MAIMON WORKING PAPERS No. 9 JULY 2021

ABANDONING EUGENICS AND THE QALY

Abstract

Should decision making in health care, notably in respect of the allocation of resources between individuals and disease states, rest on notions of the burden of disease and denial of care as assessed by societal evaluations or on the extent to which the need of patients are caregivers is fulfilled. The prospect of the denial of health care, for those deemed 'unworthy' has a long history in the eugenics movement. Many have assumed that this 'utilitarian aberration' has long been discredited. Unfortunately, once the question of the allocation of limited health care resources is considered it reasserts itself; manifested in the creation of health state preferences and states worse than death, and application of the cost-per-QALY calculus driving claims for pricing and access. In the US, this focus on cost-per-QALY claims is most closely associated with the Institute for Clinical and Economic Review (ICER) with is regular clinical assessments and modelled imaginary simulations supporting recommendations which, in many if not most cases, give support to the denial of care. The purpose of this commentary is to point to the unfortunate similarities between 'eugenic' decision making and the application of thresholds in burden of illness cost-per-QALY exercises. If we are to finally rid ourselves of a 'eugenic' approach to health care resource allocation, then we must abandon preferences and the QALY calculus.

Introduction

In health technology assessment there is an important dichotomy between the notion of a societal evaluation of the burden of disease in the assessment of competing therapy interventions and the role of health care interventions to meet the needs of patients, including caregivers and family members. The Institute for Clinical and Economic Review (ICER) in the US has taken upon itself the role of arbiter of drug pricing and access to therapy based on ICER's assessment of the burden of disease expressed in multiattribute preference scales and quality adjusted life years (QALYs). Recommendations for access to health care, including recommendations for pricing, are based on assumption driven modeled simulations which have been shown to fail the standards of normal science; they are pseudoscience. Yet ICER, in its self-appointed position as arbiter for cost-effectiveness claims in the US, sees itself in pole position to endorse as well as deny access to health care. Applying multiattribute ordinal preferences that reflect societal valuations of health states, ICER produces clinical evaluations and model driven evidence reports that are intended to inform decision makers. An adverse ICER report can have devastating implications for patients, care givers and family, all too often in rare diseases.

As allocation of health care resources must involve denial as well as support, the ICER burden of disease metric has much in common with societal support for eugenic solutions in access to care; allocating healthcare resources on perceptions of the 'worth' of the individual and denying support for those considered 'unworthy' [1].By the early decades of the 20th century 'a process had taken hold in which descriptive assessments based upon physical difference, clinical aesthetics and social values had combined to create a progressive measure of 'worthy versus unworthy life' [1]. While it would be possibly a step to far to describe societal multiattribute health state preference scores as a latter day Wannsee Protocol, the application of generic clinical status markers to define health states Given the possibility that on the burden of illness calculus, the costs of care or the 'worth' of a health state may exceed what society is willing to pay, it is possibly not too much of a stretch.

In the last few decades the earlier version of the eugenic ‘worthy’ life have been replaced by the notion of life’s quality, based on scaled physical and cognitive attributes, but (again the eugenics tradition) valued from a societal perspective. The key to this is the generic perspective to support assessments across disease states. There is no attempt to define disease specific measures of the burden of disease but rather as in the ICER case, to apply a minimal clinical scorecard to force the denial or sanction of access to health care resources. Denial is straightforward: directly through prior authorization and prohibitive personal costs; indirectly through reducing the anticipated returns on investment in new products through recommended price discounting so that potential palliative interventions and even cures are abandoned.

The purpose of this commentary is to point out that, as in the case of measures of ‘worth’ in eugenics, which are considered pseudoscience, ICER’s ongoing attempt to provide a common metric for evaluating the burden of disease, to support denial of care in resource allocation, is also pseudoscience. Accepting the ICER metric lets eugenics in by the back door.

