MAIMON WORKING PAPER NO. 4 APRIL 2021

FLOGGING A DEAD HORSE: A-MAPPING WE WILL GO

ABSTRACT

A foundation belief in technology assessment is that the quality adjusted life year (QALY) must be supported at all costs. The meme will collapse if the QALY is abandoned. Unfortunately, at least for believers, that denial is in place. The QALY is an impossible mathematical construct. Efforts to support the QALY through mapping from ordinal scores to more 'advanced' incarnations are not a solution. Attempts to map ordinal scores from the EQ-5D-3L to the EQ=-5D-5L are a last ditch attempt to rescue the QALY. What the erstwhile mapping fraternity fail to appreciate is that as the EQ-5D-3L fails to meet required measurement standards, the resultant 'mapped' EQ-5D-5L as an ordinal scale takes us nowhere. The downside is claims for cost-effectiveness which are meaningless based upon mapping claims which fail to recognize the limitations of fundamental measurement,

INTRODUCTION

One of the more absurd features of the technology assessment meme is the effort devoted to mapping from one preference or utility score to another. As these scores are ordinal it may perplex the observer to ask why try to map from one to the other when the focus should be on, the probably impossible, task of creating ratio measures for utility scales. Focusing on ordinals scales, and mistakenly thinking they have ratio properties is an endeavor fueled by a basic ignorance of the axioms of fundamental measurement. If you want to create a measure of quality adjusted life years (QALYs) then this must be the product of two ratio scales. It is mathematically impossible to create a QALY from an ordinal scale [1]. Unless a ratio utility or preference measure is created, the QALY is an impossibility. The result is dramatic: the edifice of cost-per-QALY health technology assessment collapses.

THE MULTIATTRIBUTE DEAD HORSE

Multiattribute utility instruments are an analytical dead end. It is not just a question of a dead horse but the more disturbing thought that the multiattribute instruments should never have been conceived of in the first place. Why attempt to add together scores on potential attributes defined by symptoms when the end result is a dimensionally heterogeneous measure that lacks not only unidimensionality but also construct validity. To add to this mélange, the scores can score utilities worse than death on a nominal 0 = death and 1 = perfect health scale. The various scales are ordinal.

This being the case it seems absurd given that the individual scores which are to be mapped are not only ordinal but have no possibility of being transformed to a ratio scale; without this the QALY collapses. What is missing is any appreciation of measurement standards in the physical sciences. The appreciation

that in the physical sciences the focus is on measuring single attributes, with either ratio or interval properties is entirely overlooked. Instead the focus in health technology assessment, in the creation of health related quality of life (HRQoL) scores, in the bundling of symptoms (which may not be single attributes) to capture a generic marker for therapy impact. Those undertaking mapping studies start on the back foot with symptom lists and levels that are then cobbled together to create ordinal scales that fail to allow assessment of therapy response other than though the analysis of rankings utilizing non-parametric statistics[2] . The absurdity is further enhanced when the impossibility of mapping between single attribute measures is recognized as nonsense. What are you going to map it to? Certainly, single attribute ratio measures can be combined, but not to the extent of assumption driven mapping functions.

PANDORA'S BOX

Idiomatically, the term Pandora's Box refers to ' any source of great or unexpected troubles ' and 'a present that seems valuable but is really a curse'. Both characterizations fit the proposal in the early 2000s to enhance the responsiveness of the EQ-5D-3L instrument to the EQ-5D-5L instrument. This involved modification of the wording to describe symptoms but more significantly the move from a 3-level response categorization to a 5-level. The result has been, not just confusing, but disastrous. It has laid bare the inherent precariousness of the two EuroQoL instruments and the futility of basing any comparative claims for cost-effectiveness on these multiattribute scores.

The fundamental error was in mapping from an existing multiattribute instrument that generated only ordinal scores. This doomed mapping from the start. Consider the often cited copula-based model estimates of Hernandez-Alva and Pudney (H-AP) [3]. The starting point is the EQ-5D-3L scores for individual respondents. An algorithm is then created to translate these to 'equivalent' imaginary EQ-5D-5L scores. This has the effect of building one algorithm on top of another. The original EQ-5D-3L algorithm is anchored at unity with rule base decrements creating utilities towards zero and beyond (the range is 1.0 to -0.58). The mistake here is a failure to recognize the ordinal character of the scale and the presence of negative values. The H-AP algorithm was built on top of this. This involved fitting the mapping algorithm to the data. What was overlooked is that if ordinal data are to be analyzed then the only avenue is through the application of nonparametric statistics. The four standard arithmetic operations are not allowed. Claims, therefor, for mean utility scores are nonsensical. A fact that the two authors overlook. They claim, for example, based on the EuroQoL (EQD) dataset that the average UK utility value for the EQ-5D-3L is 0.628 compared to the average of 0.703 for the EQ-5D-5L. As nonparametric statistics cannot support mean values, these numbers are meaningless because in an ordinal scale we have no idea of the distances between scores [4]. We might believe we do but that is only because the scores are put on an interval number line as a matter of convenience; any other number line could be used.

