MAIMON WORKING PAPERS No. 19 (VERSION 2) SEPTEMBER 2021

THE PURSUIT OF IGNORANCE: THE INSTITUTE FOR CLINICAL AND ECONOMIC REVIEW'S ASSESSMENT OF TEZEPELUMAB FOR UNCONTROLLED CHRONIC ASTHMA

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

Abstract

The Institute for Clinical and Economic Review's (ICER's) persistence in bringing to the table assumption driven simulation modelling claims for newly developed and approved therapies continues. ICER is well aware that their reference case methodology fails the standards for normal science, notably for the denial of the axioms of fundamental evidence. The recently published ICER draft evidence report for tezepelumab in chronic asthma falls squarely in this absurd tradition. Once again ICER, or more accurately experts at the University of Colorado Anschutz Medical Campus, continue with this modelling charade. In all fairness, they have no other option. To admit that their approach lacks any semblance of scientific credibility would be a disaster; both for their reputations and cash flow. The ICER claim to fame as the cost-effectiveness and pricing arbiter in the US would collapse. ICER's position and those of the expert model builders is relativistic; that is, they apparently subscribe to the belief system that truth is consensus and that health technology assessment can be explained sociologically. To believers such as ICER, the position is to maintain that science is not concerned with coming to grips with reality. By assumption, evidence is never discovered; it is constructed. As weird as this may appear, the upshot is that no one body of evidence is superior to another. ICER, in other words, is not in denial, it just lives on another planet. The purpose of this brief commentary is to explore planet ICER, its belief system and its denial of the scientific method; spelling out the implications for the imaginary ICER modelled assessment and claims for tezepelumab.

INTRODUCTION

The survival of the Institute for Clinical and Economic Review (ICER) is not particularly surprising; it rests on two elements. First, a concordance with the broader relativist position of practitioners in health technology assessment to invent claims and, second, a profound ignorance among believers in health technology assessment of the standards of normal science shared with the audience for ICER's relativist claims, including media outlets [1] [2]. The latest example of ICER's commitment to relativist imaginary claims in health technology assessment is in the draft evidence report for tezepelumab in severe asthma, supported by the expert modeling team for imaginary simulations at the University of Colorado Anschutz Medical Campus [3]. ICER typically avoids serious criticism; where criticisms are raised they are just brushed off. The situation is little different from the belief system sustaining intelligent design: an ignorance of the principles and evidence for evolutionary biology and an external belief system that shuts out criticism, or at least the willingness to accept criticism which is in conflict with an established belief system [4]. The two also share a further characteristic: they are both pseudoscience; or at least pseudoscience (or bunk) as judged by those who recognize the contribution of the 17th century scientific

revolution. ICER's claims fail the demarcation test; they cannot support credible claims for empirical evaluation and replication [5].

FUNDAMENTAL OUTCOME MEASUREMENT

If a patient reported outcome (PRO) instrument is to meet the appropriate measurement requirements to capture response to therapy, then it must meet the following standards [6]:

- Patient centric (issues relevant to patients in disease areas)
- Acceptable to patients
- Unidimensional (reporting on only one attribute)
- Reliable
- Able to demonstrate construct validity (is it measuring what is intended and is the instruments conceptual model appropriate)

Few advocates of either generic or disease specific PRO instruments recognize these standards. In the case of the multivariate generic instruments favored by ICER to create QALYs and then support the imaginary simulated claims, none meet the most important standards: patient centricity and unidimensionality. Any instrument fails at the first hurdle if neither of these standards are met. Multiattribute PRO instruments, whether a direct or indirect, fail because they fail to meet required measurement standards. Health states cannot be described by a bundle of attributes, with ordinal response levels for each symptom, with a community asked to value them to support algorithms to yield ordinal preference scores. By definition, unless the separate attributes are related to a conceptual model or construct and can demonstrate ratio measurement properties, they cannot be combined. The preference scores are meaningless: they are neither unidimensional nor dimensionally homogeneous, they lack construct validity and report only positive and negative ordinal scores. Yet groups such as ICER continue to utilize them in simulation modelling and the construction of QALYS, and other outcome measures, without recognizing the limitations imposed by the failure to meet ratio measurement requirements. This ignorance of fundamental measurement is allied to a bizarre belief held by ICER in a mystical measurement property where ordinal scores are ratio measures in disguise.

It is important, however, to ask why ICER and others continue to promote QALYs and the invention of claims from assumption driven simulations. To answer this we need to consider the case that ICER is relativist; rather than subscribing to the standards of normal science, ICER looks elsewhere and accepts the belief that information is invented not discovered, that truth is consensus and that formulary decisions for pricing and access can be supported by imaginary models that lack credible outcome claims. ICER may not recognize these terms but ICER clearly subscribe to the relativist position that truth is consensus.

