MAIMON WORKING PAPERS No. 3 MARCH 2024

WILL WE FINALLY ABANDON THE MATHEMATICALLY IMPOSSIBLE QALY WITH H.R. 485 PASSING IN THE SENATE?

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

OVERVIEW
H.R. 485 PASSING THE SENATE

The narrow success of H.R. 485 (211 to 208) in the House does not augur well for passage in the Senate. The argument presented here is that the supporters of this Act must build upon the disabilities or ableist argument, making a more complete and technically sophisticated case: it is not only the fact that the QALY is a mathematical impossibility but that it is supported by imaginary modeled simulations. This follows from the standards of normal science and fundamental measurement. The latter have been in place for patient reported outcomes for over 60 years with Rasch measurement; the former since the scientific revolution of the 17th century. Recognition of these makes clear that the QALY must be rejected in providing imaginary claims, to the detriment of patients in Federal and state health programs, together with modelled claims for pricing and access.

Abstract

Passing of H.R. 485 in the House of Representatives on February 7, 2024 represents a welcome and anticipated development in the abandonment of the multiattribute quality adjusted life year (QALY) for the allocation of health care resources in the US. Even so, the vote was close (211 to 208). This does not augur well for the likelihood of the Act passing in the Senate. While there is no case to be made for retaining the QALY, the case made in terms of disabilities can be strengthened by making the further case, a straightforward one, that the QALY, or more specifically the preference or utility score that creates the QALY, is only an ordinal metric. As such, it cannot support multiplication or any other arithmetical operation, so that the QALY is an impossible mathematical construct. In short, it is not the case that the QALY is ableist, but that the QALY has no role whatsoever to play in judging the extent to which those with disabilities are disadvantaged as it cannot capture response to therapy. This simplifies the case for abandoning the QALY selfevident; we have no need for extended debates over the eugenic implications of an impossible measure. But the case against the QALY, or more properly utilities is more fundamental: we must embrace not only fundamental measurement which yield true estimates of therapy response but the standards of normal science.

If, rightly, we abandon the QALY because it is mathematical nonsense, then this abandonment will embrace all proposed measures for therapy response and claims for products and devices for a place on Medicare and Medicaid formularies that do not support the recognized standards for normal science, empirical evaluation, and the standards for fundamental measurement which have been in place for over 60 years. Any value claim for a therapy intervention must, and there is no debate, utilize an instrument, and one which is patient centric, that is unidimensional, linear, interval and invariant with, ideally, a true zero.

The potential obstacles in the Senate can be overcome if those proposing build on the disability and related arguments presented in the House to demonstrate the need for a paradigm shift in health technology assessment in the US. We have to consider a new start. Fortunately, the framework for a new start has been already established. The purpose of this paper is to detail how this case and transition can be made and the new standards we are seeking for formulary evaluation and product acceptance.

INTRODUCTION

The purpose of H.R. 485 *Protecting Health Care for all Patients Act of 2023* is to amend Title XI of the *Social Security Act* to prohibit the use of quality-adjusted life years (QALYs) and similar measures in coverage and payment determinations under Federal health care programs. While no one would, hopefully, deny that human life has value and the allocation of healthcare resources must respect this even with budgetary constraints that may make allocative criteria inevitable, to focus on the QALY misses the impossible role of the QALY as a measure that should be taken seriously. As will be pointed out in this paper, the QALY is not a true measure; it fails the required standards for Rasch or fundamental measurement [1]. Prohibition in health care decision making of what is described as the QALY is long overdue. More to the point, the QALY should not have been introduced in the first place.

Irrespective of one's belief as to how health care resources are to be allocated, the QALY should play no part in that decision making. The reason is quite simple: from the perspective of sound scientific measurement the QALY is a mathematically impossible measure [2] . It is meaningless to rely on the QALY as a basis for denying access to health care and in the therapy pricing. While it may seem surprising, the fact that the QALY is a meaningless construct has been recognized for decades. There is a substantive literature which makes the case that we must not confuse measurement with observations [3]. For a construct such as the QALY to meet the required standards, we have to apply those that ensure that we are creating a measure that is unidimensional, linear, interval and invariant with a true zero. Unfortunately, without thinking this through, it is at this point that people switch off. The result is that the QALY continues to flourish. Opponents of the QALY, notably for those concerned with the allocation of health care resources to those with disabilities, falling back on taking the QALY on its own terms as a 'measure' without appreciating that they are in fact making sure that the QALY will to continue to be used when it is nothing more than a mathematical absurdity. This is the risk of a failing Senate vote.

