credibility, are impossible to empirically evaluate and replicate [5]. The ICER exercises, as has been detailed, is pseudoscience; yet this is the paradigm that has characterized health technology assessment for the last 30 years [6] [7]. It appears that the 'discipline' of health technology assessment is the only one in the social sciences that bases its credibility on inventing evidence for cost-effectiveness claims.

THE TEZEPELUMAB IMAGINARY MODEL

Tezepelumab is a monoclonal antibody that targets thymic stromal lymphopoietin (TSLP), in severe asthma. It is administered by subcutaneous injection every 4 weeks. It is presently designated a breakthrough therapy with approval expected from the FDA in quarter 1, 2022. Its price has yet to be determined but with ICER and the expert group creating the assumption driven simulation model, assigning a placeholder annual net price of $28,000 (based on the price of dupilumab) for the purpose of making imaginary pricing recommendation.

The model involves a lifetime Markov framework, utilized in previous ICER imaginary simulations and evaluations of asthma therapies, where a hypothetical asthma population proceeding, by assumption, through two recurring health states: an asthma non-exacerbation state and an asthma exacerbation state. The final absorbing state was death (it cannot recur). Assumption driven estimates of time spent in each health state are multiplied by the ordinal preference or utility score to create impossible QALYs. In this model, as detailed below, the 'preferences' are created by a mathematically impossible 'mapping' from the Asthma Quality of Life Questionnaire (AQLQ) ordinal scores.

The model then proceeds, as detailed in previous commentaries, to generate lifetime imaginary QALYS and lifetime imaginary costs. The term imaginary is used advisedly because these QALY and cost estimates, as they are by assumption and for the future, are just one of a multitude of possible alternative assumptions. Assumptions as to an unknown future cannot be justified by past observations; the problem of induction. For the base case imaginary modeling of tezepelumab plus standard of care versus standard of care yields, respectively, precise yet imaginary $697,000 lifetime costs and $228,000 lifetime costs respectively; the corresponding imaginary lifetime QALYs are 15.00 years and 13.91 years respectively or $430,000 per incremental QALY gained; the base case imaginary value claim.

Applying the impossible ICER imaginary cost per imaginary QALY threshold criteria, the annual price of tezepelumab required to reach thresholds between $50,000 and $200,000 per QALY range from $6,200 to $15,000 per annum. ICER proposes a socially acceptable imaginary (HBPB) price for tezepelumab between $9,000 and $12,000. While ICER claims it can validate its model, validation can be claimed equally for any other model and would be equally unconvincing given the nature of assumption driven simulations and the impossible QALY. The only accepted validation criteria is

empirical evaluation of the claims made which in the expert group model are designed to be impossible; clearly, these recommendations by ICER should be rejected out of hand.

A DIGRESSION ON MEASUREMENT

Accurate measurement is the key to value claims that are credible, evaluable and replicable; if measurement fails to meet required standards then the value claim fails. The often quoted statement by Lord Kelvin (William Thomson 1824-1907), inscribed above the entrance to the University of Chicago, Social Science Research Building, is the touchstone: *If you cannot measure, your knowledge is meager and unsatisfactory* [8] *.* Value claims for comparative response to therapy require in health technology assessment can only survive if they respect the axioms of fundamental measurement; this they have singularly failed to do [9 10].

Following the formalization by Stevens and others in the 1930s and 1940s, scales or levels of evidence used in statistical analyses are classified as nominal, ordinal, interval or ratio [11]. Each scale has one or more of the following properties: (i) identity where each value has a unique meaning (nominal scale); (ii) magnitude where values on the scale have an ordered relationship with each other but the distance between each is unknown (ordinal scale); (iii) invariance of comparison where scale units are equal in an ordered relationship with an arbitrary zero (interval scale) and (iv) a true zero (or a universal constant) where no value on the scale can take negative scores (ratio scale). The implications for the ability to utilize a scale to support use of arithmetic operations (and parametric statistical analysis) are clear. Nominal and ordinal scales do not support any mathematical operations; only nonparametric statistics. Interval scales can support addition and subtraction while ratio scales support the additional operations of multiplication and division as they have a true zero. This zero point characteristic means it is meaningful to say the one object is twice as long as another. To measure any object on a ratio scale it has to be demonstrated that all criteria for an interval scale have been met with a true zero. It is impossible to take an ordinal score and translate that to a ratio score via a simple linear transformation. If a ratio scale requirement is dictated by the need to create QALYs then that has to be designed from the get-go in instrument development [12]. Unfortunately, creating ratio measures for latent constructs is far from settled.