BELIEF IN THE IMPOSSIBLE

Previous commentaries by the author have challenged ICER on the grounds that, knowingly or not to ICER, the position taken by ICER is relativistic [7 8]. ICER can be viewed, in its dogged adherence to a belief

in the creation of invented evidence and subsequent claims for cost-effectiveness as subscribing to the view put forward by Feyerabend that 'anything goes' [9]. In other words, a belief in democratic relativism where different societies, or groups within societies, may look at the world in different ways and regard different things as acceptable; hence the ICER reference case where evidence is invented, not discovered. Against this view, with its links to the question of underdetermination, where evidence available to us at a given time may be insufficient to determine what beliefs we should hold in response to it, is the accepted standard of normal science. If evidence is incomplete then we should aim to remedy the situation through a program of discovery of new facts, rather than adopt a relativist position and invent evidence to support claims through assumption driven lifetime simulations.

This conflict of beliefs raises the further question of the standards for episistemic appraisal. In the case of ICER it is not, however, the question of underdetermination *per se*, of whether or not we have sufficient evidence to support a conclusion, even if we need to refer to auxiliary factors, but the response of 'inventing' evidence to support a conclusion; the acceptance of a reference case where a claim for cost-effectiveness is made from an assumption driven simulation model. The assumptions are taken from the literature, clinical trials and even 'informed' (or uninformed) guesswork to create a vision or fantasy of the claimed future response of a hypothetical population to competing therapy interventions with the deliberate choice of response measures that are incapable of empirical assessment. Induction is assumed, as it has no logical basis, with assumption driven non-evaluable claims stretching decades into the future and proposed to support formulary decisions for pricing and access. ICER is only interested in non-evaluable claims.

In ICER's case it is not a question of the difference between empirically successful or empirically unsuccessful theories and their respective advocates, but of rejecting any need for an empirical basis for judging the merits or otherwise of competing therapy interventions to support a blanket claim for cost-effectiveness. The absurdity of this position becomes clear once the issue of a multiplicity of competing models arises [10]. Are we to put 'success' or otherwise to one side and focus on the talents and resources of advocates such as ICER in technology assessment? Should we accede to a relativistic position and argue that 'truth is consensus'? If those in this health technology assessment world, planet ICER, hold to this singular belief then we should presumably have some mechanism to 'judge' the relative merits of these imaginary constructs. In the UK this is undertaken by the National Institute for Health and Care Excellence (NICE) contracting with academic centers to review imaginary claims submitted by manufacturers. The entrails of these imaginary constructs are examined by academics with years of experience in entrail reviewing. While ICER has not hired entrail assessors, it does subcontract to academic modelling groups who presumably share ICER's relativistic world view and are willing to produce imaginary pseudoscientific constructs, defying the standards of normal science.

If methodological and evidential considerations are insufficient to account for the choice of analytical frameworks what non-cognitive and social factors explain the ICER choice?

Certainly, others have taken this route; but the ICER claim that they do it because (almost) everyone health technology assessment does it is a weak defense. In ICER's case is it that they don't know any better; imitation is the sincerest form of flattery, and ICER is into flattery. They simply jumped on the bandwagon with no thought to standards of epistemic appraisal.

Scientific knowledge is not relative to its context of production nor does underdetermination give any support to relativism. Certainly, we should concede to the notion of fallibilism: that all scientific claims are provisional and liable to fail. We can agree that the totality of evidence available may not provide sufficient evidence to assess the worth of a given therapy. Yet there are good reasons, if we follow the standards of normal science, to agree that underdetermination is not a barrier. We have good reasons for provisionally embracing the best available explanation, noting also that there are objective facts about the universe even if we are presently unaware of them. However, there are no equally acceptable principles for accepting theories; the standards of normal science are not just one of other equally acceptable principles. We do not invent evidence to support a theoretical or analytical framework for predictions. Planes fly; magic carpets and broomsticks do not.

We must also reject a more radical form of relativism: social constructionism [11]. For this extreme form of relativism, reality is not simply 'out there' to be discovered; it is constructed. In this context, ICER is a willing partner: claims are constructed through a variety of norm-governed, socially sanctioned cognitive activities such as interpretation, description and manipulation of data. In relativistic terms, different social forces lead to the construction of different 'worlds' with no basis for adjudicating between them. The ICER reference case is a world apart, but one that has equal standing with a structured program of discovery to evaluate therapy response in the real world. In social constructionism there are no objective facts; only socially constructed ones, created and modified in particular times and places courtesy of prevailing theoretical and conceptual frameworks which may vary with the belief systems of different groups or societies. All are equally valid. Different social forces, belief systems or memes, lead to construction of different worlds. Including, presumably, the decision in health technology assessment to overcome limited data at product launch by inventing modeled claims for cost-effectiveness; even if those claims rest on models driven by dubious and false assumptions, in particular a failure to consider the evaluation standards for PRO instruments.

