MAIMON WORKING PAPER NO. 3 APRIL 2021

THE HERD AND THE MEME: WHY ORDINAL SCALES HAVE TO BE RATIO MEASURES IN DISGUISE

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

ABSTRACT

The fact that for over 30 years, the discipline of health technology assessment has rested on a belief that ordinal scores are ratio measures in disguise may come as a surprise. While the practitioners in the discipline may not recognize this representation of their belief system, given their apparent lack of knowledge of the axioms of fundamental measurement, there is no doubt that there has been a deliberate decision to reject the standards of normal science. The purpose of this brief note is to demonstrate how this has transpired and why, contrary to common sense, it continues to this day.

INTRODUCTION

In what must be one of the most bizarre and continuing episodes in the history of the social sciences, is the continued embrace by those in health technology assessment of a meme or belief system that rejects the standards of normal science in comparative cost-effectiveness claims. This embrace is of long standing; indeed it goes back over 30 years. The reasons for it are reasonably clear: the role of quality adjusted life years (QALYs) in assumption driven modeled simulations. In the case of formulary committees, who typically accept this belief system, the basis for cost-effectiveness claims is the incremental cost-per-QALY calculus and the application of cost-per-QALY thresholds to support pricing recommendations and access.

Hypothesis testing is rejected in favor of decisions based on imaginary approximate information; not approximate information *per se* but invented imaginary information on cost-per-QALY claims stretching decades into the future [1]. Models which have a tenuous link to reality in the form of limited, protocol driven evidence from pivotal clinical trials and limited evidence to create some assumptions from the literature. The rest is guesswork. Yet these constructs are believed, forming the basis for dominant textbooks in the discipline [2] . Imaginary model based recommendations for social pricing and patient access by the Institute for Clinical and Economic Review (ICER) in the US are lapped up by the media and formulary committees where the knowledge of how these imaginary claims are generated is minimal [3]. Standing over this, with its global reach as guardian of this meme, is the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Welcome to the world of imaginary formulary decision making in the 21st century.

NORMAL SCIENCE

Science is distinguished from pseudoscience by a commitment to the discovery of new yet provisional facts [4]. For claims to have any merit they must be credible, empirically evaluable and replicable. This establishes the demarcation line between science and pseudoscience; health technology assessment joins, with intelligent design, the pseudoscience fraternity [5]. This rejection of normal science is the essence of the meme; the meme has significance, the transmission fidelity is high and the 'mystery' of the QALY with ratio properties a key part of this fidelity. Indeed, as with many religions, the less plausible a 'mystery' claim the greater the strength of belief. Challenging the belief is that more difficult. To this extent, the meme in technology assessment is best seen as a sociological phenomenon. Science is about rhetoric, persuasion and authority; rationality is culturally relative. If claims are made we should seek a psychological or sociological explanation; refutation by an appeal to evidence is illegitimate. This relativist position misses the point of what science is about. It is also nonsense.

If we are to answer unresolved questions regarding claims for cost-effectiveess then the focus must be on evidence platforms to support hypotheses regarding one or more single attributes of competing therapies. This is relatively straightforward when clinical claims are being assessed that relate to measurable quantities with agreed standards. It becomes more complex where latent attributes are concerned. We should, however, reject self-reporting symptom based instruments that attempt to capture a complex of attributes; these are not only dimensionally heterogeneous, they lack unidimensionality and construct validity. Notions of instruments to capture health related quality of life (HRQoL) should be rejected, notably direct and indirect preference instruments. These create scores not measures. This is a critical distinction: the term 'score' refers to ordinal scales while the term 'measure' relates to interval and ratio scales.

EVIDENCE IS INVENTED

While inventing evidence, or approximate information, may have seemed a novel response to the limited evidence to support cost-effectiveness claims at product launch, it has created a potential nightmare for its supporters. How are its supporters to be informed that they have been misled? There are now thousands, if not tens of thousands of papers, commentaries and reviews that have embraced the imaginary technology assessment meme. Add to this the more gullible health technology assessment agencies, the professional groups such as ISPOR who have rigorously promoted the meme with its practice guideline publications, and groups such as ICER who have enjoyed media success with its imaginary pricing and access recommendations. Last but not least are the academic centers who have trained some few generations of graduate students in this fatally flawed methodology.

Modelling is a revered part of the meme; models driven by assumptions making unprovable claims about unknown future realities. While this has obvious religious parallels, the problem is that if the model is driven by assumptions a multitude of other models may follow. Zipping into existence like virtual particles but not disappearing. One defense against a multitude of lifetime simulation models is

to establish reference guidelines with a monitoring police force or inquisition. In the UK, for example, NICE engages with academic centers with extensive experience of reviewing imaginary models to award a good housekeeping seal of approval to manufacturer's submissions. ICER in the US has gone in another direction by establishing a cloud simulation base where its models can be re-imagined to provide a discussion on 'appropriate claims' [6]; both illustrating the futility of inventing evidence to support cost-effectiveness claims.