The purpose of this note is to make, first, quite clear why the QALY is a charade, a mathematically impossible construct that lacks required measurement properties and, second, to point out that we have an alternative analytical framework to support value claims for products in target patient populations which meet not only the required standards of normal science but also fundamental measurement. It will become clear that the QALY fails on its own terms; there is no need to devote significant time and resource to prohibiting a presumed 'measure' which is self-defeating. All we need is to point out that health technology assessment must recognize, not imaginary and false claims, but the standards of normal science and fundamental measurement.

UNDERSTANDING MEASUREMENT

For the past century there has been agreement on measurement classification: the scales of measurement. There are four scales: nominal, ordinal, interval and ratio. Each provides a different type of information in assigning numbers. Specifically:

- Nominal: data or observations are sorted and categorized within a variable into mutually exclusive categories (e.g., zip codes, gender, employment status)
- Ordinal: data or observations are classified into categories within a variable that have a natural order but where the distance between the categories is uneven or unknown (e.g., responses to a question with five categories: never, rarely, sometimes, often, always)
- Interval: data or observations are sorted and categorized within a variable while measured on a numerical scale that has equal distances or intervals between adjacent values; there is no true zero, the scale can take negative values, where zero is an arbitrary point (e.g., Fahrenheit or Celsius temperature scales)
- Ratio: data or observations are sorted and categorized within a variable while measured on a numerical scale that has equal distances or intervals between adjacent values and a true zero where there is a total absence of the variable being measured (e.g. length, area, population)

The key point to note is that for interval and ratio scales the measure refers to a single attribute (it is unidimensional) with linear, interval and invariance properties. This cannot be assumed from observations. The Rasch mathematical model for transforming observations or ordinal data onto interval and ratio measurement means that rules have to be applied to ensure the measure has these properties . This has been known for over 60 years with software programs available to apply the Rasch criteria to create interval ratio measures for target patient populations. The Rasch model does not support composite or multiattribute observations. It is the failure to recognize this point and the need for measure to be created that spells disaster for the QALY.

UNDERSTANDING THE IMPOSSIBLE QALY

Few advocates of QALY removal seem to have an understanding of how the QALY is constructed and why that creates a mathematically impossible construct. The presumed impact on persons with disabilities is irrelevant as there can be no measurable impact in the first place that meets required

measurement standards. All that needs to be said is that instead of the *QALY and similar measures* all that is required is to prohibit any metric or score that fails the standards of Rasch or fundamental measurement.

There are two components to the false QALY construct: first, estimated time spent for a hypothetical target population in the simulated disease stages that map the assumed natural course of the disease to death and second, an assessment of the extent of which quality of life is discounted in that disease stage. These QALYs are then aggregated over disease stages. Consider a quality adjusted discounting for time T spend in a disease stage (say 12 months). For the discounting to occur such that the discounted time (the 'quality' time) a discount factor, call it 'k', has to be a proportion in the range 0 to 1. If the k value is 0.8, then the quality adjusted time is kT (k x T) equals 0.8 x 12 months = 9.6 months. A therapy intervention which increases the value of k yields more quality months. If k = 1 (perfect health) then the number of months (12) will be all quality months.

But the devil is in the details: how do we calculate values for k? This is where we part company with traditional ways of estimating k. Following the Rasch rules for measurement k should have the properties of representing a single attribute (e.g., needs fulfillment – a patient perspective) with linear, interval and invariant properties but with a true zero. In other words, a ratio measure. It is thus perfectly possible to create a 'true' QALY but with two provisos: (i) the values reflect those of patients in a particular disease state and (ii) the instrument to capture items to form the instrument must be an expression of a latent variable. This second point is not just abstract: if we think quality of life is important and if the needs of patients represent an expression of that quality of life, then we might construct items for an instrument, from patient interviews, asking what needs (which will be health related) are important. The instrument would then comprise items which can be in binary (yes/no) form and the respondent is asked, before a new therapy, which needs are met and then after therapy to see if a greater number of needs are met. We can use either number of responses or transform to probabilities: this is our 'k' value.

A 'k' measure (range 0 – 1) following Rasch rules for instrument development where, in probabilistic terms, we rank items of increasing difficulty and match against the ability of respondents to respond positively to an item is straightforward and such instruments have had hundreds of applications. But the authors and supporters of the existing QALY equivalent to the k preference will have nothing to do with Rasch. This is where the disaster unfolds. There are three errors: (i) there is no concept of the requirement that, as in the physical sciences, a measure must represent a single attribute (e.g., length) with Rasch properties; (ii) the QALY utility or preference value is composite capturing health states defined in the various multiattribute instruments by symptoms and dimensions with ordinal responses; and (iii) in defining a utility or preference from instrument responses as utility decrements from perfect health (k = 1) the resulting aggregate scores defined by preference or utility algorithms inevitably undershoot or overshoot zero. The Rasch framework ensures that the measure cannot take negative values.