Social constructionism has a strange appeal: if there are no facts 'out there' to be discovered only invented then if claims are to be made for therapy response by target populations in disease states based on reference case simulations, those claims are valid for decision making until, at some later stage, they may be revised. There is no 'mind-independent reality' or any attempt to discover it. There can be no incentive to seek facts that are not there to be discovered, only constructed. Constructed facts are not meant to

be observed. Are the constructed facts the end of our endeavors? Does the reject and notion of empirical evaluation of claims? Are ICER's claims never to be challenged? Constructed facts are self-referential; they are not to be challenged unless by another group whose constructed facts are a challenge. The situation becomes even more strange when we consider the status of the 'constructed' facts. Are these constructed facts meant to have the property of confirmation; should we, if we feel so inclined attempt, in some sense, to validate those constructed facts? Welcome to a universe of pointless debates between constructed models from alternative elief systems.

THE MEASUREMENT SUBTEXT

In the case of ICER there is a critical subtext: denial of the relevance of the axioms of fundamental measurement. We might, if unwisely, dispute with ICER over the relativistic role of invented evidence versus a program of discovery, a search for new yet provisional facts for selected therapy interventions, but we cannot skirt the issue of measurement. It is undeniable that when the various direct and indirect preference instruments were proposed, few gave any though to the required ratio measurement standards; an instrument that would support standard arithmetic operations of addition and subtraction (invariance of comparisons), and multiplication and division (a true zero). This is not a question of belief (or its suspension) but of the axioms of fundamental measurement, the standardization of which in terms of levels of evidence has been accepted for over a century.

A further point is that if we are to create a measure with interval or ratio properties, this must apply to a single attribute, whether physical or latent. This, again, was overlooked. The direct and indirect preference instruments ask respondents to evaluate health states defined by a bundle of attributes with ordinal response scales for each attribute. This, from the perspective of the axioms of fundamental evidence is wrong. We cannot aggregate over attributes to create a single score unless the each attributes has a coherent construct supporting it and is on a ratio scale. Otherwise we are constructing a score that lacks dimensional homogeneity, unidimensionality and construct validity; and, to be more specific, a generic score that lacks content validity with a bundle of attributes that may be irrelevant to target patient populations in specific disease states.

The time bomb in the multiattribute universe is not just the ignorance of the role of single attributes in claims for response to therapy but the attempts to sideline the existence of negative health state preferences with multiattribute generic preference instruments: negative preference scores that are classified as 'states worse than death'. This makes clear that, apart from a lack of construct validity, these various preferences score have only ordinal properties, a characteristic that ICER goes to lengths to deny.

Of course, in the tradition of social constructionism, ICER could respond by claiming that the Planet Earth axioms of fundamental measurement are irrelevant as ICER is conforming to axioms developed on the Planet Dim by leading Dimwit experts who can demonstrate that what we consider ordinal generic preference scores are actually ratio measures in disguise; a mystical position ICER has willingly endorsed.

ICER's position (if somewhat confused) is that:

> *As we have expressed before we (and most health economists) are confident that changes in the EQ-5D (and other multiattribute utility instruments) do have ratio properties. The EQ-5D value sets are based on time trade-off assessments (which are interval level), with preference weights assigned to different attributes. We fail to see why this should be considered an ordinal (ranked) scale. The dead state represents a natural zero point on a health related quality of life. Negative utility values on the EQ-5D scale represent states worse than dead. We do not find this lacks face validity* [12].

Confidence can be misplaced; in this case fatally. Consider the term 'natural zero'. As stated, the term is meaningless, unless we can identify zeros existing in the real world? No one goes out to buy zero fish; although some might be confident, by diligent search, that they can find null fish [13]. There **is** no inherent or natural zero starting point (where none of the quantity is present). In mathematics 0 is the integer preceding 1. It is an even number, but neither positive nor negative; it is the only natural number that is not positive. It quantifies a count or an amount of null size. It is not the same as the integer zero. It may or may not be considered a natural number but is both a rational and a real number as well as an algebraic number and a complex number. It is usually displayed as the central number on a number line. If ICER views the preference scores as represented as a number line with zero as the central number then the number 0 is the smallest non-negative number.

ICER is incorrect is asserting that the dead state, zero, represents a natural zero. It is simply a score generated by the preference algorithm. Depending upon the algorithm preference weights and the characteristics of patient population there may or may not be an aggregate of health states reported by that population that create a zero score. The sole function of the preference algorithm is to produce, by happenstance, positive and negative scores with, possibly, a zero score recorded. The zero score has no more meaning than any other score; it is just an implicit scoring artifact. We might wish to describe negative scores as states 'worse than death' or more appropriately, scores ≤ 0. The fact that there is no lower bound at zero means that the developers of the algorithm failed to consider this requirement in the first place, thus allowing inadvertently (but quickly realized) negative scores. The result is a ranking of ordinal scores with positive, possibly zero and negative values, capped at 1. There is no natural zero. Calling negative states 'worse than death' is just a device to bolster the assertion that zero represents death, a null value where none of the 'health' quantity is present. This is a

contradiction because with negative scores some of the 'health' quantity is still present. Death is defined in terms of the symptoms or attributes bundled to create health states. Different preferences have a different definition of death defined by the attributes the developer choose; even with a negative value for each attribute in a particular algorithms, someone could be very much alive.

