MAIMON WORKING PAPERS No. 4 March 2023

REJECTING PSEUDOSCIENCE: IMAGINARY COST-UTILITY CLAIMS CREATED BY THE REFERENCE CASE SIMULATION MODELLING OF THE NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE)

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

ABSTRACT

The commitment by the National for Health Care Excellence (NICE) in the UK to the creation of assumption driven modeled simulated cost-utility claims over the past 20 plus years has been a staple of health technology assessment. The NICE reference case has been widely imitated in a significant number of other single payer health systems as well as by advocacy groups such as the Institute for Clinical and Economic Review (ICER) in the US. The problem, which has been pointed to on a number of occasions, is the concept of cost-utility which is central to the NICE reference case; interval measures with a composite or multiattribute structure deny the standards of fundamental measurement. This means that as utility scores are multiattribute ordinal constructs, the QALY is mathematically impossible. Add to this the commitment to hypothetical reference case lifetime simulations; these simulations fail the standards for normal science by ignoring demarcation. Reference case cost-effectiveness modeled claims cannot be falsified. This relegates reference case models to the non-science or pseudoscience league, alongside intelligent design. The purpose of this review is to make the case that the NICE 2022 technology assessment manual, in its advocacy of the reference case, is both misleading and irrelevant in its support for formulary decisions by denying both the standards of normal science and fundamental measurement.

Keywords: NICE technology assessment failure, falsification, impossible interval measurement, simulated pseudoscience

INTRODUCTION

It has been noted on a number of occasions that the current belief system or meme in health technology assessments puts front and center in its analytical armamentarium the creation of assumption driven simulated modelled claims for a single non-empirically evaluable generic metric: cost-utility [1] [2]. Judged by the standards of normal science and fundamental or Rasch modern measurement this is an untenable situation, relegating HTA to the category of pseudoscience [3]. This relegation stems from an apparent lack of interest in meeting the standard for demarcation: a criterion that distinguish science from non-science [4] . The criterion

is well established: the appeal to falsification [5] . While the boundary may be fuzzy, such as assessing historical evidence to support claims, the evidentiary reference is clear cut. We must be in a position, in Popper's terms to support the evolution of objective knowledge, through the ability to appeal claims to superior evidence [6]. There is an asymmetry between proof and disproof. We can never confirm or prove a statement or value claim; we can only make the claim for provisional acceptance given a failure to disprove the claim.

In HTA there is no opportunity, by design, to empirically evaluate a multiattribute cost-utility claim. Rather we face the application of thresholds, sensitivity assessment and even the application of probabilistic sensitivity analysis to validate assumption driven modeled imaginary claims for cost-effectiveness. All that is occurring is a modelled rationalization for an assumption driven claim; a story euphemistically labelled as 'approximate' information' [7]. This is an analytical dead end, but one that is enthusiastically embraced in the thousands of peer-reviewed published studies that have proposed cost-utility claims. A framework that has been recently endorsed by the CHEERS 2022 guidance for authors submitting imaginary cost-utility modelled claims to journals [8 9]. A framework that fails to recognize the opportunity for the development of self-serving assumption driven claims where, with a judicious choice of assumptions, the sponsor's product can always be claimed, at their preferred price, to be cost-effective [3].

The purpose of this review is to point out the fundamental errors in the NICE reference case while asking an equally relevant question: why does this commitment to pseudoscience, rejecting falsification, persist? After 20 plus years how has this commitment and the rejection of the standards of physical science continued to be globally accepted?

THE NICE REFERENCE CASE

The stated purpose of the NICE reference case guidelines is to ensure that economic evaluations are consistent as guides to an efficient use of health system resources [10]. The reference specifies the methods NICE considers to be most appropriate for estimating clinical effectiveness and value for moneys while ensuring that there is consistency in the evaluations. While non-reference case claims can be made, a reference case analysis must always be performed.

The reference case submission for a product or device must define the decision problem, detailing the choice of comparators and with outcomes capturing all relevant health effects. Costs must be defined from the perspective of the health

system; productivity costs should not be included. NICE accepts two types of submission: a cost utility analysis when a full analysis of health benefits and costs are included or a cost-comparison analysis if the technologies are likely to provide benefits of similar or lower costs that the relevant comparator(s). Cost-effectiveness or specifically cost-utility is applied to determine if differences in expected costs between technologies can be justified in terms of health effects changes. Health effects are defined in terms of quality adjusted life years (QALYs) which are seen as the most appropriate generic measure reflecting both mortality and health related quality of life effects. Claims should be expressed in incremental cost-utility ratios (ICERs). Claims for net health benefits should be presented applying values of £20,000 and £30,000 per QALY gain.