It cannot be assumed, *ex post facto*, that a given scale has interval, invariance of comparison, or ratio properties. This point is made Bond and Cox [12] in their discussion of Rasch Measurement Theory (RMT) theory and its contribution to fundamental measurement: in traditional test theory (TST) and item response theory (IRT) the observed data have primacy; results are exploratory and descriptive of those data. Rasch models are, on the other hand, confirmatory and predictive; a confirmatory model requires the data to fit the model where following the principles of conjoint measurement are sufficiently realized to claim the results are a measurement scale with interval measurement properties [12].

THE ASTHMA QUALITY OF LIFE QUESTIONNAIRE

Although widely used over the past 25 years, the AQLQ fails to meet the standards of fundamental measurement; it is a multiattribute ordinal scale. The reasons for this are obvious: in the instrument development in the 1990s no though was given to the required measurement properties (other than adding up integer scores) and the limits imposed by fundamental measurement standards. If you require an instrument to support multiplication then is must be dimensionally homogeneous or unidimensional, with construct validity and ratio properties [13]. The AQLQ score cannot support claims for response to therapy. The fact that it is widely used is testament to the widespread failure to recognize the limitations imposed by the axioms of fundamental evidence, despite a number of warnings going back over 30 years [14] [15] [16]; to which should be added the seminal contributions of Rasch, and Luce and Tukey to fundamental interval measurement for detecting measurement structures in non-physical attributes, in the 1960s [17] [18].

Consider how the AQLQ is assembled and the measurement implications of Likert scales [19]. This is a 32 item-questionnaire used to assess the physical, occupational, emotional and social qualities of adults 17 to 70 years exhibiting mild to moderate asthma. It is a multiattribute instrument with four domains: symptoms (12 items), activity limitation (6 generic and 5 patient-specific items), emotional function (5 items), and environmental stimuli (4 items). Each item response is on a 7 point Likert scale with responses ranging from 1 = maximal impairment to 7 = minimal impairment. The items are in the form of questions with each of the scale points anchored on a word or phrase and not just the extreme values; descriptors include “totally”, “extremely”, “very”, “moderate”, “some” “a little”. As Wilson et al note: some of these scales may be confusing to respondents as they mix adjectives with other grammatical elements and that there is no published evidence that the anchor words and phrases can be consistently ordered independently of their numerical positioning on the response scale or that the relative positions of different phrases represent approximately equal psychometric intervals [20]. A common feature of Likert scales. The fact that the AQLQ has shown strong classical measurement properties based on integer ratio assumptions, is irrelevant; this only occurs if you ignore the axioms of fundamental measurement and assume the AQLQ has interval properties for the Likert scores (which could equally well be designed on a 7 point scale as A >B >C >D >E >F > F >G rather than with a numeric assignment 1 >2 > 3 >4 >5 >6 >7 > 8 or even an emoji for each Likert space). In other words, if traditional or classical statistical operations are to be attempted with an instrument such as the AQLQ, then the developers need to demonstrate: (i) that all items are of equal difficulty and (ii) that the spaces between each Likert item are of equal distance [12]. This is patently not the case with these polytomous data.

Likert scales do not have interval properties (i.e. invariance of comparisons) between adjacent spaces. Just as we can’t interpret the ‘numerical or response’ distance between A and B, we can’t interpret the distance between 1 and 2. The easy way out is to ignore the question of the failure to address invariance of comparisons and just assume they exist. In other words, that each Likert scale is on a ratio scale because we need to claim after addition of the scores for each individual Likert scale that the overall scale has ratio properties together with the scales for the 4 domains. As there is no proof or any statement of intent from developers, we cannot assume that the Likert scale for each question

has interval, let alone ratio properties, or that the limitations imposed by the axioms of fundamental evidence actually occurred to the developer.

As Likert scales are ordinal scales this means the AQLQ combines 32 Likert scales none of which, if we follow the axioms of fundamental measurement, can support averages as the spaces between the integers or letters are unknown. At best ordinal scales can only support medians and modes and non-parametric statistics. As noted, an ordinal score, unless you assume otherwise, cannot support the arithmetic operations of addition, subtraction, multiplication or division. Nevertheless, the scoring of the AQLQ ignores these requirements of fundamental measurement and treats the scales, either through ignorance or design, as if they had interval properties. This allows an average score to be created for each ordinal Likert scale with domain and aggregate scores created by merging the average Likert values for each for each item.