Formulating Societal Preferences for Access to Care

Whether expressed in value or utility terms, societal preferences are central to generic multiattribute instruments such as the EQ-5D-3L/5L, the HUIMk2/3 and the SF-6D. Constructing these instruments requires, first, agreement (typically by clinicians) on the collection of symptoms that measure perfect health. These are necessarily limited to minimize respondent fatigue. As such they can hardly be thought of as a measure of perfect health (which no one in clinical practice would agree with). The second step is to ask a sample of the population to value health states defined by combinations of response levels for these symptoms. The EQ-5D-3L, for example has five symptoms (mobility, self-care, usual activity, pain/discomfort, anxiety/depression) with 3 response levels for each symptom (no problem, some problems, major problems). This yields 243 individual health states with a small number of these evaluated to give a benchmark valuation, which is then modeled to give preference scores for each of the response levels within each symptom (15 scores). These preference scores (sometimes referred to as TTO tariffs) are then combined in an algorithm to yield a single score, hopefully in a range 0 = death to 1= perfect health.

Importantly, while the scoring algorithm is capped at 1 with preference weighting scores applied as decrements, all multivariate instruments generate negative scores or what are described as ‘states worse than death’ (the eugenics overtones presumably unintended). For any target patient group, therefore, the possibility exists that there will be a ranking of individual response scores as a distribution that can give negative values. Some respondents are ‘alive’ with the scores in the positive range to unity while others (who are presumably not dead as a response would require a Lazarus-like transformation) occupying an odd societal ‘limbo’ state that is worse than death; or where death is the preferred societal state.

If society, or the dominant political party, decides that care must be denied (or ‘rationed’) to those ‘less worthy’ or, in the modern vernacular, those with a ‘poor’ quality of life with no hope of remission or improvement, then the filter scorecard must meet minimum measurement standards. One approach is to consider a scorecard construct that, applying a utilitarian ‘denial’ calculus can identify patients who are experience highly adverse health states, objectively ‘states worse than death’. This has been achieved, although more by accident than design, with both direct and indirect preference constructs. Starting out in the 1980s they embraced quality of life or, more narrowly, clinician determined criteria for health states. The presumption (or at least the prayer) was that all health states would fall conveniently in this range; none would eventuate that were ‘worse than death’. They also assumed that the resulting scale would have ratio properties, with interval scores, and able to support the ’gold standard’ or Holy Grail of health measures to support the allocation of health resources, the quality adjusted life year (QALY). In the event, this wishful thinking never eventuated for a simple reason: if you want to have a measure with specific properties, then the developer has to build these into the measure from the ‘get-go’. The level of ignorance of measurement theory was profound: not only did they fail on the ratio property front, but they also failed on the self-proclaimed ‘multiattribute’ front. The axioms of fundamental measurement deny the possibility of adding together scores on different health attributes; if you try then you end up with a dimensionally heterogeneous scale that lacks construct validity.

States Worse than Death

The presence of negative scores for both direct and indirect generic preference instruments was recognized at early stages in developing the various preference instruments. Yet the developers persevered, paying scant attention to the adage ‘if you are in a hole stop digging’. The result was a series of preference scales that yielded negative scores. Perhaps the best ‘own goal’ was the EQ-5D-5L sensitivity revision. At the end of the 1990s it was argued that the EQ-5D-3L was too coarse; it lacked sensitivity as a generic measure. The solution was, not to increase the number of health symptoms or health dimensions from the existing five, but to increase the number of response levels from 3 to 5 ordinal ranks. Pandora’s Box was opened when the EQ-5D-5L was released in 2009. Chaos ensued because going from a possible 243 health states (3^5) to 3125 (5^5) increased the number and proportion of states worse than death (the issue of the ordinal nature of both scales was never raised). Doubts about the EQ-5D-5L persist with the National Institute for Health and Care Excellence (NICE) in the UK forbidding (in 2019) its use in modeling. The upshot, with an attempted US valuation in 2019, is that for the EQ-5D-3L the 243 health states, 10 (4.1%) had negative scores while the EQ-5D-5L the 3125 health states had 624 (20.0%) worse than death scores; the respective value range was -0.109 to 1.0 and -0.573 to 1.0. It is possible to envisage a eugenics equivalent distribution of scores which, to acknowledge the ‘founder’ of eugenics (Sir Francis Galton: 1822-1911) that might be called the Galton [or GALT-EU] Scale for evaluating a ‘worthy’ health state [2].

