MAIMON WORKING PAPERS No. 13 AUGUST 2021

WHERE BLISS IS IGNORANCE: THE INSTIUTE FOR CLINICAL AND ECONOMIC REVIEW (ICER), THE CENTER FOR PHARMACOEPIDEMIOLOGY AND PHARMACOECONOMIC RESEARCH, COLLEGE OF PHARMACY, UNIVERSITY OF ILLINOIS AT CHICAGO AND IMAGINARY CLAIMS FOR MAVACAMTEN IN HYPERTROPHIC CARDIOMYOPATHY

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

Abstract

It has been conclusively demonstrated in a number of papers and commentaries that the analytical framework imposed by the Institute for Clinical and Economic Review (ICER) is pseudoscience; it fails the standards for normal science including, most importantly, the axioms of fundamental measurement. ICER is not alone, its commitment to assumption driven imaginary lifetime simulations is shared by its contracted academic modelling groups with, unfortunately, a wider audience of clinicians who seem oblivious to the constraints imposed by fundamental measurement. The fact that is overlooked (or conveniently ignored) is that where preference scores typically have only ordinal properties; they can only support non-parametric statistics (median and modes but not averages). This applies to both generic and disease specific instruments. In the present case, the just released draft evidence report on mavacamten for hypertrophic cardiomyopathy, this failure to appreciate the standards of normal science is shared by the Center for Pharmacoepidemiology and Pharmacoeconomic Research University of Illinois at Chicago College of Pharmacy (UIC-CPR) as consultants to ICER. They appear to have no idea of the constraints imposed by ordinal scores; or, possibly, what is an ordinal score. The result is that the conclusions of the evidence report should just be ignored. ICER and their stable of academic consultants persists in treating ordinal preference scores as if they had ratio properties and then launching forward with cost per

quality adjusted life year (QALY) modeling when ICER knows full well that the QALY is an impossible mathematical construct. The purpose of this commentary is to reflect on this non-belief, pointing to the major errors and shortcomings in the recent ICER evidence report for mavacamten, with particular reference to two widely used instruments in cardiovascular disease, the New York Heart Association (NYHA) functional classification and the Kansas City Cardiomyopathy Questionnaire (KCCQ) both of which produce ordinal metrics.

INTRODUCTION

The Institute for Clinical and Economic Review (ICER) draft evidence report to assess the effectiveness and value of mavacamten (Myokardia Inc; acquired by Bristol Myers Squibb October 2020) in hypertrophic cardiomyopathy (HCM) demonstrates, once again, ICER's commitment, despite overwhelming evidence to the contrary, to pseudoscience; joining intelligent design in the Dover courtroom[1] [2] [3]. A commitment apparently shared by the College of Pharmacy, University of Illinois Center for Pharmacoepidemiology and Pharmacoeconomic Research (UIC-CPR), as consultants to ICER. Rather than discovering new facts the ICER business model dictates the creation of evidence through assumption driven simulations that lack credibility. Claims are made that are impossible to assess empirically, not least because they end decades into the future, but with a commitment to quality adjusted life years (QALYs) that rests on a strange belief that ordinal scores have a mystical property of being ratio scales in disguise. The fact that the QALY is an impossible mathematical construct is ignored [4].

The UIC-CPR mavacamten simulation model is designed to invent evidence; it is not the basis for the discovery of new evidence in heart disease. The stated primary claim of the analysis is to estimate the incremental cost-effectiveness of mavacamten used along with first-line standard of care treatments for patients with symptomatic HCM. Two de novo semi Markov models were developed, informed by key clinical trials and prior relevant economic models which also are designed to invent imaginary evidence.

At this stage two points are worth noting: first, the model's claims for assessing cost-effectiveness are invalid because the effectiveness measure is not a single attribute but a composite of attributes which, as detailed below, lack dimensional homogeneity. Second, the model assumes that claims from the past are the basis for assumed future claims. This defies an elementary logical rule: the problem of induction. We cannot assume that a prior observation is the basis for making a claim for a future observation; otherwise known as Hume's problem this has been recognized since the mid 18th century.