A further error is to present preference scores on a number line. This is invalid because constructing a number line requires invariance of comparisons. Preference scales do not have this property; they are just ranked scores with an unknown distance between each score. For ICER, scores with a negative value representing states worse than death have face validity. This is the weakest form of construct validity and seems to suggest that ICER knew there would be negative preference and community assessments of health states worse than death. Perhaps, ICER is content to stay with 'states worse than death' (rescaling each set of ordinal numbers following each application of the preference algorithm would be illegitimate) as these emphasize the eugenic component and support denial of care.

If ICER believes in the inevitability of negative scores than it follows that ICER accepts the absence of a true zero and any pretense to ratio properties. Yet, the interpretation is a little odd. If a patient with a negative preference score is treated (and not denied care on eugenic grounds) and improves, they have to pass through a null state (death) before achieving a positive preference score. In other words, they slowly improve as their negative scores decrease towards zero, then suddenly and unexpectedly they die, entering the null state (and the absence of the preference score you are measuring) and then, Lazarus like, are reborn to a positive community preference state (possibly in Boston) and continue to improve.

ICER might even consider a Flatlining Service where patients in states worse than death can be continually monitored as their conditions improve and they approach the null state; ICUs could be warned of an impending null transition and then shepherd the patients from the negative preference world to the positive preference world (as judged by community preferences); a *Through the Looking Glass* transition! ICER cannot dismiss negative preferences as an unavoidable community calculation; if so, null preferences can exist (at least in ICER's existential world) and health services should be aware of this possibility in their acceptance of the QALY (and negative QALYs) and prepare accordingly.

What ICER, and its expert advisors, fail to recognize is that if 0 indicates the absence of the variable being measured then it has a true zero and is on a ratio scale; if not, it is on an ordinal scale. With ratio data there is an order, there is invariance of comparisons (meaningful differences) and a true zero starting point. ICER cannot argue that the preference scores are ratio scores in disguise. The mystical ordinal scale with ratio properties that is an item of ICER belief is just nonsense.

Preference scales are ordinal for the simple reason that no one thought to ensure that invariance of comparisons has to be part of the design, not assumed *ex post facto*. Certainly, it gives the appearance of an interval scale because the health state scores are

presnted on an interval number line. This says nothing about our inability to make claims about the differences between health state scores produced by the scoring algorithm. Is the difference, for example, between the five health state scores of 22221 and 22222 of the EQ-5D-3L the same as between 22222 and 22223? What does the differences in scores mean, other than to rank the scores in ascending/descending order (i.e., 22221 > 22222 > 22223)? By how much is one health state preferred to another? If there are 3 response levels for each of 3 symptoms what is the difference between the null health state 33333 (where 3 is the 'worse' health state for that attribute) and 23333? In fact, the EQ-5D-3L algorithm will assign 55555 a negative value (range -0.5 to 1 in the US). Depending on the algorithm there may be none or a number of null states given community preferences or TTO weights [14]

The real problem is that we routinely mistake absolute or actual differences between fractions or percentages (e.g., difference on VAS scale anchored at 0 and 1) as having direct interval scale properties when all we may infer is their ordering of scores (i.e., we seek invariance of comparisons between scores) without taking into account the size of the score. We need a way of representing the size of the gaps, which is ignored in developing preference scores. This issue was resolved by Thurstone in the 1920s with the notion of relative difference [15]. To avoid compression at the ends of a scale (floor and ceiling effects) the log-odds scale is typically employed (that is, a natural logarithmic transformation of the odds ratio or probability). This transforms an ordinal scale to an interval scale. Note that the integer or counting scale cannot have negative numbers. Given this argument it is impossible to transform any ordinal preference scale to an interval scale if negative numbers are a possibility.

The point that is overlooked is that if you want to create a mathematically impossible QALY by multiplying a preference score by a year spent in a disease state, then the existence of negative values means there are negative QALYs [16]. A response to a preference instrument such as the EQ-5D-5L may include both negative and positive responses with preference weights attached to patient responses. Indeed, the EQ-5D-5L is the apotheosis of states worse than death [17]. The instrument yields, with five attributes and five response levels for each attribute, a total of 5^5 or 3,125 health state attribute combinations, each generating through the preference score algorithm, a single score. Unfortunately, of the 3,125 health states some 20% or 615 create negative scores. For a scale to be recognized as having ratio properties there must be no circumstance under which a negative score is possible; this clearly not the case with the EQ-5D-5L or any other multiattribute instrument.

