MAIMON WORKING PAPERS No. 4 2022

AN ORDINAL OVERSIGHT: ABANDONING THE TUFTS MEDICAL CENTER COST-EFFECTIVENESS ANALYSIS (CEA) DATABASE

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

Abstract

One of the more disturbing issues with health technology assessment is the tenacity with which analysts believe in the construction of assumption driven imaginary simulations and the importance of these non-evaluable model claims to formulary decisions. This is epitomized in the recent publication of the CHEER 22 guidance for assumption driven simulations. It has been pointed out on numerous occasions that this belief, promoted by professional groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) defies the standards of normal science, notably the axioms of fundamental measurement. A key element in this belief system or meme is role of ordinal preference scores. These play a central role in the creation of incremental cost-per-QALY claims with supported of the approximate information modeling apparently also believe, incorrectly, that these preference scores have ratio properties; which they certainly do not. The result is the impossible QALY where time spent in a disease state is multiplied by an ordinal preference scale. This is mathematically impossible. The result is a disaster for model building. At the same time the neglect of the axioms of fundamental evidence is evidenced in the Tufts Medical Center Cost-Effectiveness Analysis (CEA) database which, for the past 46 years has actively promoted the belief that ordinal preference scales can support cost-per-QALY claims. The purpose of this brief commentary is to make the case that the Tufts CEA database with its pay-for-access health state preferences to support model building and non-evaluable value claims is not only irrelevant but an obstacle to a new start in health technology assessment that abandons the CHEERS 22 methodology for one that meets the standards of normal science. With over 10,000 studies featured in the Tufts CEA database since 1976 and, more recently the TUFTS GH-CEA disability adjusted life year (DALY) database, where the DALY suffers the same fate as the QALY, there will no doubt be a backlash; the sunk costs and the reputations of these database are too important for the developers to abandon them without objections.

Keywords: Tufts CEA, Tufts GH-CEA, abandoning Tufts CEA, Tufts ordinal error

INTRODUCTION

The imperative that drives health technology assessment is the commitment to an endless future where analysts produce one assumption driven modeled imaginary value claim after another for the same and multiple pharmaceutical products and devices; the Tufts CEA is a willing partner in this commitment in providing ready access to generic ordinal preference scores to support imaginary incremental cost-per-QALY claims. This commitment to imaginary claims sets health technology

assessment apart and in a unique position among the physical and mature social sciences in promoting the importance of basing formulary and other health care decisions on imaginary, and by definition and intent, non-evaluable value claims covering timespans of decades into the future. This has been, given the Tufts CEA timeframe, the commitment of the approximate information meme for almost 50 years. After over 350 years since the founding of the Royal Society of London (1660-2) with its motto *nullius in verba* (take nobody's word for it), the Tufts CEA has taken the backward step of supporting a meme that insists we take somebody's word for value claims in health care decisions that can never be empirically challenged.

The Tufts University Medical Center's Cost-Effectiveness Analysis (CEA) data base was established in 1976 [i]. Since then the database has grown to include some 10,000 studies involving some variation of cost utility analysis and QALY belief. The common element has been the present of preference or utility scores, typically from multiattribute instruments such as the EQ-5D-3L/5L with the database reporting on these together with various preference ratios and health state utilities. In addition, there is the more recently developed disability adjusted life year (DALY) GH-CEA database which exhibits the same fatal flaws as the Tufts CEA database. The health state utilities from the Tufts CEA have been used widely with the database receiving commendations for its contribution.

The purpose of this brief commentary is to make the case that both the Tufts CEA and Tufts GH-CEA database are deeply flawed; to the extent that both are irrelevant and, from the standards of normal science, should be abandoned. These databases, despite the efforts put in to create and maintain them, are irrelevant in value claims and formulary evaluations and always have been. The fundamental error has been to ignore the axioms of fundamental measurement and assume that preference scores have ratio properties. In fact, as it has been conclusively demonstrated, these preference scores are ordinal and cannot support claims for response to therapy applying the standard techniques of statistical analysis, let alone the creation of quality adjusted life years (QALYs) [ii] [iii] [iv]. The result is obvious: if the QALY is an impossible mathematical construct then assumption driven simulation models, where health state ordinal utilities drive the false mathematically impossible cost-per-QALY claims by stage of diseases, also have to be abandoned [v]. There is no place for either of these two databases in technology assessment.