If a scale is to have a true zero then there must not be any circumstance in which a negative value can emerge. The fact that multiattribute instruments such as the EQ-5D-3L yield negative scores is unsurprising. If, as is the case in the physical sciences, you want to design an instrument with ratio properties, that is a scale with a true zero, then that has to be built into the design of the instrument [3]. This was not done. The reason is unclear, but certainly none of the developers (or the current owners of the EQ-5D franchise) took this into account. The result is a degree of embarrassment, with the hope that inquiries can be put to one side as the focus is on the impossible ‘average’ of the individual ordinal scores and not their distribution.

We can add to this debacle the fact that as multiattribute instruments, the EQ-5D-3L and others fail the axioms of fundamental measurement as they are dimensionally heterogeneous, failing the test for unidimensionality, and construct validity. As such they just aggregate ordinal scores over five symptoms, creating an overall ordinal score. This means they can only support nonparametric statistical operations; they cannot support standard arithmetic operations. To support a QALY, where time is a ratio scale, we require the preference score to have also ratio scale properties to support multiplication. Ordinal scores cannot support multiplication because we can only rank the scores; we have no information on the distance between them. Hence the impossible QALY [4] .

Transparency

ICER hides behind ordinal multiattribute preference scores sourced from the available quality of life literature. ICER's coterie of academic model builders show not the slightest interest in going behind these scores to assess their measurement properties; there only justification is support for one more assumption to plug into the model [5]. As ordinal scores, we can report only on the median value with both negative and positive values, but not the 'average' score. Not surprisingly, this rather trivial point is overlooked by ICER and other model builders, where they report average QALYs and average utilities. But there is a serious downside: if there are negative preferences these will 'depress' both median values and the ersatz 'averages' (which are accepted at face value rather than as a distribution which may give a picture of states worse than death). If a rare disease state exhibits a significant number of negative 'states worse than death' the utility increments will be less and the cost-per-QALY figure will be greater. This may put the ICER mandated discount in a range that makes further investment in the product unattractive. Unfortunately, few who 'review' ICER models recognize that in many cases the 'average' preference score combines negative and positive values where some respondents are more 'worthy' than others to receive health care.

Although ICER would have to have access to the underlying data sets where respondents report their individual preference scores, it would be salutary for formulary committees to ask what the distribution of preference scores was for the 'average' preference score (e.g., by stage of disease). It is presumably one thing to have only positive ordinal preference scores versus another where, say, 20% of respondents report a state worse than death. After all, if ICER is focused on the allocation of health care resource to those who, in some sense, are 'worthy of attention' then a case where, say, 20% of respondents report a state worse than death, is a rather odd position to be in; if these respondents could be 'taken out', then the calculus may change significantly.

ICER is not alone in promoting and making claims from imaginary modeled simulations. A key player is the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Founded in the late 1990s, is now with a global reach promoting inventing approximate information rather than hypothesis testing; an implicit support for eugenics and burden of disease. The word 'approximate' is a fudge; there is no such animal as 'approximate information' when the basis for claims are assumption driven imaginary simulations. One product of the commitment to imaginary simulations, promoted assiduously by ICER, is probabilistic sensitivity analysis (PSA) [6]. This is a framework to produce imaginary, assumption driven non-evaluable claims for the cost-effectiveness of products but expressed in terms of allowing uncertainty to enter the calculus. PSA allows an analyst to create blanket claim as to the likelihood that a product will be cost effective versus a comparator (QALYs are a key input) at different prices. In principle, this allows a health system to establish a cost-effectiveness threshold and deny access to care for products that are below that threshold. While, to many, PSA is seen as a gold standard for denying access to care on cost-effectiveness grounds, its intrinsic appeal should not hide the fact that it fails the demarcation test between science and pseudoscience; it creates non-evaluable claims driven by assumption. Any number of competing PSA's could be envisaged for any product comparison.