Neither, as noted above, was thought given to relative differences; is the difference between an ordinal preference score of 0.2 to 0.3 the same as the distance between 0.8 and 0.9? The first yields a percentage difference of 50% the latter 13%; although as we don't know the actual difference (no invariance of comparison) these percentages are actually meaningless. How does this translate, if possible, to the relative difference in QALYS in response to a therapy intervention?

However as Bond and Cox point out, in their comparison of Rasch Measurement Theory (RMT) with Classical Test Theory (CTT) in the construction of fundamental measurement, this seals the fate of instruments such as the EQ-5D-5L [18]. In CTT the data have primacy, models are fitted to data; the model is descriptive of those data. They must account for all the data. No thought is given to measurement properties. Following RMT we take the opposite approach: the data are required to fit the model where the results can be used as a measurement scale with invariant, interval level properties. However much ICER might attempt to argue (or at least falsely claim) that the preference scores are ratio measures, the axioms of fundamental measurement dictate otherwise. RMT also emphasizes the need for all measures to report on single attributes to capture a coherent construct such as need-fulfillment quality of life.

DESTRUCTVE ORDINAL IMPLICATIONS

For those who have engaged in building stacks of playing cards, the awkward feature is that if you remove a foundation card then the edifice collapses. ICER is in the same position: once the ordinal nature of preference scores is recognized, with the preference score also lacking construct validity, the entire reference case simulation modelling collapses. Given the preference score is an ordinal and not a ratio scale is the key to this collapse. Central to the reference case is the QALY; a construct that is, as noted, mathematically impossible. Unsurprisingly, once the QALY is judged an impossible construct the notion of lifetime QALYs across disease stages must be dismissed, together with ersatz notions of cost-per-QALY thresholds and incremental cost-per-QALY claims, where lifetime costs are a non-evaluable figment of the collective ICER imagination and the instrument attempts to combine attributes, each on an ordinal scale. To add a final nail to the reference case coffin is the fact that the reference case fails the demarcation test between science and pseudoscience.

A first order candidate for the collapse of the ICER reference case is the much lauded (at least by ICER) social pricing reference point: the Health Benefit Price Benchmark (HBPM)[19]. For all simulated imaginary reference case models an HBPM is developed for a new intervention to reflect prices aligned with commonly cited long-term cost-effectiveness thresholds ranging from $100,000 to $150,000 per QALY gained and per equal value of life year. These prices are discounts or premiums from wholesale acquisition costs (WAC) that would be required to reach these cost-effectiveness thresholds. The ICER belief is that prices at or below these thresholds, which violate the required axioms of fundamental measurement, help ensure that the health benefits gained by using new treatments are not outweighed, in this neo-eugenics universe, by health losses due to long-term cost pressures that lead individuals, health systems and the government, to delay, abandon or deny care [20].

For those concerned with the implications of the single metric driven neo-eugenic approach to access and denial of care, the saving grace is that it is complete nonsense. With ordinal preferences the QALY is an impossible construct; hence the entire ICER edifice collapses. Long-term cost-per QALY thresholds are impossible hence discounts or premiums to reach these are also impossible.

MEANINGLESS UTILITIES

The ICER reference case models demand utilities to survive; they can come from any quarter to include the creation of ersatz utilities by mapping to predict reference scores such as those created by the EQ-5D-5L algorithm from disease specific PRO instruments that claim to measure quality of life. This is a pointless exercise because to accomplish this feat requires, once again in this imaginary world, either a willful disregard or a ignorance of the standards for fundamental evidence. As an example, a paper aiming to predict EQ-5D values from a frequently used instrument in COPD and asthma, the St Georges Respiratory Questionnaire (SGRQ makes no mention of any possible constraints imposed by the axioms of fundamental measurement where the SGRQ has in fact only ordinal scoring and ranking properties [21]). The authors apparently assume, without any proof, that the SGRQ creates ratio scores; this is patently false. In all fairness, however, the authors advocate the direct elicitation of preferences from the target population; a recommendation that has been ignored for the last 10 years. This is perhaps also fortuitous as any preference based multivariate instrument would still only produce ordinal values; that is, a waste of time. Suffering from the same defects as the SGRQ the other frequently use instrument is the Asthma Quality of Life Questionnaire (AQLQ) [22]. Both are applied in the ICER modeling.

As detailed above the focus must be, if quality of life is the latent construct of interest, to consider need fulfillment as we have had for many years existing instruments that meet the required RMT properties in both asthma and COPD. These are the Asthma Life Impact Scale (ALIS) and the Living with Chronic Obstructive Disease (LCOPD) questionnaire [23] [24]. With recent developments in the transformation from RMT scores to a bounded ratio scale for each instrument, a need fulfillment quality of life score (N-QOL) can be applied to evaluate response to therapy and the extent to which need is met by adult patients in each of the disease states [25].