It is important to note also that there is not a globally agreed QALY construct. Obviously, time is a measure which meets ratio criteria, but the time estimate typically used in the calculation of costs-per-QALY is an artifact of the underlying assumption driven simulation model. The utility

or preference score that is also part of the model comes in a number of varieties depending on the health state dimensions and symptoms that form the health state description which is valued by a community sample and not by patients in a disease state. For example, the most widely used ordinal preference score the EQ-5D-3L is based on weights attached to five health dimensions (mobility. self-care, usual activity, pain/discomfort, anxiety/depression) with 3 response levels (no problems, some problems, extreme problems). This gives 243 possible health state descriptions (bundles of health dimensions and symptom levels). Other preference scores may encompass more dimensions (the EQ-5D-5L has five) while others may have more symptom levels with unique descriptions of each level. The Rasch model does not rely on health state descriptions.

Where multiattribute preference or utility scores are created, the QALY is meaningless as it is impossible to multiply T by a composite reference score +/-k $\leq$ 1 to give a QALY with positive and negative values. For those concerned with target disability patient populations the solution is clear: abandon the present menageries of multiattribute generic ordinal preference and utility scores in favor a of single attribute Rasch standard measure for specific populations. This is the only framework which will give a meaningful account of the impact of therapy interventions where 'k' measures the results for a single attribute instrument that represents a manifestation of quality of life. In a clinical trial 'k' might be a primary endpoint; it seems redundant to expand this to a time-based equivalent.

SIMULATION FOR IMAGINARY DISABILITY CLAIMS

For those making the case to abandon the QALY the misapplication of traditional multiattribute composite ordinal QALYs extends to the creation of assumption driven simulated modeled imaginary claims which form the core of the cost-QALY modelling [4]. The best example in the US is the reference case assumption driven simulated modelling applied by the Institute for Clinical and Economic Review (ICER). Thes models generate claims that are entirely imaginary. The rejection of such models should receive equal billing for those attempting to convince lawmakers (the Senate) to abandon the QALY. Claims for incremental costs-per-QALY are false, not only because of the use of the QALY but because the simulation models fail the standards of normal science; they provide no empirically evaluable claims and were never designed to do so over the lifetime horizon for modelling. Recommendations from such models for pricing and access must be rejected.

In short, if the supporters of the disability case want to abandon multiattribute QALYs they should also advocate abandoning assumption driven simulations that fail to provide empirically valuable product claims. While the case for abandoning assumption driven simulations might be a bridge too far for many in the disability case, there must be a cleansing of the stables. For over 30 years access and pricing decisions have been driven by imaginary claims for cost-effectiveness. Allied with multiattribute community preference scores, target patient groups, including those with disabilities, have been doubly disadvantaged. It is not just the disadvantage implicit in the potentially biased false QALY values but the choice of model structure and dubious assumptions that populate these models.

QALY STATES WORSE THAN DEATH

As decrements from a baseline utility of preference score of unity (= 1) is the basis for calculating preference or utility scores considered, with an algorithm creating scores from responses by patients with predefined clinical states and symptoms. Each symptom has a weight and these are combined by the algorithm defining only allowed clinical health states. As such, it is inevitable that just as much the utility or preference algorithm is tweaked, the worst health states will either being greater than or less than zero. An exact correspondence of states worse than death, judged by community preference of the symptoms reported, and the death of zero will be just happenstance. In fact, in all cases there are negative scores indicating a state worse than death. This follows, inevitably, from preferences being determined as utility decrements from perfect health (= 1). It is worth noting that while positive utilities have an upper limit of unity, negative utilities have no such constraint; they can extend to minus infinity. For example, the value sets for health states with the EQ-5D-5L for the US yielded a preference or utility range from -0.573 to 1 with 20% of the predicted health states having a negative value; death is an arbitrary point on an ordinal number scale [5].

Allowing negative values or preference scores that create negative QALYs, is an insidious way of allocating resources to make sure patients are excluded to the benefit of those with positive QALY scores; which is a modern twist on eugenics (and equally false). The fatal flaw is the reliance on community preferences for clinically determined health states rather than on the interests and needs of the patient. Claims can then be made that to exclude those with negative QALYs will increase overall population health; time spent in a health state with negative QALYs reduces overall population health; it is a loss to the community. It is a system that prioritizes the needs of nondisabled people, viewing non-disabled people as more valuable, where economic evaluations lead to a systematic underestimation of the value of life extending treatments [6] .