Interestingly (or not) the Tufts University Medical Center supports databases of preference scores and cost-utility claims based on the health technology assessment literature: the Cost-Effectiveness Analysis (CEA) Registry [7]. While the organizers of these databases, which can be accessed by a small fee, seem singularly unaware that they are promoting ordinal scores not ratio measures (the eugenics portfolio), they offer the erstwhile model builder the opportunity to choose from a variety of preference scores to validate constructing imaginary claims, including PSA, to create invented comparative claims. For those who are attracted to non-evaluable modeled ordinal preference scores from a variety of multiattribute and similar instruments as inputs to modeled simulations, there are some 9,080 cost-utility papers that have been summarized and cited between 1979 and 2019. Given preferences (utilities and values) are ordinal scores for populating imaginary simulations, this treasure trove of ordinal scores from a variety of preference instruments and stages of disease seems a monumental waste of time (but you do get $25 for each paper you summarize and submit; enjoy, it's probably higher than the minimum wage).

The Appeal of the QALY

It is surprising, but few critics of the QALY appear to fully understand the role and construction of value or utility preference scores. This limited understanding extends to the lack of appreciation of the embrace by ICER of pseudoscience in the application of cost per QALY simulation models [8]. These fail not only the standards of normal science but also the elementary logical point that inductive inferences cannot support future claims (Hume's problem of induction first proposed in 1748; (David Hume 1711-1776) [9]. On this argument alone, the ICER analytical framework collapses.

Yet, supported by the leaders in health technology assessment, the role of simulation models to create approximate imaginary information has been accepted with an explicit rejection of hypothesis testing [10]. Imaginary information modeling is the soft or easy option; why wait for more conclusive evidence of the impact of a new therapy, when you can invent claims? There was a receptive audience. The concept of approximate invented evidence, it could be argued, took sufficient hold of the attention of those in health technology assessment, with enough force that they remembered it to transmit to others. For this 'meme transmission fidelity' the concept of creating imaginary information was sufficiently novel but not so outrageous that the prospective audience would immediately deem it ridiculous [11]. Unfortunately, if there had been sufficient appreciation of the standards of normal science, as evidenced in the acceptance of the role of hypothesis testing drug in development and the limitations imposed by the axioms of fundamental measurement, this ridiculous analytical dead of cost-per-QALY value claims could not have persisted for over 30 years.

Against this appeal to the axioms of fundamental measurement, the QALY retains its position in the lexicon of technology assessment. In current political terms it is the equivalent of the 'big lie' or the 'big steal'. Sufficiently large number of health technology assessment practitioners maintain their allegiance to a mathematically impossible construct, together with cost-per-QALY simulation modelling that fails the standards of normal science. ICER is not alone: after all, if everyone else does it as they claim, this is presumably sufficient reason for ICER to persist with its business model to invent non-evaluable evidence. If not, the business model collapses.