Our odyssey to boldly go into an unknown imaginary simulated future starts with a hypothetical cohort of patients treated with standard first-line therapy, standard first line therapy alone, myectomy along with standard first-line therapy, septal ablation along with first-line therapy, or disopyramide along with first-line therapy. To round out this smorgasbord of structural treatment assumptions, an assumed cycle length of 4 weeks is applied. Model assumptions and their rationale are detailed in Table 4.2 of the report (e.g., the absence of any account for discontinuation of therapy or serious adverse events).

Treatment effects are assumed to be characterized via observed changes in New York Heart Association (NYHA) class [5]; an assumption based on trial preference data (EXPLORER-HCM NCT03470545) for scores and with data points to support assumptions selected from the literature. The subjectively assessed NYHA classes are: Class 1 – No objective evidence of cardiovascular disease; Class 2 – Objective evidence of minimal cardiovascular disease; Class 3 – Objective evidence of moderately severe cardiovascular disease; and Class 4- Objective evidence of severe cardiovascular disease. It is not clear who assigns the 'objectivity'.

Within the assumed semi-Markov hypothetical framework the hypothetical cohort cycles (figuratively) through the various NYHA classes over its member hypothetical lifetimes as the disease progresses with mortality rates captured from literature reviews and expert opinion. The UIC-CPR model actually collapses Class 3 and 4 Classes although the EXPLORER-HCM trial separately

identifies these classes as a primary endpoint. The hypothetical patients remain in the model until the time of their hypothetical death. The imaginary simulation creates estimates, still imaginary as they could never be validated, of lifetime spent in NHYA classes in 4-week increments. These are the basis for the imaginary QALY claims.

THE MYSTICAL EQ-5D-3L RATIO UTILITY

To understand the mystical nature of the ICER preferences a brief digression to measurement theory is required as too few, including UIC-CPR, seem to have grasped the rudiments of fundamental measurement[6]. Following the formalization by Stevens and others in the 1930s and 1940s, scales used in statistical analyses are classified as nominal, ordinal, interval or ratio. 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 (addition, subtraction, multiplication, division); 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 have been met with a true zero. Finally, any scale must apply to a single attribute, trait or construct. Multiattribute scales are unacceptable as they lack dimensional homogeneity, unidimensionality and construct validity. These requirements are, it appears,

totally foreign to ICER and UIC-CPR, although a standard in the physical sciences and the more mature social sciences such as economics. If attributes are to be combined to produce a further ratio scale then the components must have ratio properties (e.g., weight and height to create the body mass index).

If these standards are applied to generic preference scales then we have to conclude: (i) the scales are ordinal (aggregating over ordinal symptom and response scales); (ii) the scales lack dimensional homogeneity and construct validity (combining different attributes); and (iii) the scales produce negative values or states worse than death (with unfortunate implications for denial of care and eugenic policies). This last point is true for both the EQ-5D-3L and EQ-5D-5L as negative scores for health states are produced. The possibility that in certain disease states and target populations responses may elicit only positive values is irrelevant as there is always the possibility of a negative health state. In the case of the EQ-5D-5L US valuation, as the most egregious case, of the 3,125 possible health states (5^5: 5 symptoms; 5 response levels) 20% have negative scores [7]. This yields the impossible negative QALY although it might prove possible that when the imaginary QALYs are being aggregated over the hypothetic treatment population's hypothetical life spent in the model that the aggregate QALY count could be deflated by the negative QALY count. States worse than death, although not mentioned in the report have unfortunate eugenic implications for an ICER criteria directed denial of care[8]. To re-emphasize: the preference instruments are not ratio scales and cannot support multiplication. As Richard Dawkins notes, in reference to what he describes as extreme versions of mind viruses, Tertullian's *Certum est quia impossible est* (it is certain because it is impossible) [9]. Belief in the mystery of the ratio preference score puts ICER, UIC-CPR and many others in august company.