The choice of time horizon for assessing ICERs should reflect all important differences in costs and outcomes., recognizing that many technologies may impact of the patient's lifetime, so a lifetime horizon is often appropriate where there are differences in survival benefit that last a lifetime. In consequence, it is recognized that there is often the need to extrapolate using statistical models and choice of assumptions.

Utility scores to assess changes in health-related quality of life (HRQoL) should be based on public preferences using a choice-based method where the QALY combines time spent and preferences for that health state relative to, usually, perfect health or death. Patients should measure HRQoL directly with responses valued by community preferences. As different methods yield different utility values, the EQ-5D-3L instrument should be used to maintain consistency. If EQ-5D-3L values are not directly available from relevant studies they can be sourced from the literature or mapped from other measures. Where the EQ-5D-3l is not considered appropriate, the utilities should be based on other preference measures, condition specific measures, vignettes or direct valuation. There is no recommended specific measure for quality of life in children and young people. However, NICE does not recommend the EQ-5D-5L value set for England. If EQ-5D-3L preferences are not available, these should be constructed by mapping from the EQ-5D-5L, with a NICE mandated mapping function.

Estimates of resource use and costs are for those resource units under the control of the health system, valued by unit prices relevant to the health system. Costs should be included that relate to the condition of interest and incurred in additional years of life gained. The present value of the stream of costs and benefits accruing over the

time horizon of the reference case analysis should be discounted at a rate of 3.5% per year.

Models are needed for most evaluations. The chosen type of model (e.g., Markov cohort model and model structure should be justified. Modelling results are to be presented in a disaggregated form and should include: (i) life years gained; (ii) mortality; (iii) frequency of selected outputs predicted by the model. Extrapolation should borrow information from similar enough classes of technologies , populations and settings. Care has to be taken to consider the plausibility of surrogate modelled biological endpoint claims and extrapolated final outcomes as well as surrogate endpoint claims for long term cost-effectiveness as long as they can be demonstrated to be validated. Validation would be supported through sensitivity analysis including probabilistic sensitivity analysis.

Assumptions used in models should be checked, having both internal and external validity External validity of the extrapolation should be as well as its coherence with external data sources; while the internal validity of alternative models of extrapolations should be based on their relative fit to the trial data and alternative model scenarios.

An overall assessment of uncertainty should be explored and should demonstrate the relative effect of different types of uncertainty (parameter, structural) on cost-effectiveness estimates. The preferred cost-effectiveness estimate should come from probabilistic sensitivity analysis.

DECONSTRUCTING THE NICE REFERENCE CASE

The NICE reference case can be deconstructed in terms of what it fails to address in a meaningful framework for evaluating therapy response. There are three issues that place it firmly in the category of non-science. These are: (i) the absence of any reference to the need to meet the standards of normal science for therapy value claims: falsification and demarcation; (ii) the failure to understand the limitations imposed by fundamental measurement where value claims must be presented as single unidimensional attributes as measures with demonstrated linear, interval and invariance properties; and (iii) the failure to recognize the problem of induction in the choice of assumptions to support non-evaluable simulated modelled outcomes.

The failure to address the question of demarcation and falsification is the most egregious errors in the formulation of the NICE reference case. Absent any commitment to the notion of falsification, the reference case is simply non-science

or pseudoscience; it cannot be taken seriously as a basis for creating or rejecting product therapy claims for a composite cost-effectiveness measure which itself is a claim immune to the application of Popper's demarcation criterion [11] Falsification is central to the standards of normal science, accepted since it

NICE does not question the measurement properties of utility scores or the QALY. It merely states that it considers it to be the most appropriate generic measures of health benefits with assumed properties such as constant proportional trade off and additive independence between health states. Constant proportional tradeoff is just an assumed and unjustified property; one that cannot be defended as the utility scores are ordinal [12]. The assumption of additive independence between health states (attributes) is irrelevant; the QALY is multiattribute and this fails standards for dimensionality and dimensional homogeneity [13] .