Not to be deterred by considerations of fundamental measurement, the ICER simulation model estimates predicted preference values by mapping from the AQLQ. We are told that:

> *Without commonly used utilities reported in the tezepelumab trials, we relied on evidence of patient reported outcome instruments with known utility mappings. The non-exacerbation health state utility value is specific to the evidence for tezepelumab plus SoC versus SoC alone. Evidence from tezepelumab trials (NAVIGATOR, PATHWAY, and Amgen data on file) include the responses from the Asthma Quality of Life Questionnaire (AQLQ) to derive utility values using the conversion from the AQLQ to the EQ-5D-5L The least squares mean change and 95% confidence intervals from the AQLQ for tezepelumab plus SoC versus SoC alone provide the inputs for the aggregate*

> *mapping algorithm (EQ-5D = 0.14 + 0.12*AQLQ score). Disutilities for the exacerbation health states and for chronic OCS use were assumed to be the same across treatment strategies (pg. 19).*

The mapping algorithm is nonsensical as the AQLQ scores are ordinal. The EQ-5D ersatz scores detailed in Tables E2.1 and E2.2 of the supplement to the modelling report are imaginary artifacts and mathematically impossible. As both the SGRP and AQLQ generate multiattribute ordinal scores, the question of mapping to predict EQ-5D-5L equivalent ordinal scores, is clearly a waste of time. Ordinal scores cannot support the statistical operations necessary to generate mapping algorithms. If we are to attempt to model a transformation of the form Y = a + bX then the continuous independent variable, X (the AQLQ sore) must have interval properties. This is a requirement for regression modeling. Unfortunately, the experts at the Anschutz Medical Campus, University of Colorado appear to have overlooked this in their application of AQLQ scores as the independent variable in their attempt to create ordinal utility scores from ordinal AQLQ scores. A further point is that if you want to create a dependent variable with a property of a range from 0 to 1 then simple regression modelling must be replaced by bounded regression modelling. In this case, and in previous applications, the claimed utility scores are a sham. Indeed, it is more than a waste of time, if the ultimate objective is to predict scores which have only ordinal properties; the ICER utilities created by the expert modeling group are nonsensical. They show fealty to a particular description of reality.

Both the SGRP and AQLQ fail to meet the standards of fundamental measurement; a situation common in disease specific outcome instruments. The reason is the same as that found with generic multiattribute instruments: no thought was given to the required measurement properties of the instrument. They were developed in the context of CTT; the thought that RMT might be a more useful framework for evaluating response apparently was not considered (if they were aware of RMT in the first place).

In the case of the SGRP the authors clearly had no idea of fundamental measurement. For each item in the question weights were derived from a visual analog (0 – 1) assessment of the various Likert scale items under two parts: perception of recent respiratory Symptoms (Questions 1 – 7), and the patient's current state (Questions 8 – 14). A further dimension was also proposed: an impacts component (Questions: 8,10,11,13,14) which can be combined to a soft measure. There is no doubt that the resulting score-addled SGRP has only ordinal properties; this can be traced back to the initial choice of Likert response scales and application of weights which have ordinal properties.

The AQLQ comes in a number of versions; all fail to meet required measurement standards creating only ordinal scales. The reason is simple: the scoring is based on Likert scales, none of which meet standards for invariance of comparison between the numbers assigned to those scales (from 7 = not impaired at all to 1 = severely impaired). The distance between 7 and 6 cannot be compared to the distance between 2 and 3 as we fail to have an interval

scale. Aggregating over ordinal Likert scores produces an aggregate ordinal score. This, as noted, cannot support parametric statistics as the distance between the aggregate scores is unknown. The scores only support ranking, nothing more. This is an unfortunate observation given the numerous applications and language versions of the AQLQ.

TEZEPELUMAB: THE EDIFICE CRUMBLES

How are the manufacturers associated with Tezepelumab (AstraZeneca, Amgen) to respond to the draft evidence report and its assorted outcome claims, culminating in recommendations for pricing and access? One option would be to refuse to engage, notably in terms of challenging the assumptions that drive the reference model. Unfortunately, AstraZeneca and Amgen will still have to deal with insurers and health system decision makes who may have a touching belief in the importance of the ICER contribution to pseudoscience. They have to be accommodated. They have to be shown, again, that planes fly; that magic carpets and broomsticks don't.

The key point in a rebuttal is that the ICER reference case fails to meet the standards for demarcation between science and pseudoscience: claims made lack credibility, empirical evaluation and replication. In short, no detailed rebuttal challenging assumptions is necessary, nor is it necessary to consider the relevance of generic QALYs to the target patient population. Challenging impossible constructs is a waste of time. The challenge should be on the standards of normal science and the axioms of fundamental measurement. This is not an easy task as we are dealing with 30 years of imaginary QALY claims and 'experts' who have been trained to accept constructed, imaginary evidence, ignoring the importance of focusing in specific attributes and overlooking the question of construct validity. To question the ICER reference case is to question their core beliefs. There will be pushback.