It should be noted that this assessment of the Tufts CEA rests, not just on the denial of the axioms of fundamental measurement, but on the mistaken belief in its contribution to the creation of imaginary evidence from assumption driven modeling simulations. These, as noted previously, defy the standards of normal science as they ask formulary committees to base decisions on imaginary value claims rather than on value claims that are credible, evaluable and replicable. Denying this standard puts the imaginary value claims modeling on the non-science side of demarcation, relegating it to the world of metaphysics and pseudoscience to join intelligent design and other odd belief systems.

FUNDAMENTAL MEASUREMENT

Given the apparent lack of understanding of the axioms of fundamental evidence by those charged with developing and promoting the Tufts databases, it is worth making clear what the axioms support. Following the formalization by Stevens and others in the 1930s and 1940s, scales or levels of evidence used in statistical analyses are classified as nominal, ordinal, interval or ratio [vi] . Each scale has one or more of the following properties: (i) identity where each value has a unique meaning (nominal scale); (ii) magnitude where values on the scale have an ordered relationship with each other but the distance between each is unknown (ordinal scale); (iii) invariance of comparison where scale units are equal in an ordered relationship with an arbitrary zero (interval scale) and (iv) a true zero (or a universal constant) where no value on the scale can take negative scores (ratio scale). Nominal and ordinal scales do not support any parametric statistical operations; only nonparametric statistics. Interval scales can support addition and subtraction while ratio scales support the additional operations of multiplication and division as they have a true zero. This zero point characteristic means it is meaningful to say the one object is twice as long as another. Given these limitations, the only acceptable empirically evaluable value claims are those designed for single attributes with interval or ratio properties.

The reasons for labelling multiattribute preference scores ordinal scores (including standard gamble and time trade off) is clear cut: to support multiplication or division you need a ratio score; no other measure will suffice. This means that the Tufts CEA will have to demonstrate (which is impossible) that the various preference scores in the database all have ratio properties. In addition, as these preference scores are proportions they will have to demonstrate that the preferences scores are bounded ratio measures with a cap of unity and a true zero. As the preference algorithms can produce negative values (states worse than death), they lack a true zero. The cap of unity is fixed because the preference scores are decrements from unity, but overshoot (despite econometric tweaking) to have a true zero. This means that in any disease state or disease stage, there is the possibility that all respondents to the respective questionnaire could report negative preference values or states worse than death. In addition, there is the issue of reporting preferences as averages. This is, of course, disallowed as these are ordinal scores; even so, averages are presented which may include negative values. It is not clear how these are incorporated as there is no presentation of the distribution of score only, where reported in the original study, distribution summaries. As the Tufts CEA database reports on the range of preference values, it would be of passing interest to see how many studies actually produce distributions of ordinal scores and the number of respondents with negative health states.

Composite scales attempting to bundle attributes are disallowed unless they are constructed from ratio scales. One example is body mass index which comprises the two ratio scales of height and weight. This immediately rejects preference scales as they are bundled symptoms with each symptom scale itself having ordinal properties. In health technology assessment physical attributes would be expected to have ratio properties, which allow value claims assessment from randomized trials and other real world data as long as they meet criteria for single attributes. Value claims that attempt to

capture latent attributes are more difficult to measure as the instrument for data collection must be designed to generate either ratio or interval properties as demonstrated by Rasch Measurement Theory[vii] . Clearly, imaginary claims are rejected, to include blanket claims for incremental cost-per-QALY, cost-per-QALY thresholds, probabilistic sensitivity analysis and overall cost-effectiveness.