Nice also fails to question the fact that the modelled claim for cost-effectiveness is based entirely on assumptions drawn from the literature including pivotal trial-based claims for the therapy under review. This raises the question of the criteria to be applied to choose assumptions to populate the model. The problem is one of induction, first raised by Hume in 1748 [14]. Successive confirmation does not support the case for one assumption over another; as Russell pointed out in a 1912 monograph [15]. We can neither prove nor disprove the principle of induction. Put simply: the fact that past futures have resembled past pasts does not mean that future futures will resemble future pasts [16]. There can be no claim that our choice of modelled assumption driven simulations can be justified by the choice of 'realistic' assumptions over any other selection of assumptions; we cannot justify one set of assumptions over another [17]. This means that one modelled claim cannot be preferred to another on the choice of assumptions. The only exception is where the modelled claim is empirically evaluable and if falsified may lead us to challenge certain assumptions. Absent falsification of competing claims we are left with only options in assumption choice to support models as marketing claims to support a sponsor's product [18] [19]. The fact that NICE employs academic invigilators whose professional life has been devoted to challenging the choice of assumption in reference case models submitted by manufacturers does not resolve the issue; it merely highlights it.

NORMAL SCIENCE AND RASCH MEASUREMENT

The standards for the evaluation of product value claims are well established: all claims must be credible, empirically evaluable and replicable. This is the antithesis

of the NICE-HTA framework which puts hypothesis testing to one side in favor of non-evaluable approximate information claims. Although not considered in the NICE technical manual, there are well established standards for fundamental measurement, notable the application of conjoint simultaneous measurement for patient reported outcomes in Rasch measurement [20] . The key point, which must be emphasized, is that these standards for fundamental measurement apply to value claims expressed as single attributes, nor the HRQoL bundled health states and response levels which form the basis for generic utility and QALY scores. These instruments and resulting algorithm output produce only ordinal utility scores; a ranking of outcomes that can only be evaluated with non-parametric statistics.

This embrace of composite utilities points to a failure in HTA and the NICE reference case to understand the distinction between observations which are always ordinal and measurement that requires interval scales [21]. Observations are just numbers; subjective ordered counts of events, responses or performance levels. The key step with the rules of Rasch measurement is to provide the necessary and sufficient means to transform ordinal number counts to interval, and under some circumstances, ratio measure for a single attribute. Original observations, provided by instrument responses, are not a measure that supports anything other than non-parametric statistics. The Rasch focus is on capturing a single attribute as a manifestation of a latent construct; a number which must be transformed, if possible, to a calibrated interval measure with a defined origin and a unit of measurement [22]. The NICE utility score fails as a measure with interval properties because it bundles attributes (health state dimensions with response levels), and thus lacks unidimensionality and dimensional homogeneity; it fails construct validity.

The imperative of developing and justifying Rasch criteria standards for interval measures means creation of an interval scale must precede anything other than non-parametric statistical analysis. The implication for the reference case mapping to support EQ-5D-3L utilities is quite clear: it is mathematically impossible to map via regression modelling from one ordinal multiattribute scale to another. Mapping requires interval or ratio measures. Efforts devoted to creating EQ-5D-3L/5L value sets with mapping for multiattribute ordinal scales over the 20 years have been a monumental waste of time and resources.

If the commitment is to single attribute patient reported measures, the further key property of Rasch measurement is that the transformation to an interval score not only ensures order and invariance on an approximation to a linear interval scale but

accommodates the interaction between item difficulty and respondent ability [20 22]. An instrument must retain its calibration character irrespective of what it is measuring or who is responding to it, recognizing at the same time the importance of the interaction between the ability of the respondent and the difficulty of an item. A requirement that is totally absent from the notion of multiattribute utility preferences.

Rasch measurement recognizes that there is an element of chance in response to a questionnaire item; the distinction between item difficulty and respondent ability supports a measuring system that indicates the likely success on an item; the more difficult the item, given ability, the less likely is a successful response [20 22]. This is the cornerstone of Rasch measurement, first proposed as a mathematical model in the 1950s, and widely applied globally in educational testing for the past 60 years and in graduate programs for the social sciences, yet essentially ignored in HTA. The sole exception being the disease specific Rasch measures for quality of life defined as a single attribute needs fulfillment and developed over the past 25 years [23].