Consider the outcomes that are presented as the result of the ICER assumption driven imaginary simulation model. In this case the outcomes reported (Tables 4.4 and 4.5; p 21) for the base case tezepelumab plus standard of care versus standard of care are:

- Intervention cost
- Other nonintervention costs
- Total cost
- Quality Adjusted Life Years (QALYs)
- Life years (Lys)
- Expected value of life years (evLY)
- Expected Value of Life Years Gained (evLYG)
- Percent responders
- Cost per QALY gained
- Cost per life year gained

- Cost per evLY gained
- Cost per responder
- Health Benefit Price Benchmark

As should be apparent, none of these outcome claims are remotely capable of empirical evaluation let alone replication, and none meet required standards for normal science, including standards for fundamental measurement with single attributes. They are neither credible nor able to be empirically evaluated, but are driven by the choice of imaginary model structure and assumptions for the lifetime simulation, among thousands of other possible choices. We are still on Planet Dim where everyone takes the Dimwits word for the outcome claims.

Claims for lifetime costs are not intended to be evaluated; there is no possibility of linking these to the any units of resource use defined by CPT code. Thus, rather than opting for resource utilization claims (e.g., changes in hospitalization CPT codes) which, in a short term model may be able to be tracked and validated, ICER is perfectly aware that, over a 20 to 30 year timeframe that none can be contradicted other than by building competing models which in turn would have the same characteristic (including applying discount factors). That said, the application of costs as numerators renders the ratio product cost-per-QALY nonsensical: a ratio of two imaginary constructs. Claims for cost-effectiveness are just imaginary claims.

Given the fact that the preference scores, in this case the impossible mapped preferences from the ordinal, multiattribute AQLQ score, any claim involving these is also impossible. The mapping algorithm is nonsensical and the claimed utilities that are produce impossible constructs. The QALYs are, as noted above, also impossible mathematical constructs as time (a ratio measure) cannot be multiplied by an ordinal measure. In the case of the tezepelumab reference case modeling, it is not a question of preferences having ordinal properties but that the preference (utilities) are mathematically impossible.

Outcomes expressed as QALYs or combined with QALYs are unacceptable. This includes any ratio involving QALYs (with costs an imaginary construct) including the HBPB. The emphasis placed on the HBPB and by ICER and others who should know better, is a true red herring; although for ICER it is not meant presumably to be misleading or distracting. Anyone who thinks that the mathematically impossible threshold driven HBPM is a sure guide as an input to resource allocation across the US health system is sadly mistaken.

Note should also be taken of the evLY and evLYG outcomes; again lacking any possibility of empirical evaluation. As they rely on QALYS to support their calculation they are also to be ignored. Even so, they are worth considering in more detail as the involve the magic EQ-5D preference score of 0.851 in a (failed) attempt to ensure equality in claims. This is an adjusted perfect health score for the US population created by adjusting age and gender preferences

from the EQ-5D-5L preferences. The problem, once again, is that the EQ-5D-5L only yields ordinal scores. As an ordinal score that lacks, by definition, invariance of comparisons it is impossible to 'adjust' to create a US perfect health score; the 0.851 score is entirely arbitrary, lacking any coherent foundation.

The proportion and number of responders to therapy are determined by the model. As a single attribute this could be expressed in credible terms, but would have to be supported by a framework of assumptions regarding compliance and therapy uptake over the few months following product entry. As it stands ICER does not provide a protocol for how this claim could be evaluated. None of the ICER claims could be supported by a claims assessment protocol; a critical requirement if we are to have a realistic framework for evaluation therapy response; this approach is detailed in version 3 of the Minnesota formulary guidelines [26].

As the price for tezepelumab has yet to be determined, ICER makes one more fantasy assumption: that the placeholder price is the ICER invented net pricing of dupilumab ($27,859.88 per annum). At this placeholder price the imaginary QALY based threshold analysis yields annual prices to achieve imaginary costs per QALY of $6,200 for $50,000 per QALY ranging to $15,000 per year for $200,000 per QALY for those with uncontrolled severe asthma. Corresponding imaginary evLY based thresholds yield essentially the same prices. This implies, although ICER will not state an HBPM price, a discount in the annual price of tezepelumab ranging from 78% (at $50,000 per QALY) to 46% (at $200,000 per QALY).

ICER also promises that, on its imaginary budget analysis, substantial denial of care to those with uncontrolled severe asthma (an estimated 1.3 million) is inevitable. ICER assumes (one more assumption) 20% of these 1.3 million patients would initiate treatment in each of the five years of the budget impact analysis, or approximately 270,000 patients per year. On this basis, the purpose of the budget impact analysis, with its eugenics undertones for denial of care is, in ICER's words: *to document the percentage of patients who could be treated at selected prices without crossing a potential budget impact threshold that is aligned with overall growth in the US economy* (p. 29). The extent of ICER determined denial of access to tezepelumab is substantial:

> *Approximately 3.8% of the roughly 270,000 patients could be treated each year without crossing the ICER budget impact threshold of $734 million per year over five years at the annualized placeholder price of $27,860. At the three threshold prices (approximately $11,927, $9,077, and $6,226 per year of treatment, respectively) 10.9%, 16.5% and 37.6% could be treated with tezepelumab without reaching the potential budget impact threshold.*

Patients will, no doubt, be thrilled with this news. They should be informed that the analysis is entirely imaginary and denies the standards of normal science.

