MAIMON WORKING PAPERS NO. 20 SEPTEMBER 2021

NULLIUS IN VERBA: TRUTH IS NOT CONSENSUS

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

Abstract

The last 30 or more years have seen health technology assessment wedded to a meme that is a perversion of the standards of normal science. Rather than focusing on the discovery of new facts through a coherent structured research program for specific therapies in disease areas to fill evidence gaps and provide a sound basis for pricing and access decisions, groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Institute for Clinical and Economic Review (ICER), with the National Institute for Health and Care Excellence (NICE) in the UK as global role model, have insisted on inventing evidence. This commitment to approximate information rather than hypothesis testing has resulted in the creation of impossible mathematical constructs, such as the quality adjusted life year (QALY), embedded in nonsensical assumption driven simulation models that purport to guide formulary decisions; each simulation as one of many hundreds of possible competing assumption driven simulations producing diverse and contrary claims for pricing and access for the same therapeutic products. None of the simulations would meet the standards in normal science for credibility, empirical evaluation and replication. The purpose of this note is to address the key elements in re-educating those who subscribe to this relativist meme. There are five core learning issues. These are: (i) the axioms of fundamental measurement; (ii) the standards of normal science; (iii) assumptions and the problem of induction; (iv) value claims as single attributes; and (v) value claims protocols.

INTRODUCTION

In 1660 The Royal Society for the Improving of Natural Knowledge (now The Royal Society) was founded (Royal Charter 1662) with the motto: *nullius in verba* or 'take no man's word for it'. In the 360 years since the founding of the Royal Society, with this commitment to empirical knowledge, acceptance of the role of empirical evaluation has become the cornerstone of the natural sciences and the mainstream social sciences such as economics. In contrast, we find in the area of health technology assessment a perverse commitment to the invention of knowledge or, more accurately, the recirculation of old facts to support assumptions to drive imaginary simulation models to support invented claims for cost-effectiveness [1] [2]. This absence of intellectual curiosity, an explicit rejection of the notion of hypothesis testing in favor of the assembly of approximate imaginary information to support formulary submissions, is held by those practicing health technology assessment; an implicit commitment to relativism where 'anything goes' [3]. In this relativistic world view the content of science is to be explained sociologically; the so-called strong program. As detailed by Wootton this program follows from the principle of symmetry: the same sorts of explanation must be given for all types of knowledge claims whether they are successful or not [4]. It denies the feature that distinguishes science from pseudoscience (or pure bunk), the appeal to superior evidence. The strong program denies that science is a way of

coming to grips with reality; no one body of evidence is superior to another. Evidence is never discovered, it is constructed within a social community; the technology assessment meme is entirely in this domain. The research program that embraces the invention of comparative therapeutic claims depends not on its ability to generate new knowledge but its ability to mobilize the support of the health technology assessment community where their commitment to pseudoscience is about rhetoric, persuasion and authority; an outdated and, like Aristotelianism, a cramping system of belief that truth is consensus [5] [6].

This embrace of pseudoscience, the belief in the invention of evidence by assumption, possibly because the model builders believe the assumptions are realistic or reasonable, must be a key focus of attempts to re-educate individuals; a re-education of persons who may have a spent a professional career in promoting the pseudoscience of modeling claims in health technology assessment and have never questioned or even remotely thought about the limitations of fundamental measurement and elementary logic. This commentary focuses on five core learning issues that re-education must address: (i) the axioms of fundamental measurement; (ii) the standards of normal science; (iii) the problem of induction; (iv) value claims as single attributes; and (v) value claims protocols in formulary submissions.

AXIOMS OF FUNDAMENTAL MEASUREMENT

Conspicuous by their absence in the health technology assessment literature are any discussions of the levels of evidence, the limitations imposed by the axioms of fundamental measurement and, of particular relevance, conjoint simultaneous measurement and Rasch Measurement Theory [7]. Whether this absence is through a lack of appreciation of measurement theory or just by design, is an open question. In either case the result has been disastrous for those followers of the imaginary claims meme [8].

Science, and the more rigorous social sciences, can only progress if the question of measurement has been resolved. If we are concerned to capture accurately response to therapy, then the relevant instruments must have acceptable measurement properties whether they are focused on clinical markers, patient reported outcomes (to include quality of life) or resource utilization. Put simply, the objective must be to design an instrument that has ratio measurement properties (or at least interval measurement properties) for the outcome of interest [9] [10].

For those with a limited or more likely total incomprehension when it comes to the axioms of fundamental evidence, a short primer is in order. Following from the work of Stevens in the 1940s four main types of measurement scale (or levels of measurement) are recognized [11]. Nominal or categorical scales which report on distinct variables such as gender (male/female) where each variable has an equal value, so no numbers are involved. No statistical analyses can be conducted on such scales. However, tests (such as the non-parametric Chi-square) comparing the number of entries in different categories are possible. Next, ordinal scales show the order of responses to latent variables such as satisfaction, happiness or pain. However, they do not inform on the distance between scores on the instrument. It is not valid to calculate total scores, means or standard deviations with ordinal scales and parametric statistical tests should not be used.