ICER and UIC-CPR are no alone in believing that ordinal scores have ratio properties. A recent health status analysis of the EXPLORER-HCM focused on the Kansas City Cardiomyopathy Questionnaire (KCCQ) [discussed in more detail below] reporting increments for mavacamten therapy impact from baseline to week 30 of the trial [10]. Compared to a placebo arm the claim made was for a

change in KCCQ overall score for the mavacamten arm with a mean score difference of +9.1 (ceiling score 100) and for those with a KCCQ score greater than 20 an absolute difference of 21%. Even though confidence intervals are attached, these claims are invalid because the KCCQ creates an ordinal score; distance between scores are unknown, only the rank. This error is repeated in the pivotal phase 3 report of EXPLORER-HCM results, but alongside single attribute clinical measures that have the required measurement properties [11].

THE MYSTERY UNFOLDS

The last step in the UIC/CPR modeling is to create an imaginary claim for incremental cost effectiveness. This involves two steps: a translation of time spent in an NYHA class by a preference score assigned to that class to create QALY equivalents and (ii) an assumption as to the drug and other direct medical costs for the NYHA class over the hypothetical lifetime of the hypothetical population. This yields the basis for imaginary incremental cost per QALY claims between mavacamten and the standard of care, together with other comparisons with disopyramide, myectomy and septal ablation.

The utilities for the NYHA classes (strictly speaking values as there is no uncertainty in framing the preference questions[12]) were provided by the manufacturer. Although not stated explicitly these are presumably EQ-5D-3L preferences rather than EQ-5D-5L values. As these were taken from trial data they include disutilities from the small number of adverse events noted. For those who believe in the mystical ratio property of EQ-5D ordinal preferences, a concern is the absence of a systematic review of the literature on heart disease, including HCM. Existing systematic reviews of the EQ-5D-3L instrument in heart disease point to a substantial variation in ordinal scores by stage of disease to include the NYHA stages. To give one example: Dyer et al found that scores decreased from a mean (not allowed with ordinal scores) of 0.78 (SD 0.18) to 0.51 (SD 0.21) from mild to severe disease in heart failure patients [13]. As this review was for the period to 2010 there will be additions not just for the EQ-5D but for the HUI Mk2/3 and the SF-6D among other scales.

In the ICER/UIC-CPR model average preference ordinal scores for mavacamten were 0.95 for NYHA class 1, 0.87 for NYHA class 2, and 0.71 for NYHA classes 3 and 4 combined (Table 4.2). Corresponding utilities for standard of care were 0.95, 0.85 and 0.70. Note again that as these are ordinal scores we can only rank them as we have no idea of the distance between the scores. This means that the range for the scores presented is meaningless (e.g., we have no idea if the distance values for the upper (+0.05) and lower (-0.15) range described for the 0.95 NYHA utility for mavacamten is capturing three times the same distance).

In this imaginary UIC-CPR simulation the claimed improvement in QALYs drives the cost-per-QALY calculations. Comparing mavacamten to standard of care yields a QALY improvement of only 0.97 (from 12.54 to 13.51 QALYs). Alternative generic preference instruments would yield different ordinal scores and impossible QALY claims. Note that in all multiattribute scores the preference are community preferences and not the preferences of patients; the notion of the need of patients is absent.

Although often seen as a justification for hypothetical claims, one way sensitivity analyses and, the icing on the cake, probabilistic sensitivity analysis (in this case with 1,000 simulations), simply obscure the fact that changing the model assumptions opens the floodgates to any number of competing models. Indeed, ICER makes this possible through its *ICERAnalytics* cloud-based platform which, when complete, will allow any one (at least for a small fee) to indulge in changing assumptions to see how any hypothetical cost-per-incremental QALY can be created or even reverse engineered; perhaps to while away a hypothetical wet Sunday afternoon at your hypothetical beach front property in the Hamptons you might indulge in *ICERAnalytics* [14].