THE HOLY GRAIL QALY METRIC

If the commitment is to developing assumption driven simulation to produce a decision-friendly claim for cost-effectiveness, then the QALY is seen as the focus for target patient populations. It is proposed by leaders in HTA as the undeniable robust measure to support resource allocation decisions between therapies and their access for target patient populations. With that commitment the fact that the QALY is an impossible mathematical construct can be put to one side as an unnecessary academic inconvenience, and one that can be safely ignored, puts agencies such as NICE in an untenable position.

Assumption driven modelled simulation claims exist only because of the QALY; absent the QALY as the focal metric to be matched to imaginary costs, to create incremental cost-per-QALY claims, the entire edifice collapses [2]. Simulation modelling to support cost-per-QALY claims and thresholds require a single QALY metric; a requirement that in turn drives a commitment to impossible ordinal mapping. The reason for this is clear cut: for HTA to play a role in the allocation of resources in a health care system and a wider budgetary environment we presumably need a metric that is common across disease states and stages of disease. There must be a gold standard, a Holy Grail, that yields a single utility measure of multiattribute health benefit, presumably a ratio measure embodying a starting point (a true zero) and expressed in invariant standard units. There must be agreement on the construct

that is being captured and manifested in the metric (e.g., multiattribute HRQoL). Assuming this is possible, given a rejection of the standards of fundamental measurement, the metric can be factored into imaginary cost-per-unit metric terms (e.g., incremental cost-per-QALY) and resources allocated by disease area. A process of denial of care and expansion of access to care within a fixed budget to ensure equal imaginary marginal benefits across all health states and stages of disease. This allocative process is driven by assumption supported simulated modelled claims; hence the central role of the favored QALY metric and its acceptance as an easy modelled option by NICE and many other single payer health systems. It provides within a short time frame assumption driven modelled claims for composite cost-effectiveness with no basis for empirical evaluation, falsification and reporting to a formulary committee.

This is, quite clearly, an impossible scenario, although of interest no doubt to those who advocate a central planning model for health systems (a Soviet 1920s vision). The fundamental flaw is a failure to define the required metric characteristics, given agreement on the relevant latent construct, to create an invariant, unidimensional, generic composite multiattribute interval scale that can be transformed to a bounded ratio scale in the range 0 to 1. The confusion or lack of awareness of this is evidenced in the last edition of the most popular textbook in HTA by Drummond et al. There is a brief discussion of measurement scales with the claim that the metric (in this case the QALY HRQoL utility weights) have composite valued interval properties (pgs. 129-131). The fact that an interval scale, as it is not anchored on zero, can take negative vales is recognized, but the most convenient scale is an interval scale bounded by zero and unity; while recognizing that there could be states worse than death and those with greater than perfect health the required measurement scale is interval (with possible negative values); a ratio scale is put to one side. If follows, therefore, to take ratios of differences in two interval scales provides the justification for incremental QALY ratios and claims (e.g., the difference in a QALY weight between -0.10 and +0.10 is the same as the difference between +0.70 and +0.90). Both are interpreted in this generic metric as yielding he same additional health benefit.

The problem with this defense of comparing utility weights is that there is no recognition in Drummond et al of the required measurement standards for the measure to capture meaningful claims for therapy response: the measure must be for a single construct which is unidimensional, linear, interval and invariant. To claim, by assumption, that an interval scale can capture bundled or composite attributes is

a contradiction in terms. This false measurement property cannot be assumed; it has to be demonstrated through the application of Rasch rules which, of course, is impossible. This is not to advocate a quasi–ex post facto assessment of consistency with Rasch criteria but a conscious effort to develop a unidimensional instrument applying Rasch rules. As noted, the fact that the NICE reference case fails to meet the standards of normal science and Rasch measurement means it fails the demarcation criterion. In fact, the question of demarcation and falsification is never raised either in the Drummond et al textbook or the NICE technology evaluations manual. Following Drummond et al, all health technology assessments are required to do is to inform social decisions rather than prescribing social choice with composite multidimensional scales that fail the requirements of fundamental interval scale measurement. Yet, this is achieved through assumption driven non-evaluable imaginary reference case modelled claims that, following the demarcation criteria, are informing decisions with an analytical framework that is categorized as pseudoscience. A position that is reinforced in the CHEERS 2022 guidance for submitting non-falsifiable modelled reference case claims to journals [8 9]. Once again, there is no mention of the required standards of normal science, fundamental measurement and the demarcation criteria in the checklist for CHEERS 2022 authors. The mathematically impossible QALY still holds center stage.