Despite this, it is common practice to report, incorrectly, such statistics derived from ordinal scales such as the EQ-5D-3L. ICER in its modelling makes this mistake. Most PRO instruments yield ordinal data and consequently, are limited in how they can be used [12]. Perhaps the best example of the misapplication of ordinal scores is in the construction of QALYs [13] [14]. This requires multiplying scores on an ordinal scale by time. Interval scales show both the order of items in a scale and the distance between these variables. Valid means and standard deviations can be calculated with interval scales and parametric statistical tests can be used with data they generate. Addition and subtraction are possible with interval scales, but not multiplication and division. A ratio scale is like an interval scale but has a meaningful or convenient zero point that no values can fall below. Ratio scale data can be multiplied or divided by other variables; for example, distance travelled divided by time gives speed. Few PRO instruments measure at the ratio level.

The critical point that is overlooked by those coming from a classical test theory (CTT) background is the need from day one to construct fundamental measures. In CTT the data have primacy and any analysis is exploratory and descriptive of those data (e.g., econometric modelling); accounting for all the data. The model has to fit the data so a 'poor fit' results in model tinkering or reformulation. In Rasch Measurement Theory (RMT) the model has primacy; the data are required to fit the model with the focus on the size and structure of the residuals to ensure, if possible, for the construct of interest, that the principles of conjoint simultaneous measurement have been sufficiently realized in practice to justify that the results can be used as a measurement scale with invariant, interval measurement properties and, in special cases, ratio properties.

If no thought was given to required fundamental measurement properties when an instrument was developed, then it is impossible to attempt *ex post facto* to reverse engineer to an interval, let alone a ratio scale. The instrument is doomed to have ordinal scores with no possibility of further refinement to an interval let alone a ratio scale. This means that the generic multiattribute preference instruments and the multitude of disease specific instruments must fail to achieve interval or ratio measurement properties. The multiattribute preference scales lack dimensional homogeneity (or unidimensionality) and construct validity and, as ordinal scales, are impossible to support response to therapy claims.

At the same time the CTT approach to creating algorithms to yield preference scores on a bounded scale of 0 = death to 1 = perfect health (a misnomer) for the health states defined by a limited number of symptoms (the term 'attribute' is confusing as it implies they meet the required ratio measurement standard) is a waste of time. The result is as expected; the algorithm will yield negative scores. Scores greater than unity are excluded as the algorithms are designed to be capped at unity with decrements to capture states worse than perfect health. The classic example is the latest attempt to value EQ-5D-5L health states for the US with the 3,125 health state preference scores [15]. The result is that 20% or 625 health states yield negative scores.

From the perspective of the true believer (e.g., ICER) the implications of denying that multiattribute preference instruments and the companion multitude of patient reported disease specific instruments have ratio (or a least interval) measurement properties is catastrophic: the ICER business case collapses

together with any attempt at academic rigor for the various expert modeling groups in the US at universities and Colleges of Pharmacy, to include the University of Washington at Seattle, the University of Utah, the University of Arizona, the University of Illinois at Chicago, the University of Colorado and the PORTAL group at the Harvard Medical School.

STANDARDS OF NORMAL SCIENCE

After some 350 years, it seems odd that we need to remind those in health technology assessment of the standards of normal science. This may, of course, reflect a lack of awareness rather than a decision to abandon hypothesis testing in favor of approximate information or inventing evidence. The key term is 'demarcation' as proposed by Popper in the 1930s; what separates 'good science', a heterogeneous activity, from non-science, to include pseudoscience [16]. It is important to define our terms and our reasons for asserting that the health technology assessment meme fits the term 'pseudoscience': 'a collection of beliefs or practices mistakenly regarded as being based on scientific method' or as Pigliucci describes it 'rampant irrationality' [5]. To this we might add that the supporters of the meme are apparently unaware of the standards for normal science, where science makes progress by eliminating an increasing number of wrong theories; this is in contrast to pseudoscience where its 'theories' are so flexible that they have no explanatory teeth because they can include any possible observation. The boundary is not hard and fast as the term science includes a number of disciplines; what delineates them as science is the ability to produce and test hypotheses based upon systematically collected empirical data; an investigation of nature based on the construction of empirically verifiable theories and hypotheses [5]. A necessary component of science is the presence of coherent conceptual constructs in the form if theories or hypotheses. What distinguishes science from non-science is empirical evaluation. If a claim lacks the ability to be empirically evaluated then it fails (or is put to one side until the required data are accessible). The ability to distinguish science from pseudoscience should be part of science education and literacy; it apparently is not.