It is surprising that, after some 30 years, the limitations of fundamental measurement have not been addressed in respect of the AQLQ. To describe the average AQLQ score as a 'score' is a misnomer; it is a value that results from illegitimate manipulations of Likert scales to produce a 'number' that is meaningless in response to therapy terms. Put simply, you cannot take an integer value from one sub-domain of a Likert sale and add it to another Likert scale. The AQLQ cannot support claims for response to therapy applying aggregate or subdomain scores. This does not mean the rejection of statistical techniques but the assurance that we are dealing with correct interval measures before application.

THE IMPOSSIBLE ALGORITHM

Mapping from the AQLQ to the EQ-5D-3L is accomplished by application of the following algorithm for each respondent where AQLQ is the aggregate (annual) score [21]:

$$\text{EQ-5D-3L} = 0.14 + 0.12\ \text{AQLQ}$$

The first point to note is that the AQLQ as an ordinal or ranked score cannot support multiplication; the mapping falls at the first hurdle. Certainly there are integer values but these fail to have interval let alone ratio properties. Aggregating over Likert-based integer scores (which is itself invalid) yields an AQLQ ordinal score. This is captured in the algorithm to create the EQ-5D-3L score (somewhere in the range 0, = death and 1 = perfect health. There is an issue: is this transformation (which is invalid) intended to create ordinal scores? There is no mysterious alchemy that allows an ordinal score to be translated by a single algorithm to a ratio EQ-5D-3L; it is an impossible score. A failure which characterizes later efforts at mapping from the AQHQ [22 23].

It is worth noting that similar issues are faced in trying to map from aggregate scores created by the St George's Respiratory Questionnaire (SGRQ) to EQ-5D-3L preferences in COPD studies [24]. Again, we face the issue of ordinal scores (the aggregate SGRP score) being applied to create, presumably, ordinal scores; again a mathematical impossibility, where the 'aggregate' SGRQ score is a dog's

breakfast of integers attached to polytomous and binary responses. Again, where the SGRQ and EQ-5D-3L are included in the same study as ordinal scores, it is easy to map (ignoring measurement theory) individual responses from one to the other but this assumes that the two are ratio and not ordinal scores for the respondents. As it stands, they are both ordinal scores so that mapping is mathematically impossible. Whatever transformations are attempted the AQLQ and the SGRO will never support a ratio transformation as the output data are only on an ordinal score. Given that the EQ-5D-3L/5L instruments support only ordinal scores, it seems odd that analysts want to pursues creating ordinal scores through mapping, if it is to be believed, from one ordinal score (the AQLQ) to another (EQ-5D-3L) unless, of course, they firmly believe that both instruments are ratio scores without any necessity of proof. The problem is that you cannot believe, on the one hand, that in mapping we are creating a ratio score (from an ordinal score?) while on the other hand, direct algorithmic measures of the EQ-5D-3L/5L have negative values and fail the standard for a ratio score.

The implications for the ICER and, in particular, the expert modelling group, are dire. The utilities which are assumed by application of the mapping algorithm to have ratio properties are only ordinal scores. You cannot just disregards the axioms of fundamental measurement. The mapped impossible ordinal preferences cannot support the creation of QALYs. The model should be withdrawn.

NEXT GENERATION QUALITY OF LIFE VALUE CLAIMS

It may come as a surprise to those wedded to multiattribute generic (ordinal) preferences that we have had, for some 25 years, a focus not on health related quality of life (HRQoL) defined by a short-list of symptoms or attributes that are bundled together to create a dimensionally heterogeneous, multidimensional score that lacks construct validity, but single attribute interval measures of the latent construct need fulfillment quality of life. These measures meet the required standards of RMT applying simultaneous conjoint standards of measurement theory, as well as capturing the patient the patient voice [25]. In the case of asthma we have the Asthma Life Impact Scale (ALIS) that was developed some fifteen years ago [26]. The rationale for the ALIS is that the focus of patient reported outcomes measures in asthma, such as the AQLQ, should not be exclusively on symptoms and functioning (which should be captured as separate unidimensional attributes), rather we require a holistic, single latent construct approach, with the question: to what extent are the needs of asthma patient's being met under various therapy intervention regimens. In other words, what is the overall impact of a therapy on the patient's quality of life; the conceptual framework is that quality of life is dependent on an individual's ability to fulfill fundamental needs and that their quality of life is high when these needs are met. With the application of RMT, the instrument items are selected to reflect a single underlying unidimensional construct with face and content validity, with overall construct validity. Scores on the final version of ALIS range from 0 to 22 with a high score indicating a major negative impact of asthma where each item elicits a binary response of True/Not true to create an interval scale. This allows for value claims regarding interval response to therapy, but not a ratio scale.