THE NICE DENIAL

While there is no acknowledgement in the description and application of the NICE reference case of the limitation imposed by fundamental evidence or the standards of natural science that have been in place since the scientific revolution of the 17th century, epitomized by the motto of the Royal Society '*nullius in verba'* [take no persons' word for it] while NICE insists that we take their imaginary assumption driven word for it. An acceptance that sets NICE and HTA apart from physical sciences in focusing on multidimensional, non-evaluable ordinal scores that fail Rasch standards and unsupported QALY claims. NICE, in practice terms, absolves reference case modelled claims for cost-effectiveness from the standards for falsification and fundamental measurement that characterize the physical sciences, including the standards for product development and the structure of pivotal trials which recognize falsification, and the more mature social sciences. For NICE, with its advocacy of the reference case, we can only come to grips with an unknown future-reality by creating assumption driven stories of non-evaluable ersatz cost-effectiveness claims.

Even so, is the NICE reference case the answer for coming to grips with a non-existent and unknown future reality in therapy choice and impact? This can be interpreted as a relativist position where rationality is always culturally relative, constructed and transmitted with high fidelity by a particular community of leading HTA scholars. If HTA, for the many believers in reference case modelling it is about rhetoric, persuasion and authority, then different HTA belief systems (or memes) are equally credible; for a social group truth is consensus. All that is required is to believe that multiattribute utility scales have ratio properties. An assumption that is shared with the Institute for Clinical and Economic Review (ICER) in the US where their business model is built on reference case modeled recommendations for product pricing and access [24].

For those who reject what they may see as the relativistic, culturally specific belief system or meme in HTA, deconstructing the NICE reference case is an easy starting point as it clearly fails the required standards, notably falsification. All we need ask is that the proponents of the utility ratio measurement scale to provide a coherent demonstration that, starting with multiattribute health bundles they can transform observations to a unidimensional, linear, interval scale. It must be demonstrated to have these properties; hence the importance of Rasch rules for interval measure development and the property of order. Any defense of QALY weights must be able to demonstrate how the developers of the instrument have focused on the manifestation of a single attribute from a credible latent construct, transforming subjective responses to the interval scale. Ignoring Rasch measurement, where the Rasch measurement model is not cited, is not an option as it has been conclusively demonstrated that Rasch measurement provides the necessary and sufficient means to transform ordinal counts to interval measures; a requirement that was re-cognized some 50 years before the NICE-HTA ‘breakthrough’ in modelling generic ordinal scales rather than the focus of disease specific Rasch measures [21].

RUNNING ON EMPTY

The NICE reference case is an analytical failure; producing nothing more than assumption driven fictional stories of non-evaluable cost-effectiveness claims. This failure has been apparent for a number of years; yet NICE perseveres. It has no option. To admit to the precedence, the failure to meet standards for normal science and fundamental measurement, is an admission of failure. There is too much to lose; a quandary faced by the leaders in the field of HTA. Overthrowing, in this case a

meme rather than a paradigm, is an unconscionable step; a step too far that none are prepared to acknowledge.

We face an entrenched belief in the fidelity of the NICE reference case. After almost 30 years the belief is self-sustaining: truth, for those preparing modelled claims following NICE reference case standards, for cost-effectiveness, is by consensus, a rejection of the essential principle of evolutionary progress that such a view is at odds with the evidence . A refusal, in effect, to recognize falsification as the criterion for demarcating science from non-science. Imaginary cost-effectiveness claims have been the mainstay and they must continue. There is too much at stake to accept that decisions for therapy value claims cannot be driven by imaginary outcomes. Belief, not science, for those subscribing to the HTA meme is about nothing more than rhetoric, persuasion and authority. The leaders in HTA, over the past 30 plus years, have produced a generation (or more) of believers in reference case modelling.