If we consider or deconstruct the currently held meme in health technology assessment its manifest failures, as judged by those who subscribe to the standards of normal science are surprisingly clear cut. First, the claims made for pricing and access completely lack credibility; second, the claims are impossible to evaluate empirically and, third, the claims cannot be replicated empirically. The lack of credibility is manifest in the failure to meet the axioms of fundamental measurement. Starting from the decision to value health states defined by bundles of attributes and symptom levels, and then to the ordinal nature of preference and then the impossible QALY, any attempt to define the value claims as credible simply evaporates. Add to this the logical error of lifetime simulation modeling where imaginary evidence is invented to make non-evaluable value claims. Finally, the possibility of a multitude of imaginary value claims from competing models renders the entire exercise worthless.

Judged by the current standards for normal scientific enquiry, the surprise is that all too many subscribe to a meme that is so easily exposed as a sham. This is not intended to suggest that this was deliberate, but an easy way out; unfortunately, the decision to abandon the rigor of hypothesis testing with credible claims, in favor of inventing evidence to support non-evaluable claims, Pandora's box was opened. This

revealed a multiplicity of opportunities to construct models that generated statements of probable cost-effectiveness and the invention of data for claims. This is the sham, with modelers asking clients to take their word for it, turning their backs on 350 years of science.

THE PROBLEM OF INDUCTION

It is not only a failure to grasp the importance of the standards of normal science and the limitations of fundamental measurement that have to be reprogrammed, but a failure to grasp a simple logical issue: the choice of assumptions and Hume's problem of induction [17]. Put simply, Hume's problem as first stated in the mid 18h century is, as Magee eloquently puts it: *The whole of our science assumes the regularity of nature – assumes that the future will be like the past in all those respects in which natural laws are taken to operate – yet there is no way this assumption can be secured. It cannot be established by observation, since we cannot observe future events. And it cannot be established by logical argument since from the fact that all past futures have resembled past pasts it does not follow that all future futures will resemble future pasts [18].* Claims for inductive predictions are psychological; one assumption about the future is equal to any other assumption about the future; we cannot justify the choice of assumptions by reference to the literature. This point was made clear by Russell over 100 years ago; inductive claims are by assumption [19].

Yet the meme and the ICER business case rest on assumptions. One model builder might claim their assumption about an unknown future is more realistic than another; this may be countered by a model builder committed to reverse engineer for an opposite conclusion. Attempting to justify assumptions, as ICER does by reference to the literature and the beliefs of key opinion leaders, is both illogical and a waste of time. To this should be added the obvious point that models built on assumptions cannot create credible claims. The only exception is if the assumption is an element to support a short-term, credible and empirically evaluable claim; if the claim fails we can then consider the impact of our choice of assumption. Looking forward 20 or 30 years is clearly nonsensical; even to the extent of promoting sensitivity analysis and probabilistic sensitivity analysis (PSA) to create a false impression of the likelihood of a product being cost-effective for the more gullible and less informed audiences. Any number of competing assumptions can support any number of PSA competing likelihood claims with formulary committees faced with a choice of imaginary outcomes. The challenge in re-educating a belief in inductive reasoning will be difficult to overcome; the psychology of a belief system that has accepted for 30 years the choice of 'reasonable' assumptions, drilled into the belief system by professional associations and academic post-graduate centers, that cannot support quasi-realistic and reasonable claims over decades into the future is a challenge. Add to this a common feature of religious beliefs or viruses of the mind; Tertullian's *Certum est quia impossibile est* (it is certain because it is impossible) [20]. Perhaps a fitting epitaph for the QALY.

VALUE CLAIMS AS SINGLE ATTRIBUTES

It is difficult to believe that the commitment to valuing 'health states' was ever taken seriously; this applies to both multiattribute ordinal preference scores such as those produced by the EQ-5D-5L

instrument as well as multidimensional patient reported outcome disease specific instruments. Attributes can only be combined under certain conditions; each attribute must represent a coherent construct and the attribute measure must have ratio or interval properties. If single attributes have met these requirements then it seems pointless to combine them into a 'health state' construct with 'generic' application and ask that the bundles can be assigned a preference by a community sample. A more sensible approach is to provide a measure for an attribute and, depending on the disease state, base a separate value claim for that attribute and others that are relevant to the target patient population or the disease area. If depression is an attribute then it should be presented as a latent construct with ratio measurement properties; applying tariffs or preference scores to the presence of depression on a three response level ordinal scale is nonsensical. Formulary committee may advise on the required attributes for the therapeutic area.