Importantly, more recently a transformation algorithm has been developed to translate disease specific interval measures such as ALIS into bounded ratio measure in the range 0 to 1 [27]. This gives, for the first time, a coherent disease specific unidimensional measure of quality of life that evaluates the extent to which need is fulfilled and the response to therapy options in disease specific quality of life terms. We are now in a position to abandon instruments such as the AQLQ and the ordinal EQ-5D-3L/5L preferences (including the applications of impossible mapping algorithms to create one ordinal scale from another) in favor of instruments to capture quality of life which meet required fundamental measurement standards.

ICER and the expert academic group were made aware of the contribution of RMT and the ALIS instrument, together with the ability to create a true ratio score to assess response to therapy [28]. Yet they continued with a mapping algorithm which was mathematically impossible in order to create ersatz utilities to populate their assumption driven simulation model.

CONCLUSIONS

After 30 or more years it is puzzling that so many practitioners in health technology assessment appear 'ignorant' of the constraints imposed by the axioms of fundamental evidence on claims for therapy response. Although it is over a century since the standardization of evidence levels and 60 years since the introduction of conjoint measurement theory for capturing, if possible, latent constructs such as need fulfillment quality of life, there is a dogged persistence to support the existing imaginary value claim paradigm which is an imperfect and mathematically impossible, framework for evaluation true measures of response to therapy. Algorithms such as that proposed in the tezepelumab expert group model are, from the perspective of measurement theory, pointless. Yet senior researchers in health technology assessment, academic groups and ICER continue to believe that measurement theory can be put to one side in developing ersatz value claims for therapy response and health benefit price benchmarks. Perhaps, after 30 years, there is just too much to lose in abandoning an entrenched belief system.

REFERENCES

[1] : Rind D, McQueen R, Herron-Smith S et al. Tezepelumab for Severe Asthma; Evidence Report. Institute for Clinical and Economic Review, November 4, 2021. https://icer.org/wpcontent/uploads/2021/05/ICER_Severe-Asthma_Evidence-Report_110421.pdf

[2] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348

[3] Langley P. Peter Rabbit is not 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. 2 https://pubs.lib.umn.edu/index.php/innovations/article/view/3992/2855

[4] Langley PC and McKenna SP. Measurement, modeling and QALYs. *F1000Research.* 2020; 9: 1048 https://doi.org/10.12688/f1000research.25039.1

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

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

[7] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care *Programmes. New York; Oxford University Press, 2015

[8] Kuhn T. The function of measurement in modern physical science. *Isis*. 1961;52(2):161-93

[9] McKenna SP, 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

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

[11] Stevens S. On the theory of scales of measurement. *Science*. 1946;103: 677-80

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

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

[14] Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*. 1989;70:308-12

[15] Tennant A, McKenna S, Hagell P. Application of Rasch Analysis in the development and application of quality of life instruments. *Value Health*. 2004;7(Supp 1):S22-26

[16] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice. *J Rehabil Med*. 2012;44:97-98

[17] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960

[18] Luce R, Tukey J. Simultaneous conjoint measurement. A new type of fundamental measurement. *J Math Psychol*. 1964; 1(1);1-27

[19] Juniper E, Guyatt G, Epstein R et al. Evaluation of impairment of health-related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax. *1992*;47:76-83

[20] Wilson S, Rand C, Cabana M et al. Asthma Outcomes: Quality of Life. *J Allergy Clin Immunol*. 2012;129(3 0): S88-123

[21] Tsuchiya A, Brazier J, McColl E. Deriving preference-based single indices from non-preference based condition-specific instruments: Converting AQLQ into EQ5D indices*. White Rose Research Online*. 2002.

[22] Young T, Yang Y, Brazier J et al. The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D.*Med Decis Making*. 2011;31(1):195-210

[23] Yang Y, Brazier J, Tsuchiya A et al. Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. *Med Decis Making.* 2011;31(2):281-91

[24] Starkie H, Briggs A, Chambers M et al. Predicting EQ-3D values using the SGRO. *ValueHealth*. 2011;14(2):354-60

[25] McKenna SP, 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

[26] Meads D, McKenna S, Doward L et al. Development and validation of the Asthma Life Impact Scale (ALIS). *Resp Med*. 2010;104:633-43

[27] Langley P, McKenna S. Fundamental Measurement: The Need Fulfillment Quality of Life (N-QOL). *InovPharm.* 2021;11(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/3798/2697

[28] Institute for Clinical and Economic Review. Tezepelumab for Severe Asthma: Response to Public Comments on Draft Evidence Report. November 4, 2021 https://icer.org/wp-content/uploads/2021/05/ICER_Severe-Asthma_Response-to-Public-Comments_110421.pdf