THE TUFTS CEA DATABASE

The inputs to the Tufts CEA database are detailed in the data dictionary. The key ratio variables of interest in in this misapplication of preference scores in economic evaluations and the creation of impossible QALY claims are: QALYs, direct medical costs and total costs, net inputs of costs and QALYs, QALYs and costs per person, per population costs and QALYs, incremental cost-per-QALY and uncertainty (including graphical analysis, probabilistic sensitivity analysis). As the modelling is imaginary, driven by assumption, none of the various cost and QALY estimates are empirically evaluable; or at least there is no indication that the possibility of empirical evaluation was a criterion for selections to the data dictionary. Central to these various ratios is the mathematically impossible QALY. However the ratios are constructed and interpreted, they are rendered irrelevant by the failure to recognize the ordinal properties of preferences or utilities (let alone imaginary costs). They make no contribution whatever to the possibility of empirically evaluable value clams for response to therapy. The attempts to present uncertainty analysis are also misplaced as the claims for cost-effectiveness which underpin these various tools, notably probabilistic sensitivity analysis and the reporting of thresholds to support pricing recommendations should simply be abandoned. They have, from an analytical perspective, zero information content. This error is maintained in one of the leading textbooks in economic evaluations in health care; it is, in fact, a primer for the construction of imaginary cost-effectiveness claims [viii] .

The data dictionary has a separate section to describe the ordinal reference weights, to include the health state defined by ICD-10, the utility weight range, direct versus indirect weight elicitation and weight elicitation method, which identifies the preference instrument (Standard gamble, Time trade off, clinical judgement, rating scale EQ-5D [3L/5L], SF-36/12/6D, HUI [3 versions], VAS, non-generic scale). There is no reference to required ratio measurement properties for each of the elicitation methods, including the requirement for interval properties for VAS scales.

These two variable lists for ratios and QALYs does not mean that the elements are captured for all studies. In most cases the lists will be incomplete with search algorithms coming up short. Even so, while the range of information collected in the data dictionary is comprehensive, allowing the user to identify specific ordinal health status weights considered appropriate for a modeling exercise. Two points are worth noting: in almost all case the descriptions or the symptoms comprising the health state description are multivariate. The EQ-5D-3L, for example, incorporates in the preference algorithm five symptoms or attributes and three response levels for each attribute. This ensures dimensional heterogeneity, a lack of construct validity and an inappropriate combination of ordinal scores for each symptom to create an overall composite ordinal score [ix]. There is the further question of whether this generic ordinal score is appropriate to evaluating response in the specific health state.

Unfortunately, if the focus is on disease or health state specific measures for quality of life are considered these typically fail to meet the required measurement standards creating, once again, ordinal composite scores.

It should be noted that, in logic, with an unknown future addressed by the modeling, the fact that a specific utility weight (e.g., by disease, age, gender, EQ-5D-3L) is captured does not mean that the fact it has been reported (as a single observation from a given study) can be used to justify future imaginary claims on the basis of being 'realistic'. This violates Hume's problem of induction: the fact that all past futures have resembled past pasts does not mean that future futures will resemble future pasts. In this sense the justification for the choice of a utility on the grounds it has been applied before is irrelevant; you could choose any other utility and it would be equally valid as a personal choice or just an informed guess. The fact that the weights are reported as proportions means that any difference is usually minimal and that, as seen clearly in the dozens of imaginary assumption driven models by the Institute for Clinical and Economic Review (ICER) the cost-per-QALY threshold assessment is dominated by differences in costs to give often astronomical imaginary claims for cost-per-QALY.

There is a further issue of competing preference scores defined by the same search algorithm, or indeed, different scores produced by the same preference algorithm in different target populations with the same characteristics. This is of interest, for example, in the application of the EQ-5D-3L and EQ-5D-5L as these produce quite different scores for populations with the same characteristics and even the same population. Once again, as ordinal scores the ability to make comparisons is limited to the application of non-parametric statistics; averages and standard deviations are disallowed.