Without going into the details of the mapping algorithms (involving age, gender and two weights), the outcomes are of interest as they illustrate the fact that if you map from an ordinal score then you end up with another ordinal score. Of passing interest, and not surprising, is that when compared to the

Devlin range of utility scores (-0.285 to 1.0) the predicted EQ-5D-5L scores cover a range of -0.225 to 0.960. Illustrating quite clearly that the mapped scale remains ordinal; there is no hint of a concern that they should be focusing on a ratio scale to create QALYs. Depending on the mapping model used , the appearance of negative utilities is a common feature varying, in four model options, from -0.218 to -0.243. Perhaps we are in some parallel reality where ordinal scales have hidden ratio properties?

THE MAPPING PANACEA

For those who believe in the pre-eminent role of invented approximate imaginary evidence to support formulary decisions, the various mapping algorithms between direct and indirect preference instruments is just one more mysterious technique that can be applied; a mystery in the technology assessment meme. Noteworthy is the decision by the National Institute for Health and Care Excellence (NICE) to require its reference case modeling to include only EQ-5D-3L scores [5]. For NICE, where the EQ-5D-5L has been used, manufacturers are required to map from the 5D-5L-5L to the EQ-5D-3L for the simulation imaginary lifetime reference case model to be accepted and presumably from other multiattribute instrument scores. NICE, apparently, is blissfully unaware that the respective scales are ordinal and that the mapping makes no sense.

The lack of awareness in mapping from ordinal scales to produce by default more ordinal scales is shared by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), the premier advocate of the health technology assessment meme. In a 2017 task force report on mapping from non-preference based outcomes measures the question of the measurement properties (or, more likely, their absence) was not raised [6]. The focus of the Report was to consider how, if during a clinical trial, a patient reported outcome (PRO) was used, how these results for this instrument could be mapped to utility scores via an intermediary and then to create QALY estimates. Not surprisingly, it was unclear from the Report whether or not the intention was to map from a PRO to utilities defined by an ordinal scale (which meant that QALYs were impossible) or to recognize the axioms of fundamental measurement and attempt to map to a ratio measure, which would support QALYs. Unfortunately, the former appeared to be the case; the authors overlooked the fact that the majority of PROs are ordinal scores (e.g., Likert scales) and the instrument cannot be construed to have hidden ratio, let alone interval measurement properties.

The Report's discussion on appropriate cross-over or linked data sets to include in the development of mapping algorithms (typically to support regression analyses) makes no reference to the required measurement properties of these data sets. Consider the case where there are data from a PRO measure, an intermediate data set that captures elements common to the PRO and the target preference measure, and the preference measure. For such a link to be operational (e.g., in a regression model) the measurement properties of these data sets must be evaluated. In fact, they must all meet ratio measure properties. This is unlikely.

The issue with the ISPOR Report recommendations is that no account was taken of the limitation imposed on the creation of mapping algorithms on the required measurement properties of the final

product. If the intent was only to produce (by default) an ordinal score with negative utilities then this has been a waste of time. The properties of the instruments supporting the mapping must be such as to support parametric statistical analyses. This criterion is not mentioned. Nor is the application of measurement applications as a check on implicit modeling assumptions.

An odd feature of the mapping decision was that by the time mapping became imperative to save the QALY the metaphorical horse has already bolted; the multiattribute EQ-5D-3L (and its foundation in time trade off [TTO]) had already been recognized as a failed instrument; an analytical dead end. This involved both ongoing criticisms of assumptions, but recognition that the axioms of fundamental evidence had been neglected (or put to one side). It was already redundant; its multiattribute structure consigning it to irrelevance.