A more recent review of studies, claiming to measure HRQoL in heart disease focused on methods to derive QALYs in trial based utility analyses, is less than enlightening [15] Again, the authors neglect any consideration of the standards of normal science and, in particular, the need to conform to the axioms of fundamental measurement. Of the 18 individual trials identified the majority

employed the EQ-5D-3L to derive QALYs (again no consideration of ordinal preference scores). The fact that the QALY is a mathematically impossible construct seems not to have occurred.

Although a meaningless exercise, the conclusion to the UIC-CPR modelling clearly disfavors mavacamten; or at least for those who take hypothetical assumption driven ordinal simulation modelling seriously. The base case reported by the analysts at UIC-CPR is a hypothetical lifetime drug cost for mavacamten of $1,258,000 and hypothetical total costs of $1,156,000 vs standard treatment of $12,600 and $434,000 respectively. Corresponding imaginary (mathematically impossible) lifetime QALYs are 13.51 and 12.54. Against standard treatment, the imaginary cost-per-QALY gained is $1,200,000 with an incremental imaginary QALY of 0.97.

The implications of the assumption driven simulation exercise are entirely expected and in line with what ICER views as its role to establish imaginary health benefit price benchmarks; in this case for a cost-per-imaginary QALY threshold of $150,000, a five-fold reduction from the placeholder price of mavacamten of $75,000.

DISEASE SPECIFIC INSTRUMENTS: THE MYSTICAL LIKERT SCALE

Unfortunately, in the majority of cases these same arguments apply for rejection of disease specific instruments in heart disease. Once again, if an instrument is to have single attribute ratio or interval measurement properties, then this has to be an integral part of the development of the instrument. Although not as extensively applied as the New York Heart Association (NYHA) functional physical activity subjective classification, it is of interest to consider the measurement properties (or the lack thereof) of the KCCQ in patients with heart failure which, together with instruments such as the Minnesota Living with Heart Failure Questionnaire (MLHFQ), relies on Likert scales to create summary scores for health status [16].

Given the lack of appreciation of the limitations of measurement theory and the fond belief that a Likert scale is a ratio scale, a brief comment is in order. The Likert scale is typically a five (or seven) point scale which is used to allow the individual to express the extent to which they agree or disagree with a particular statement. Users of Likert scales attach numbers to the Likert categories (e.g., 0 … 6; 1 …. 7). They may as well attach letters (e.g., A, B, C, ….F) or emojis (or any ideogram or pictogram). What is overlooked (or never thought of) is that whatever number, letter or image is used it only provides a ranking; we have no idea of the 'distance' or 'extent' between the letters, emojis or the numbers assigned to the Likert categories. The assumption, however, is that when numbers are assigned the difference between 2 and 3 is the same as between 5 and 6; or the distance between B and C is the same as between E and F). This is a patently false assumption; an assumption is not proof that the distances are the same. The notion of distance, or invariance of comparison (required for an interval scale) is never addressed. Unfortunately, their ignorance of measurement theory allows developers and users to aggregate scores over respondents (to give a mathematically impossible average) and to add the numbers over Likert items (also mathematically impossible). The resulting overall score is simply nonsensical. Despite its ordinal status, the NCCQ continues to be applied and reported on in the leading journals, such as the *Lancet*. Presumably, *Lancet* reviewers have no understanding of the limitations imposed by fundamental measurement.