The issue of negative values is also of interest. As noted, the various preference measures can create negative scores (excluding constrained VAS scales). Even with the VAS, there is no implication that the scale has interval properties for invariance of comparisons. The absence of this property applies across the board for the simple reason that in instrument development it was not considered a property that had to be built into the instrument from the get-go. The presence of negative values means, as noted, that there is no true zero. Even if a range for a particular study does not produce negative values, health states worse than death, it does not mean that for another application the same algorithm could produce negative values. When we are dealing with latent constructs, it is difficult to create a scale for a single attribute that has ratio properties; or, in the case of the application of proportions, a bounded ratio property. Recently, the creation of a bounded ratio score for need fulfillment quality of life as a latent attribute has been achieved, but this is a special case [x] The role of need fulfillment interval quality of life scores, which have been developed for some 30 disease states over the past 25 years is not a consideration that the Tufts CEA recognizes [xi].

It is not clear from the data dictionary whether any advice is presented to advise on the preference scores created by the different direct and indirect preference instruments (e.g., EQ-5D-3L/5L), HUI Mk2/3, SF-6D, SG, TTO). The various systems are quite different in the dimensions of health covered, symptom levels and their descriptions, definitions of severity, the populations surveyed for the

scoring and their algorithms. Presumably, if you are creating an imaginary set of value claims then they should be based on one set of preference scores, not a mixture. If so, then there should be a choice of a common ordinal preference score for each health state (e.g., the EQ-5D-3L as this is most widely used) with a crosswalk the other preference scores to this common core. This, of course, is not only mathematically impossible as we are dealing with ordinal scales but there is unlikely to be sufficient evidence presented in a study to even attempt to crosswalk. Users may, inadvertently perhaps, look at the smorgasbord of ordinal utilities and pick those that conform to the health states and stage of disease in their model without appreciating the differences.

Given that the individual study models summarized are, for the large part, capturing the lifetime or natural course of disease there is not surprisingly no mention in the database as to whether or not any value claims from the modeling are empirically evaluable. The fact that the preferences supporting the QALYs are ordinal means that the entire exercise and reporting is irrelevant, although certainly in the CHEERS 22 tradition. While the database may report cost and QALY ratios does not mean they have any relevance to real world decision making; unless formulary committees are willing to accept imaginary value claims.

DISTRIBUTION OF TUFTS WEIGHTS

The Tufts webpage allows a review of the weights presented for 100 health states; whether these are representation of the entire weights database is uncertain. Putting to one side the possibility that the weights come from the various direct and indirect preference scales which render comparisons impossible (but may be overlooked by users). the distribution of weights is of some interest. Of the 100 entries accessible through the webpage, 47% of health states have negative values or 'states worse than death'. The range of negative health states is from -0.01 to -0.55; the range for positive weights is from zero to 0.93. As these are averages, presumably, of individual responses they are technically false as only a ratio measure can support addition and division. At the same time, averaging incorrectly over individual responses implies that the underlying distribution of scores can encompass both all negative values or a combination of negative and positive values; demonstrating once again the impossibility of a true zero which would require only positive scores under all circumstances. As the preference scores are ordinal (unless you believe with ICER in a mystical ordinal scale with ratio properties) this raises the intriguing possibility of negative QALYs. It is not clear how these would be introduced into a simulation cost-per-QALY model. Indeed, the presence of negative scores for individual respondents raises the intriguing possibility of calculating QALYS at the respondent level before aggregating the results to an overall average QALY; all of which is mathematically impossible although the ranked distribution may be of interest. With such a high prevalence of negative scores it is surprising that after 46 years that no one in the Tufts Medical Center thought, apparently, about the implications of this; the insights negative scores give to the underlying measurement properties and the requirement for a bounded ratio scale to create QALYs.

THE TUFTS GH –CEA DATABASE

Associated with the TUFTS CEA database is the Tufts GH CEA Database [xii]. This is a database of the world's largest collection of cost-per-disability adjusted life year (DALY) studies. Funded by the Bill and Melinda Gates Foundation, the application of cost-per-DALY claims to estimate the resources used and health benefits gained by alternative public health interventions. The database comprises studies that have an original cost/DALY estimate as the measure of health effects, with data collected on over 40 items for each study.