At this stage, for those wedded to the relativist world of ICER and the invention of evidence through assumption driven simulation models, we might remind them of the motto of the Royal Society (formerly the Royal Society of London for the Advancement of Knowledge) founded in 1660: *Nullius in verba*: Take nobody's word for it [27]. Clearly, in terms of the scientific revolution and the focus on progress in knowledge, ICER and the experts at the University of Colorado Anschutz Medical Campus failed to receive the memo.

CONCLUSIONS

Claims based upon assumptions and invented evidence (or its absence) have a distinctly medieval cast. It is surprising that this tribute to Aristotelian philosophy, doubting the possibility of new knowledge, attracts a willing audience of academics and health system decision makers, let alone journal editors, reviewers and the ubiquitous media. It is time we abandoned relativism and the beliefs of Planet Dim and the Dimwits in health care decision making. The fundamental difference between relativists and those with belief in the standards of normal science, with the focus on credible claims, empirical evaluation and replication, is that relativists, such as ICER and its academic cheer leaders, who believe truth is consensus fail to see *that this belief is incompatible with an understanding of one of the fundamental things that science does, which is to show that a consensus view must be abandoned when it is at odds with the evidence* [5] .

REFERENCES

[1]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

[2] 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

[3] Rind DM, McQueen RB, Herron-Smith S et al. Tezepelumab for Severe Asthma; Draft Evidence Report. Institute for Clinical and Economic Review, September 16, 2021. https://icer.org/wpcontent/uploads/2021/05/ICER_Severe-Asthma_Draft-Evidence-Report_091621.pdf

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

[5] Wootton D. The Invention of Science: A new history of the scientific revolution. New York: Harper Collins, 2015

[6] McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting Patient Reported Outcome Instruments and Interpreting the Data they Produce, *InovPharm*.2021;12(2):No 17
https://pubs.lib.umn.edu/index.php/innovations/article/view/3911/2764

[7] Langley P. The Impossible QALY and the Denial of Fundamental Measurement: Rejecting the University of Washington Value Assessment of Targeted Immune Modulators (TIMS) in Ulcerative Colitis for the Institute for Clinical and Economic Review (ICER). *InovPharm*. 2020;11(3).No.17 https://pubs.lib.umn.edu/index.php/innovations/article/view/3330/2533

[8] Langley P. To Dream the Impossible Dream: The Commitment by the Institute for Clinical and Economic Review to Rewrite the Axioms of Fundamental Measurement for Hemophilia A and Bladder Cancer Value Claims. *Inov Pharm*. 2020;11(4). No. 22 https://pubs.lib.umn.edu/index.php/innovations/article/view/3585/2642

[9] Feyerabend P. *Against Method*, London: 1975; Revised edition, London: Verso, 1988.

[10] 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

[11] Latour, B., and S. Woolgar, Laboratory Life: The Construction of Scientific Facts. Princeton: Princeton University Press, 1986

[12] Institute for Clinical and Economic Review. Lupus Nephritis. ICER-lupus nephritis-comments-folio_031221.pdf

[13] Seife C. Zero. New York: Penguin Group, 2000

[14] 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

[15] Thurstone, L. A law of comparative judgement. 1927; *Psychological Rev,* 34, 273-286.

[16] 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

[17] 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

[18] Bond T, Fox C. Applying the Rasch Model: Fundamental measurement in the human sciences. Routledge: New York, 2015

[19] ICER. 2020-2023 Value Assessment Framework. 23 October 2020 https://icer.org/wp-content/uploads/2020/10/ICER_2020_2023_VAF_102220.pdf

[20] 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

[21] Starkie H, Briggs A, Chambers M et al. Predicting EQ-5D values using the SGRQ. *Value Health*. 2011; 14:354-60

[22] Juniper E. Validation of a standardized version of the Asthma Quality of Life Questionnaire. *Chest*. 1999;115(5): 1265-70

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

[24] McKenna S, Meads D, Doward L et al. Development and validation of the Living with Chronic Obstructive Pulmonary Disease questionnaire. *Qual Life Res*. 2011;20(7):1043-52

[25] 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

[26] Langley P. Value Assessment, Real World Evidence and Fundamental Measurement: Version 3.0 of the Minnesota Formulary Submission Guidelines. *InovPharm*. 2020;11(4). No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/3542/2613

[27] Royal Society https://royalsociety.org/