The focus in developing the KCCQ was on how heart failure impacted patients' lives rather than, as proposed here, to assess the need of cardiovascular patients with hypertrophic cardiomyopathy, the extent to which that need is met and the how new therapies might impact that need [17]. First developed in 1996 as a multiattribute instrument it is clear that those responsible for the KCCQ had no idea of the limitations imposed by the axioms of fundamental measurement; this extends to the present day (including the more recent 12 item version) [18]. The KCCQ is a multiattribute instrument comprising 23 items covering, it is claimed, 7 domains: symptom frequency, symptom burden,

symptom stability, physical limitations, social limitations, quality of life and self-efficacy (patients' awareness of their condition). The Likert responses (as integers) are scaled from 0 to 100 and often summarized to 25-point ranges where health status is poor to very poor for 0 to 24; poor to fair 25 –to 29; fair to good 50 to 74; and good to excellent 75 to 100 and interpreted in HRQoL terms. The domains can also be scored.

The items are presented as Likert ordinal scales. The overall score and the domain scores are, therefore, ordinal scores. Assessing change in an ordinal score (other than rank position) is meaningless. It is also meaningless to compare the KCCQ 'sensitivity' to other cardiovascular instruments such as the popular Minnesota Living with Heart Failure questionnaire. Again, this rests on a series of Likert scales, is multiattribute and in consequence another ordinal scale. Again, such comparisons are meaningless although there are clearly associations with, for example, the NYHA classification and with clinical measures of health status. The point that escaped the attention of developers is that if response to therapy is to be accurately measured then the measure must have ratio measurement properties; it must refer, as in the physical sciences, to a single well-defined attribute either physical or latent, with a true zero and dimensional homogeneity. In these respects the KCCQ is a singular failure; claims for therapy response, in either overall or sub-scale scores, and minimum clinical difference are meaningless; it was doomed, from a measurement perspective, from the start.

While it may seem odd to undertake factor analyses of instruments with only ordinal score properties, suspending belief points to the rather strange domain structure of the NCCQ. An assessment by Creber et al which attempted to (i) explore the factor structure of the NCCQ; (ii) undertake reliability and validity testing for 'best' factor (or multiattribute solution); and (iii) determine the meaningful components of health states captured by the KCCQ came to some divergent sub-scale conclusions [19]. The authors recommended that, consistent with earlier studies of the factor loadings of the KCCQ, there are four recognized subscales: (i) physical limitations, (ii) symptoms, (iii) quality of life/perception of

social interference and (iv) self-efficacy. The authors proposed that the 3 items intended to measure Q0L be reclassified under the Social Interference subscale with no QoL subscale because the items are a poor measure of quality of life and the subscale is redundant in a HRQoL tool.

Needless to say the Creber et al paper takes no account of the axioms of fundamental measurement or techniques for item identification, selection and inclusion that characterizes Rasch Measurement Theory. It treats the Likert scales that characterize each of the 21 original items in the NCCQ as ratio scales; none were reworded or otherwise changed they were just reshuffled, like the proverbial deckchairs, to a different steerage class box. There was no concept of the tasks required to focus on latent constructs or single attributes. All we have is an HRQoL instrument that focuses on symptoms and functions. The neglect or dismissal of QoL as a latent construct is ridiculous. After all, if the patient is the presumed principal beneficiary of an intervention, then we should focus on the extent to which need is met. We have to abandon these cobbled together ordinal Likert scales, focusing instead on identifying single attributes that are measured separately as inputs to formulary decisions. Even with this restructuring the point remains that the NCCQ overall and sub-scale scores are still ordinal, telling us nothing about response to therapy, other than with non-parametric assessments.

In what must be one of the more bizarre treatments of ordinal scores, Thomas et all report on predicting or crosswalking KCCQ Likert scores to EQ-5D utilities[20]. While this must be seen as a worthless exercise as both scales are ordinal, the authors with no thought given to measurement theory, claimed they could predict EQ-5D utilities directly as well as predicting individual EQ-5D items; effectively predicting one Likert scale from another. Why anyone would want to engage in creating ersatz EQ-5D values is a puzzle; certainly NICE in the UK asks for a crosswalk from the EQ-5D-5L to the EQ-5D-3L for their imaginary reference case modeling exercise, but attempting to crosswalk from an ordinal Likert based disease specific instrument to the EQ-5D-3L/5L is just nonsensical.