However, it is not just people with disabilities that are negatively impact, similar arguments can be made for inflammatory bowel disease [7] . These application of negative scores to create negative numbers and QALYs cannot be avoided.

FUNDAMENTAL MEASUREMENT AND DISABILITIES

Rasch modelling is ideally suited to assessing the impact of therapy interventions for disability target groups. Developed in the 1950s for assessing ability on school mathematics tests, it is globally accepted as the only valid basis for instrument development. The essence of the Rasch approach, which has a particular resonance for those with disabilities, is that any instrument design to capture response (e.g., for questions on quality of life) must be for a single attribute or unidimensional:

> *A person having a greater ability than another person should have the greater probability of solving an item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one* [1].

Developing a Rasch instrument is quite straightforward with the required model standards in place for over 60 years. The object is to fit the data elements (item descriptions) to the model to give the best fit which reflects the view of patients in target groups (e.g., specific disabilities). In health technology assessment, all Rasch instruments (with 30 developed and hundreds of language versions) are specific to a disease state. They are inputs to decision making as they report on response to therapy for competing interventions and assess standards for patient populations over time. There can be no discrimination against groups because community preferences for health states are excluded. The focus is on the status and needs of patients in a disease state by stage of disease. This is the true measure for therapy response, apart from purely clinical measures.

This argument applies across the board. Whenever a QALY or incremental; cost-per-QALY claim is made and set against a cost-per-QALY arbitrary threshold to support claims for pricing and access, the response is simply that the entire exercise makes no sense. This goes back to the basis for QALY estimates in specific therapeutic areas and target patient population: the creation of simulated assumption driven imaginary claims.

Rasch instrument development, in focusing on a specific disease state and patient needs or concerns, does not admit of implicit or explicit discrimination against these respondents. It is an entirely patient centric instrumentation. Needs are identified and ranked by difficulty with respondents ranked by ability to meet these needs given a particular therapeutic intervention. The Rasch instrumentation matches the abilities to meet needs of a sample of the target population set against a distribution ranked of needs by the difficulty of those needs. The positive contribution of a therapy is judged, therefore, on the increase in needs that are met for the target patient population. This gives the required fit to assess for any defined disability group (with its own Rasch instrument) the extent to which respondents believe increasingly more difficult needs are being met.

OTHER MEASURES

One of the unresolved issues with H.R. 485 is the meaning of the phrase ‘other measures’. While these are meant to be similar to the QALY no examples are given for these other instruments; the decision maker has to judge if a claim (but not for QALYs) passes muster. If we apply the standard proposed here of a measure which must respect Rasch or fundamental measurement standards, then H.R. 485 has opened Pandora’s Box for patient reported outcome instrument targeted to specific patient populations. Meeting Rasch standards means instruments must be unidimensional, linear, interval and invariant with a true zero (or acceptable approximation). In these terms there are only a handful of disease specific outcome measures which capture quality of life (the needs fulfillment instruments) with the rest, claiming to measure quality of life, failing.

The reasons for this failure have been recognized for decades. If the disease specific instrument is based on responses to questions or items (e.g., a Likert scale with 5 response levels) than the only basis for adding up these integer responses to create an aggregate score requires *a priori* assumptions that (i) all items in the instrument are of equal difficulty and (ii) the thresholds between steps are of equal distance or value. If not, the aggregate score or ‘measure’ is meaningless; it can’t capture response to therapy. This does not mean that instruments with a range

of item response levels are impossible. There are techniques in Rasch or fundamental measurement to create the required scoring standard from such responses: the Rasch Rating Scale Model and the Rasch Partial Credit Model [1] . The problem is that few have been aware of these models (software has been accessible for over 40 years) so that there is a sizable reconfiguration (and reeducation) task that lies ahead.