The DALY has become the preeminent framework for evaluating the burden of disease in regions and countries and the social impact of disease prevention strategies in both the developed and developing worlds. The DALY which captures years of lives before premature death (YLL) and years of life lost to disability (YLD) is taken as a potential marker for disease states and the impact of targeted programs The cornerstone of the YLD component has been the reliance on disability weights Unfortunately, despite increasing attention given to recalibrating disability weight estimates, the GBD disability weight calculus suffers from a fatal flaw: a failure to recognize the limitations imposed by the axioms of fundamental measurement. The oversight has significant implications. The disability weights have only ordinal measurement properties which means that YLD is a mathematically impossible measure. In turn, this means that the DALY is also mathematically impossible, unless restricted to the YLL component. If true, it goes without saying that the GBD project rests on untenable foundations and the GH-CEA database is irrelevant.

A NEW START IN FORMULARY DECISION MAKING

The manifest deficiencies in modeled cost-effectiveness analyses give no option but to reject the approximate information meme, as epitomized by the recent release of the CHEERS 22 guidance for imaginary claims, in its entirety. The term 'cost-effectiveness' is redundant; it has no information content. It fails the standards of normal science. This analytical dead end can be traced back for some 30 years where the leaders in the nascent subject of health technology assessment opted for rejecting hypothesis testing in favor of approximate information and the invention of value claims as a response to evidence gaps art product launch [xiii].

The case for a new start in the information required for formulary decisions has been made quite clear: the focus must be on empirically evaluable, single attribute, unidimensional value claims that meet either ratio or interval measurement standards; preferably the former. This requirement has been detailed in Version 3.0 of the Minnesota formulary guidelines [xiv]. These claims can refer to clinical endpoints, patient reported outcomes as well as resource utilization and even compliance with therapy. The days of cobbling together an assumption driven lifetime model are past. Modeling should be abandoned with the exception of situations where a model claim produces empirically evaluable claims which, supported by an evaluation protocol, can be reported to a formulary committee in a meaningful timeframe.

CONCLUSIONS

This critique of the established Tufts CEA database may come as a surprise. After 46 years since its inception in 1976, advocates of approximate modeled information might have thought it could stand as a critical part of assumption driven modeling and the technology assessment meme. Unfortunately, if the axioms of fundamental evidence were applied, it was doomed from the start in confusing ordinal and ratio scores, and the apparent significant prevalence of negative health state weights. Of course, the Tufts CEA is not alone; the belief in the ratio properties of ordinal scores and their application in assumption driven imaginary worlds is widespread. Any substantive criticism will invite pushback; after all, the TUFTS CEA has significant sunk costs and is widely accepted by analysts and those creating imaginary simulation driven value claims. As it stands, however, not only is the Tufts database, notably in respect of health state weights, an exercise that defies the axioms of fundamental measurement but a situation where no one thought the existence of states worse than death noteworthy.

REFERENCES

[i] Center for the Evaluation of Value and Risk in Health TMC. Cost-Effectiveness Analysis (CEA) Registry www.ghcearegistry.org https://cevr.tuftsmedicalcenter.org/databases/cea-registry

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

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

[iv] Langley P. Peter Rabbit is not a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(2): No 22

[v] Langley P, McKenna S. Measurement, modeling and QALYs. *F1000Research*. 2020; 9:1048 \https://doi.org/10.12688/f1000research.25039.1

[vi] Stevens S. On the theory of scales of measurement. *Science*. 1946;103: 677-80

[vii] Bond T, Fox C. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd Ed.) New York: Routledge, 2015

[viii] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes. New York; Oxford University Press, 2015

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

[x] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6

[xi] Galen Research Ltd. Manchester UK http://www.galen-research.com/

[xii] Tufts Medical Center, Center for the Evaluation of Value and Risk in Health. Global Health Cost-Effectiveness Analysis (GH CEA) Registry: Overview. 2018
http://ghcearegistry.com/ghcearegistry/2018_GH_CEA_Registry_Overview_Brochure.pdf

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

[xiv] 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