Finally (last but not least) is the question of the validity of physician assignment of patients to the NYHA classes. A primary outcome measures for the EXPLORER-HCM trial is the transition between the NYHA classes. Credibility for this claim rests on establishing the process by which patients are allocated to those classes. If follows, from this assignment, that the credibility of EQ-5D-3L preferences from patients assigned to those classes (at baseline and 30 weeks) rests on this assurance that to eliminate the view that the assignment is actually a random process.

It should be noted, however, that the ICER/UIC- CPR imaginary simulation takes the NYHA classes as the basis for assigning preference states and computing imaginary QALYs (or NYHA QALYs) for time spent in any of the three NHYA classes over the course of the hypothetical patient progress. This raises two issues that the ICER/UIC-CPB simulation fails to address: first, the categorization of NYHA functional status is apparently highly subjective with considerable variability among assignments, assigning patients to NYHA classes can be problematic, particularly as the ICER/UIC-CPB analysis given no indication that the preference assigned to NYHA classes are valid. [21]; second, it is unclear what the NHYA classes are actually capturing; it appears that the focus is on functional status which is assumed to be a proxy for health state preferences. If the EQ-5D-3L is applied to value the NYHA classes that patients are assigned to then we have no information other than upper and lower limits as to the distribution of preference valuations and, in the case of Class 1 versus Class 2 NYHA whether the values are significantly different (an impossible criterion as the preferences are ordinal). There are, apparently, no EQ-5D-3L preference scores that take negative values. It would have been useful to express these results in terms of both the ordinal EQ-5D-3L and EQ-5D-5L instruments; if the assignment of patients to the various NYHA classes is essentially subjective then the model framework for progression through NYHA classes is meaningless as are any transition matrices.

These doubts are present in a recent comparison of the differences between patient reported health status with the NCCQ and the physician assigned NYHA

classification [22]. There are interesting differences. Assigning NHYA classes to the 80 plus KCCQ group (assumed to represent perfect health), the analysis found that this group comprises 26.4% categorized as Class 1 NYHA 51.7% of patients categorized as Class 2 NYHA patients and 18.9% categorized as Class 3 patients. If patients are assigned to NHYA classes as the basis for utilizing KCCQ claims for therapy response then we are entering a rather strange world of comparing ordinal classifications and scores which lack measurement credibility to support pivotal claims for FDA approval where a significant proportion of patients defined by the KCCP as in good health are actually classified by the NYHA as providing objective evidence of moderately severe cardiovascular disease.

NEXT GENERATION QUALITY OF LIFE CLAIMS

The fact that the QAL:Y is an impossible mathematical construct because preference scores are ordinal as well as lacking dimensional homogeneity, unidimensionality and construct validity, means that we should look to a compelling construct to support the measurement and reporting of QoL. The required criteria are:

- **Analytically rigorous, meeting the standards of normal science for credibility, evaluation and replication**
- **Adherence to the axioms of fundamental evidence**
- **Disease specific for target patient populations**
- **Patient (and caregiver) centric**
- **Single attribute**
- **Ratio scale**

At the present there is only one candidate that meets all of these criteria: the need fulfillment QoL measure that determine the extent to which respondents can meet their fundamental human needs. The measures are derived directly from relevant patients and provide data on the value to these groups from alternative interventions. This ability is clearly related to the symptoms and

functional limitations they experience. In contrast to clinician determined HRQoL , these new measures generate a single index of patient value or QoL rather than adding together a basket of outcomes defined as ordinal scales.