AFTER THE QALY: THE WYOMING CERTIFICATE PROGRAM

Once modeling such as that exercised by ICER, including the QALY, have been abandoned the question is the choice of a framework to support submissions to the Federal (Medicare) and state government (Medicaid), together with groups such as Veterans Affairs to support negotiations, market entry and pricing of products and devices. It is of no use to abandon the QALY and just leave a vacuum. Those who have been at center stage in making the case for abandoning the QALY should make the case for an evaluation framework that meets the required standards defined in terms of normal science and fundamental measurement. There is a way forward with the transition to a new standard for product and device evaluation resting on just 3 premises:

- All value claims must refer to single attributes for defined patient populations that meet the demarcation standards for normal science: they must be credible, evaluable and replicable
- All value claims, notably for patient or caregiver reported outcomes. must be consistent with the limitations imposed by the standards of fundamental measurement: they must be unidimensional with linear, interval and invariance properties
- All value claims must be supported by an agreed protocol detailing how they are to be assessed in a meaningful timeframe

These are premises that have been proposed in a recent submission to the current Health Technology Assessment Policy and Methods Review by the Australian Department of Health and Aged Care: *Rejecting the PBAC Guidelines for Imaginary Cost-Effectiveness Value Claims: A proposed New Start for Health Technology Assessment in Australia.* A copy of the submission is attached [8].

These proposed requirements and their implementation for formulary submissions are detailed in a recently released University of Wyoming on-line Certificate Program: *A New Start in Health Technology Assessment.* This is a 14-module program comprising audio-visual presentations, notes to each module (in total 85,000 words) and short multiple-choice exams for each module. The program is recognized by the US Accreditation Council for Pharmacy Education (ACPE) with credit of 20.5 hours. A Certificate is given by the University of Wyoming to all successfully completing the program. For further information the link to the program is detailed in the references [9].

CONCLUSIONS

It can be argued that the 30 or more years over which the impossible QALY and assumption driven simulated model claims have dominated health technology assessment represents a wasted effort

with substantial costs to patient populations in denial of care. Abandoning the QALY and similar measures is a positive and overdue step in the application of new standards in health technology assessment. If value claims for competing products are to reflect the needs of target patient groups the potential success of H.R. 385 in the Senate and final Presidential approval will not only alleviate the concerns of those seeking more ethically defensible health system decisions but ensure defensible metrics for access to therapy and meaningful claims for therapy response. While concern for patients with disabilities has been a key starting point, the implications are more substantive than was probably first envisaged.

Those who wish to face down the QALY will face an entrenched opposition of believers who reject, or lack any awareness of, the standards of normal science and fundamental measurement. A level of ignorance, for want of a better word, that is unacceptable. In this lack of awareness, we must include advisors, advocacy and media groups who also lack any appreciation of these requirements; ICER recommendations for pricing and access are widely reported, particularly by media groups who have no understanding of the flawed methodology that support them.

We must not drop the ball. Certainly, there is a grim humor in noting the time and effort involved in arguing against a QALY measure that is nothing more than a chimera. It must be made clear that we have to avoid impossible mathematical constructs both in terms of measurement and modelling. The passing of H.R. 485 in the House marks a key step forward but one that is on a knife edge. The lack of bipartisanship is not a good auger for the future Senate vote. This is unfortunate, but reflects in large part a lack of understanding of the impossible concept of the QALY and of the false modelling standards in health technology assessment.

If the ball is dropped this is our own fault.

REFERENCES

[1] Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (4th Ed.) New York: Routledge, 2021

[2] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved] *F1000Research* 2020, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[3] Wright B, Linacre J. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil.* 1989; 70(12):857-60
https://www.researchgate.net/publication/20338407_Observations_are_always_ordinal_measurements_however_must_be_interval/link/5563b02408ae9963a11ef326/download

[4] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: 2 approved]. *F1000Research* 2022, 11:248 (https://doi.org/10.12688/f1000research.109389.1)

[5] Pickard A, Law E, Jiang R et al. United States Valyation of EQ-54D-5L Health States using an International Protocol. *ValueHealth.* 2019;22(8):931-41

[6] Schneider P. The QALY is ableist: on the unethical implications of health states worse than death. *Qual Life Res*. 2022;31:1545:1552

[7] Sawhney TG, Dobes A, O'Charoen S. QALYs: The Math Doesn't Work. *JHEOR*. 2023;10(2):10-13. doi:10.36469/001c.83387

[8] Langley P. *Rejecting the PBAC Guidelines for Imaginary Cost-Effectiveness Value Claims: A Proposed New Start for Health Technology Assessment in Australia.* A submission to the Health Technology Assessment Policy and Methods Review by the Australian Department of Health and Aged Care.
February 2024
https://docs.google.com/document/d/1pGbItxzYIrtIf5HiLSYnoxMwHBIhaCoD_u4OxhoiwY/edit?usp=sharing

[9] University of Wyoming. School of Pharmacy. Certificate Program: *A New Start in Health Technology Assessment.* https://www.uwyo.edu/pharmacy/resources/certificate-program-a-new-start-in-health-technologyassessment.html