THE QALY HOLY GRAIL

The fundamental error, which goes back to the advocacy of standard gamble and time trade off techniques in the 1970s and 1980s was the belief that, adopting notion of preference or utilities, it was possible to combine health state descriptions and generate a single score that had ratio properties. A single utility score that, on a scale of 0 = death and 1 = perfect health could qualify time spent in a disease state, translating it to equivalent time in perfect health. This was doomed from the outset. If health state symptoms and the current response to those symptoms, as assumed ratio scales, could be combined this, it was believed would create a score, a health related quality of life score (HRQoL) that would support the creation of QALYs. This, from a measurement theory perspective, was a false belief. Each symptom or attribute in such a composite measure has its own dimensional characteristics. Combining results in a dimensionally heterogeneous score, lacking dimensionality and construct validity, inevitably the result was an ordinal score not a ratio measure. This was, essentially, game over: if clinical attributes are important in formulary decisions, then each attribute, as applied in the physical sciences, must be designed to have its own measurement properties [7]. Asking patients to 'score' a collection of attributes and responses was equally nonsensical. The holy grail of a composite multiattribute HRQoL ratio scale was nothing more than a chimera.

TRUTH IS CONSENSUS

But why did the attraction to HRQoL scores, despite repeated criticism, continue; why were these doubts brushed aside? The reason is straightforward; evidence for cost-effectiveness had to be invented based on the universal belief in the contribution of QALYS. Preference scores such as the EQ-5D-3L/5L are seldom found in clinical trials. Even if they were their ordinal properties would disallow QALYs. So, everything has to be invented. The literature is scoured through systematic reviews (or less systematic reviews) for preference score to plug into models. Some have a fairly laid back view and take any available score irrespective of the multiattribute instrument. After all, if you are constructing an imaginary scenario, slightly dodgy assumptions are part of the game. Outside of the brigade of academic reviewers, few will challenge the assumptions of a model. In any event, the claims can never be empirically evaluated and the validation that does occur involves comparing one model to another, which seems a pointless exercise.

The bottom line is that truth in health technology assessment is by consensus. If you subscribe to the standards (or lack of) that characterize the meme, you will find a publication more than willing to accept your manuscript for peer review. The editor will accept your 'publication' fee and, following peer review

by other meme believers, ensure publication. If you step out of line the consequences are obvious. The fact that the impossible or I-QALY is a constrict that relies, in defiance of the axioms of fundamental measurement, on multiplying an ordinal utility score by a ratio time score is nothing more than a minor and easily overlooked irritant [8]. We will press on regardless or until the objections become indefensible.

THE MADNESS OF CROWDS

The current framework for cost-effectiveness claims requiring the I-QALY and assumption driven simulation incremental cost per I-QALY models is an intellectual dead end. While the news has yet to percolate down to the thousands of analysts who hold tenaciously to this belief, the case for rejection is overwhelming. The primary reason is the failure to appreciate (or willingly ignore) the limits imposed by the axioms of fundamental measurement. They can, of course, be ignored but then there is the risk of ridicule; those trained to believe may want their money back. It is not as if the meme has only just encountered criticism and has yet to gird its loins and respond: the admonitions against thinking ordinal scales are ratio scales go back over 30 years, with Rasch Measurement Theory developed some 60 years ago [9] .

Whether we should characterize the current meme in health technology assessment as postmodernism in the tradition of Derrida and Foucault is an open question. Certainly, its adherents subscribe to a belief system that for those of us who subscribe to the more traditional view of the standards of normal of science,as heirs to the scientific revolution of the 17th century and more recently the contribution of Popper to the process of discovery, conjecture and refutation find, to be charitable, weird. After all why go to the trouble of constructing an assumption driven future comparative scenario, with many sub-scenarios, capped by the application of probabilistic sensitivity analysis where none of these claims, probabilistic or otherwise, can ever be evaluated empirically? Even if we claimed the moon was made of green cheese, at least it could be empirically assessed (and has been; the answer is no, it's actually mature stilton).

A NEW PARADIGM

It is most unlikely that the existing technology assessment meme or paradigm will be overturned any time soon. This is not the question of a preceding paradigm being augmented and absorbed within the new, but of a complete rejection of the previous one. This will not be easy. In fact, the only avenue forward (given the 'egg on face' issue) is to reject QALYs as part of formulary decision making. This rejection, as has occurred in the US, may not be because of abstruse (i.e., somewhat difficult) arguments based on axioms of measurement theory but for more prosaic reasons that the I-QALY fails to capture, among other elements, disability. It is morally tainted and as such has no place in ethical formulary decision making. This opens the door to claims for alternative measures which are patient centric, rejecting generic oversimplifications, in favor of disease and target patient specific assessments of patient needs.

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 Revaluation of Health Care Programmes. 4th Ed. Neww York: Oxford University press, 2015

[3] 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

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

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

[6] Langley P. Let a Thousand Models Bloom: ICER Analytics Opens the Floodgates to Cloud Pseudoscience. *InovPharm*.2021;2(1): No. 5 https://pubs.lib.umn.edu/index.php/innovations/article/view/3606/2668

[7] 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

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

[9] Bond TG, Cox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 3rd Ed. New York: Routledge, 2015