As the first of the next generation class of QoL instruments, the development of each disease specific instrument is in two parts. First the creation through the application of Rasch measurement theory of an index that captures the ranking of items (binary response) specific to the stated need of patients or caregivers in a disease area. Item selection and ranking reflects the combined characterization of the difficulty of the item and the ability of the patient to respond to that item [23]. Second, given the index. it is then transformed to a bounded ratio scale; bounded by the number of items yielding a scale in the range 0 to 1 to indicate the extent to which need is met [24]. Negative scores and scores > 1 are impossible. It is a true ratio scale. An item score of 0 indicates that none of the identified item needs are met; a score of 1 indicates all needs are met. As an example, items for a need fulfillment instrument in atopic dermatitis, the Quality of Life in Atopic Dermatitis (QoLIAD) instrument which comprises 25 items (statements) with a binary response (True/Not True) [25].

Thinking about your experience of having atopic dermatitis …….

- **I worry about my appearance**
- **I avoid physical contact**
- **My life revolves around my condition**
- **I feel tense all the time**

The bounded ratio score, the Need-QoL (or N-QoL) can be used in both clinical trials and observational studies to evaluate response to therapy in terms of better meeting patient need. As well, it can be applied in regression modelling to evaluate possible determinants of need fulfillment. In either case it is removed from the existing ordinal preferences and the support they give to imaginary assumption driven simulation modelling.

CONCLUSIONS

It would be an understatement to maintain that there is any merit to the ICER/UIC-CPR draft evidence report for mavacamten in hypertrophic cardiomyopathy. This should come as no surprise given the belief that evidence can be created through assumption driven simulations that lack credibility and any effort at empirical evaluation and replication. At the same time, abandoning the scientific method and embracing pseudoscience, is further evidenced by a denial of the axioms of fundamental evidence. Preferences must have ratio properties; the QALY must be protected. This is axiomatic; it is the foundation of the ICER business model.

It is now widely recognized that the ICER/UIC-CPR invention of evidence for cost-effectiveness claims is an analytical dead end. As a response to limited evidence at product launch it should never have been considered. Rather there should have been a commitment to the standards of normal science; in particular fundamental measurement. If the focus is on QoL then this should have been assessed with the creation of patient and disease centric instruments. Instead, we have the commitment to assumption driven imaginary simulations to create impossible cost-effectiveness claims.

Unfortunately, ICER claims and modelling by academic groups all too frequently go unchallenged by health care decision makers. This is notable in respect of their lack of understanding of the constraints of fundamental measurement. A situation that is repeated in protocol designs for pivotal clinical trials.

As it stands, the ICER/UIC-CPR draft report is a waste of time. If fails the standards of normal science, it is designed to invent evidence for cost-effectiveness claims using the mathematically impossible QALY. Any recommendations for pricing and access, including the notion of an ICER determined social price with impossible QALY thresholds should be ignored. Simulation modelling has no place in establishing value claims for products.

At the same time, in cardiovascular disease, reliance is placed on two unacceptable measure for classifying patients (NYHA) and for clinical outcome assessment (KCCQ). The first is subjective with considerable variability in patient assignment; the second fails to meet standards for fundamental measurement as it is nothing more than a dimensionally heterogeneous ordinal scale. If we are to capture QoL, then a single attribute need fulfillment instrument is the obvious solution.

There is a way forward: abandon generic preferences and QALYs; abandon simulated assumption driven QALY claims and abandon a mystical belief in ordinal scales with ratio characteristics. Instead we should commit to look to response defined in terms of patent centric, single attribute measures

REFERENCES

[1] : Wasfy JH, Walton, SM, Beinfeld M, Nhan E, Sarker J, Whittington M, Pearson SD, Rind DM. Mavacamten for Hypertrophic Cardiomyopathy: Effectiveness and Value; Draft Evidence Report. Institute for Clinical and Economic Review, August 18, 2021. https://icer.org/hypertrophic-cardiomyopathy-2021/

[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 a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(2): No 22
https://pubs.lib.umn.edu/index.php/innovations/article/view/3992/2855