The question facing the more open minded HTA practitioners is whether or not to reject the current belief system or meme in favor of a framework that supports a new paradigm, recognizing the standards and imperatives of normal science and fundamental measurement, or retreat to the known meme beliefs. This retreat is unlikely to happen; too many people have too invested in the current HTA meme. In the UK, at least, such a rejection of the NICE reference case would invite ridicule and a possible parliamentary inquiry. The stakes are too high, challenging hundreds of pricing and access decisions, and practice guidance, which fail the required science standards, is an unpalatable prospect.

CONCLUSIONS

The denial of the criteria for demarcation is one that has been implicit in HTA for some 30 years; the reasons for this are not clear cut, yet we can speculate. The first possibility is the undisputed need for a single, gold standard composite metric to support access to and denial of care; hence the focus on the multiattribute generic EQ-5D-3L and the QALY irrespective of measurement properties. The second possibility is a willingness to ignore Rasch measurement and the creation of single attribute, unidimensional, linear, invariant interval measures; assuming that Rasch measurement was on the table in the first place. The third possibility is the belief that demarcation and falsification is an irrelevant concept when all that is required is a justification for the necessity of ersatz cost-per-QALY approximate incremental claims to support formulary access and health care allocative decisions. The final possibility is that with limited evidence at product launch the easily expedited

response is for evidence to be invented rather than waiting on a longer-term research horizon to evaluate meaningful therapy response claims; reference case models can be constructed to meet required endpoints in short order.

The fact that public monies, both to support application of the NICE reference case and specific recommendations for pricing and access (i.e., denial of therapy) have been allocated, raises a critical issue for NICE and its duty of care. If blanket claims for non-evaluable claims for pricing and cost-effectiveness have driven formulary decisions, it could be argued that health systems are at fault for taking NICE recommendations at face value. A critical issue when the few health system decision makers are ill equipped to assess the simulated modelled claims, let alone Rasch measurement standards.

NICE seems uninterested in meeting the standards of normal science and fundamental measurement in proposing standards for evaluating value claims for therapy response that focus on false composite interval measures. Instead, NICE appears comfortable with its assumption driven multiattribute simulated model claims that meet the criteria for pseudoscience; one-off ersatz modelled claims that focus on the mathematically impossible EQ-5D-3L based utility score to create equally impossible QALYs. Rather than considering value claims supported by protocols to support on-going research programs in therapy areas with regular disease area and therapeutic class reviews with options for outcomes-based contracting, the safe harbor over the past 20 years has been to promote ersatz non-evaluable claims that deny progress and the discovery of new yet provisional facts, with modelled threshold claims that can lead to denial of therapy. Surely, NICE, with the resources available, could commit to standards that are in place with the physical sciences, including pivotal trial claims, and the more mature social sciences, to propose provisional pricing supported by ongoing disease area and therapeutic class reviews, and not impossible cost-per-QALY thresholds.

REFERENCES

[1] Drummond M, Sculpher M, Claxton C et al. Methods for the Economic Evaluation of Health Care Programmes (4th ED0. New York: Oxford University Press, 2015

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

[3] 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

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

[5] Popper K. The Logic of Scientific Discovery. London: Hutchinson, 1959

[6] Popper K. Objective Knowledge: An Evolutionary Approach, Oxford: Clarendon Press, 1972

[7] 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

[8] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth*. 2022;25(1):3-9

[9] Husereau D, Drummond M, Augustovski F, *et al.*: Consolidated health economic evaluation reporting standards 2022 (CHEERS 2022) explanation and elaboration: a report of the ISPOR CHEERS II good practices task force. *Value Health.* 2022; **25**: 10–31

[10] National Institute for Health and Care Excellence. NICE health technology evaluations: The Manual. Process and Methods [PMG36]. 31 January 2023

[11] Thornton S. "Karl Popper", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Zalta E & Nodelman U (Eds.)

[12] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7

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

[14] Hume D. An Enquiry Concerning Human Understanding. 1748

[15] Russell B. Problems of Philosophy. 1912

[16] Magee B. Popper. London: Fontana, 1974

[17] Henderson, L. "The Problem of Induction", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Ed.) Zalta E & Nodelman U (Eds.),

[18] Xie F, Zhou T: Industry sponsored bias in cost-effectiveness analysis: registry-based analysis. *BMJ.* 2022; 377.

[19] 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: 3 approved]. *F1000Research* 2022, **11**:993

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

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

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

[23] Galen Research. Measures Database. https://www.galen-research.com/measures-database/

[24] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12