Focusing on single attributes points to the futility of aiming for a single 'global' metric to support cost-effectiveness claims and resource allocation in health systems (with unfortunate eugenic implications) [21]. This has been the rationale for the QALY, a construct that fails to meet measurement standards, and the integration of the impossible QALY into an assumption driven simulation, which then fails the standards of normal science. ICER claims are not intended to be empirically evaluable. It is a win-win debacle.

Claims for future costs are a classic case of misapplication. Costs are only relevant is we have detail on the cost decisions within a health care system that are intended to be applied over 20 or 30 years for resources identified by CPT codes or their equivalent, including drug units identified by NDC code. This is clearly impossible. What is possible is for value claims to be expressed in terms of resource units to allow empirical assessment. Each resource unit will be an attribute with ratio properties so that the manufacturer can claim short term resource savings defined by selected codes. It is then possible for the formulary committee to translate these to short term costs. The same argument applies to attributes defined for drug utilization, switching and compliance. Expressed as single attributes these can be evaluated empirically with real world data following product launch.

VALUE CLAIM PROTOCOLS

If a manufacturer makes value claims for its product then it is appropriate that the manufacturer presents a protocol detailing how the value claim is to be assessed and reported to the formulary committee in a meaningful timeframe. This does not mean that the assessment has to occur for each health system, but that an assessment will be undertaken and reported to the relevant formulary committees. Neither does this exclude the possibility of a re-application of a Phase 3 protocol in a target patient population, with due regard for time constraints. As noted, value claims should be presented as single attributes; the protocol may involve mining real world evidence to support the value claim; otherwise it may involve an observational study. Unless readily available, including Phase 3 trial data, proposing to undertake a Phase 4 trial may be outside a reasonable timeframe for value assessment. All value protocols must relate to credible claims, a process for empirical evaluation and replication, and reporting.

CONCLUSIONS

In a very real sense the challenge of re-educating the many believers in the static health technology assessment meme is similar to the challenge faced in the 17Th century by natural philosophers such as Gilbert, Galileo and Descartes to overcome Aristotelian philosophy and the belief that there was nothing left to discover. Rather, as prisoners of the Aristotelian system, the scholastics of the 14th century saw their efforts as filling in gaps and removing inconsistencies in a generally accepted world picture. It was not until the 15th century with the contributions of Copernicus, Oresme, Vesalius and Brahe that questions on the movement of bodies through time and space emerged to question the Ptolemaic system and lay the foundations for the scientific revolution, which came to fruition in the period after the mid 17th century with the Newtonian synthesis; a belief in progress and the power of human reason that finally rejected the Aristotelian system.

In challenging the existing health technology focus on inventing evidence, the difference is, of course, not on overthrowing a universal Aristotelian belief system, but to convince members of a sect that that has turned its back on the standards of normal science and embraced the pseudoscientific equivalent of intelligent design in its commitment to approximate imaginary information, that they are embracing an analytical dead end. The fact that this rejection of normal science was explicit and almost immediately embraced by the nascent occupation of health technology assessment points to both the strengths and weaknesses of the paradigm, with its high transmission fidelity. It is clearly from the assessor's perspective, straightforward and lucrative. After all, no assumption driven claims can be empirically assessed and, to many, it is nothing more than a marketing tool; the latest incarnation of the 19th century snake oil sales pitch with the common theme of the absence of empirical evidence to support claims. Against this belief that supports the ICER and other business cases, is the fact that just scratching the surface reveals manifest deficiencies. There is no defense to the case presented above. ICER has attempted in its convoluted arguments that the true believer recognizes that ordinal scores have ratio properties and that even with the absence of a true zero, ordinal preferences can support the construction of QALYs; this is easily deconstructed. But yet, the ICER imaginary belief system holds firm, with funding from organizations in health that should know better and a multitude of true believers. Our task must be to convince decision makers that under no circumstances should they take ICER's word for it.

REFERENCES

[1] Neumann PJ, Willke R, Garrison LP. A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *Value Health*. 2018;21:119-123

[2] 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

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

[4] Wootton D. The Invention of Science: A new history of the Scientific Revolution. New York: Routledge, 2015

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

[6] Briggs R. The Scientific Revolution of the Seventeenth Century. London: Longman, 1971

[7] Bond T, Fox C. Applying the Rasch Model: Fundamental Evidence in the Human Sciences. New York: Routledge, 2015

[8]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

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

[10] Grimby G, Tennant A, Testo L. The use of raw scores from ordinal scales: Time to end malpractice (Editorial) *J Rehab Med*. 2012;l44:97-8

[11] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-680

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

[13] 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

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

[15] 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

[16] Popper K. Logik der Forschung. 1934 [English translation 1959]

[17] Hume D. A Treatise of Human Nature (1739-1740)

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

[19] Russell B. The Problems of Philosophy. 1912

[20] Dawkins R. A Devil's Chaplain. New York; Houghton Miflin, 2004

[21] 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