[4]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

[5]https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure

[6] Langley P, McKenna S. Fundamental Measurement and Quality Adjusted Life Years. *Value Health*. 2021;24(3):461[letter]

[7] Pickard A, Law E, Jiang R et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*.2019; 22(8):931-41

[8] Langley P. Abandoning Eugenics and the QALY. *InovPharm*. 2021;12(3): No.20
https://pubs.lib.umn.edu/index.php/innovations/article/view/4291/2939

[9] Dawkins R. A Devil's Chaplain. New Yotk: Houghton Miflin, 2004

[10] Spertus J, Fine J, Elliott P et al. Mavacamten for treatment of symptomatic obstructive hypertrophic cardiomyopathy (EXPLOTER-HCM): health status analysis of a randomized, double-blind, placebo-ontrolled, phase 3 trial. *Lancet*. 2021;397(10293):2467-2475

[11] Olivotto I, Oreziak A, Barriales-Villa R et al. Macacamten for treatment of symptomatic obstructive htpertrophic cardiomyopathy (EXPLORER-HCM): a randomized, double-blind, placebo-controlled, phase 3 trial. *Lancet.* 2020;396(10253):759-69

[12] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes. 4th Ed. New York: Oxford University Press, 2015

[13] Dyer M, Goldsmith K, Sharples L et al.A review of health utilities using the EQ-5D in studies of cardiovascular disease. *Health Qual Life Outcomes*, 2010;8:13

[14] Langley P. Let a Thousand Models Bloom: ICER Analytics Opens the Floodgates to Cloud Pseudoscience. *InovPharm*. 2021;12(1): No. 5 https://pubs.lib.umn.edu/index.php/innovations/article/view/3606/2668

[15] Rankin J, Rowen D, Howe A et al. Valuing health related quality of life in heart failure: a systematic review of methods to derive quality-adjusted life years (QALYs) in trial based cost-utility analyses. *Heart Failure Reviews*. 2019;24:549-63

[16] Rector TS, Cohn J. Assessment of patient outcome with the Minnesota Living with Heart Failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Pimobendan Multicenter Research Group. *Am Heart J.* 1992;**124**(4):1017–25.

[17] Spertus J, Jones P, Sandhu A et al. Interpreting the Kansas City Cardiomyopathy Questionnaire in Clinical Trials and Clinical Care*, J Am College Cardiol*. 2020;24(20)

[18] Green C, Porter C, Bresnahan DS, et al. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: :A new health status measure for heart failure. *J Am College Cardiology*. 2000;35(5):12451255

[19] Creber R, Polomano R, Farrar J et al. Psychometric properties of the Kansas City Cardiomyopathy Qurestionnaire (KCCQ). *Eur J Cardiovascular Nurs*. 2012;11(2):197-206

[20] Thomas M, Jones P, Cohen D et al. Predicting the EQ-5D utilities from the Kansas City Cardiomyopathy Questionnaire in patients with heart failure. *Eur Soc Cardiol*. 2021; 7:388-96

[21] Bennett J, Riegel B, Bittner V et al. Validity and reliability of the NYHA classes for measuring research outcomes in patients with cardiac disease. Heart Lung.2002;3(4):262-70

[22] Tran A, Chan P, Jones PO et al. Comparison of patient self-reported health status with clinician assigned New York Heart Association classification. *JAMA Network Open*. 2020; 3(8)

[23] Bond T, Cox C. Applying the Rasch Model: Fundamental measurement in the human sciences. 3rd Ed. New York: Routledge, 2015

[24] Langley P, McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6 https://pubs.lib.umn.edu/index.php/innovations/article/view/3798

[25] Whalley D, McKenna SP, Dewar AL et al. A new instrument for assessing quality of life in atopic dermatitis: international development of the Quality of Life Index for Atopic Dermatitis (QoLIAD). *Br J Dermatol*. 2004; 24(1):105-13