The effort put into trying to create mapping algorithms (and even ensuring they were embodied in standard statistics packages), absent an understanding of the axioms of fundamental measurement, is, in retrospect, a waste of time. It is impossible to resuscitate a dead horse. But the horse will be still attracting flies even lying in the knacker's yard. Without wishing to be unreasonable, it is surprising that with the number of experts involved in developing, over 3 years, the good practices recommendations that none apparently had an inkling of the limitations imposed by fundamental measurement. A lack of appreciation that is shared by the leading textbook in the field in its advocacy of assumption driven imaginary simulations, capped by equally redundant probabilistic sensitivity analysis [7].

CLAIMS FOR COST-EFFECTIVENESS

Unless the claim relates to a single attribute and associated costs, in other words claims which meet the standards of normal science, then any other claim for cost-effectiveness must be rejected out of hand. This standard, common in the physical sciences, is no part of the health technology assessment meme. Claims for cost-effectiveness must rest on the impossible, multiattribute QALY. As noted in a previous Maimon Evidence Report, health technology assessment is the only social science (if that is the correct term) that rests its laurels on the construction of claims from imaginary simulations that not only lack credibility but fail the requirements for empirical evaluation and replication [8].

This is not good news. It points to decades of ignorance of fundamental measurement with thousands of researchers pursuing a cost-effectiveness will o'the wisp modeling framework in mistaken claims for ordinal scores. The question that must be addressed is whether or not a utility or preference scores can be constructed with ratio measurement properties. This seems impossible given the present embrace of multiattribute ordinal utility scores.

QUO VADIS MAPPING

Are there any next steps for the current technology assessment meme? Certainly, analysts can continue to promote multiattribute scores and the impossible or I-QALY modeling; increasingly this is recognized as an analytical dead end. Surprisingly, the answer lies in the treatment of measurement in the physical

sciences. A recognition of the measurement of single attributes. Attempts to combine attributes in a composite score is a waste of time; unless an attribute has ratio properties they cannot be combined. If the purpose of mapping is to create utility scores with ordinal characteristics, then it is a waste of time. Add to this the nonsense that attaches to the creation of simulated model claims that fail the standards of normal science based in large part on the impossible or I-QALY. The best that can be done is to just dispense with mapping. It serves no useful purpose. Mapping endeavors are a waste of time.

CONCLUSIONS: GROUNDHOG DAY

After this intensive effort to create a mapping from the EQ-5D-3L to a hypothetical or imaginary EQ-5D-5L we appear to be back at the starting point. Together with the EQ-5D-3L with its properties of an ordinal utility or preference scale with states worse than death (negative utilities), ceiling and floor effects, a lack of dimensional homogeneity, an absence of unidimensionality and a lack of construct validity creating an impossible or I-QALY, we now have a number of algorithms mapping to an EQ-5D-5L scale with the same properties. This is hardly a major step forward as the mapping overlooks the axioms of fundamental measurement and creates one more ordinal scale instead of, presumably, a ratio scale. This presages a flurry of activities, in response to claims for lack of response for EuroQoL scales with more symptoms and response levels. Clearly a waste of time. The horse was dead before it left the stable. Unfortunately, news of its demise has not filtered down to the thousands of believers in the hidden ratio properties of the various EuroQoL scales and the mapping options that are open.

As detailed in a number of previous Maimon Evidence Reports and associated commentaries, the puzzling feature is the that none of these authors and ISPOR task force members appear to have recognized that both direct and indirect measures for preferences and utilities are ordinal scales. As they play a key role in creating I-QALYs and imaginary simulation models to support incremental cost-per-QALY claims, recognizing their ordinal property would collapse a centerpiece of health technology assessment. Mapping is merely indicative of this level of ignorance, not only of measurement theory but of the standards of normal science.

REFERENCES

[1] 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

[2] Conroy R. What hypotheses do 'nonparametric' two-group tests actually test? *Stata J*. 2012;12(2): 182-90

[3] Hernandez-Alava M, Pudney S. eq%dmap: A command for mapping between EQ-%D-3L and EQ-5D-5L. *Stata J.* 2028;18(2)(:395-415

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

[5] NICE. Position statement on the use of the EQ-5D-5L value set for England (updated October 2019). London: 2019

[6] Wailoo AJ, Hernandez-Alava M, Manca A, et al. Mapping to estimate health-state utility from non–preference-based outcome measures: an ISPOR Good Practices for Outcomes Research Task Force Report. *Value Health* . 2017; 20(1):18-27.

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

[8] Langley PC. The Herd and the Meme: Why Ordinal Scales have to be Ratio Measures in disguise. *Maimon Working Paper* No. 3 2021 www.maimonresearch.net