Abandoning the QALY

It is not a question of waiting until a successor measure of a value claim or ICER Holy Grail is accepted; we do not have the luxury of recognizing its limitations, yet continue to accept a role for the preference score and the resulting imaginary QALY in drug pricing and access when it is seen as the basis for the denial of health care. It has no role. It has to be rejected. Surprisingly, rejecting the QALY is easy because there is no need for a replacement. Our focus should be, not on multiattribute ordinal scales to support an overarching value or 'worth' claims across respondents in various disease states, but on single attributes, determined by formulary committees, to support decision making; these may be clinical, quality of life or in terms of drug and resource utilization. The primary feature, and this is detailed in version 3 of the Minnesota formulary guidelines, is the development of protocols to support claims made to formulary committees [12]. These may support meeting evidence gaps or considering need fulfilment quality of life, but all have the required measurement characteristics. Rather than basing formulary decisions on assumption driven imaginary claims, real world evidence comes to the fore to support pricing and access negotiations between formulary committees, agents for the patient and caregiver, and physicians. We can put QALY evaluations, with their eugenic implications, to one side in favor of ongoing evaluation and replication of credible single attribute claims for target patient populations.

Patients and caregivers are the ultimate beneficiaries of health care interventions. To both patients and caregivers their quality of life rests in large part on the ability of those interventions to help them meet their need. Defining this need requires, for target patient populations in disease areas, an assessment through subjective reviews of patient responses, their need. There is one avenue open to creating a measure that meets the standards of normal science and fundamental measurement. This is the recently developed disease specific measure, called the Need or N-QOL scale, a disease or target patient population specific bounded ratio scale from 0 = no needs are met to 1 = all needs are met [13].

The N-QoL measure stands in contrast to societal preference measures: it is a measure of the need of patients (and caregivers) not of the burden of disease. This is the critical distinction; there are no eugenics implications. This is not a central planning exercise to determine who are best eligible for health care and who may be denied care on a generic societal preference calculus with clear links to the eugenics paradigm. Rather, the focus is on the extent to which the individual (and their caregiver) may be expected to benefit defined, not in clinical terms, but in terms of their need. This is a single attribute and is factored into decision negotiations. It is not a process of reassigning and denying access to care; nor is it a basis for price discounting. This is not a preference scale, so should not be compared to value or utility scales. It is a measure from the patent perspective in a target disease state of the extent to which need as defined by the patient is met. The N-QOL captures a single latent attribute: patient or caregiver need. It avoids the multiattribute approach by avoiding clinical criteria, which may be of more interest to the clinician than to the patient or caregiver. It is patient-centric within disease states, focusing on the patient as the ultimate beneficiary of therapy interventions. This is not new; some 30 instruments have been developed over the last 25 years, covering the major disease states.

Slamming the Door Shut

ICER's failure rests on a number of factors; the most important being enthusiastically accepting the health technology assessment paradigm that eschews hypothesis testing and the discovery of new facts in favor of the so-called approximate information or invented evidence paradigm. ICER is not alone; approximate information has been the accepted framework for 30 years. It is a soft option where non-evaluable model claims support formulary decisions; it is quick with a fast turnaround in consulting income. Far better than setting out an evidence program to fill gaps and discover new, even if provisional, facts.

The problem with modeling approximate information and using invented non-evaluable claims to support formulary decisions in the early stages of product market entry is that with an adverse decision to limit access and impose pricing caps, is that the model is never revisited; there is no incentive to do so; least of all by health care systems that may have egg on their face from an earlier decision to deny access to needed care. The door, for example in a rare disease, is slammed shut on future therapy investigations, with companies reluctant to invest [14]. The recipients are deemed less 'worthy' with implications stretching decades into the future.

Conclusions

The pursuit of prospective utilitarian measures to drive and rationalize healthcare resource allocation decisions ignores the need and life quality of individuals; members of target patient groups are just cyphers. They are scores on an ordinal scale which, by construct, forbids any discussion or assessment of the difference between health scores or any application other than non-parametric statistics. Treating these preferences as ratio scales in disguise makes no sense. This invalidates much of the analysis which accompanies their application including, as ICER would argue, a defensible position from which to 'objectively' argue for the allocation and denial of health care and enter the Tufts data base. Hopefully, ICER believes, this takes us away from any moral obligation to defend our model decisions even though they fail completely the standards of normal science. ICER's approach is nothing more than pseudoscience; a characterization that applies equally well to eugenics. At least they have that much in common; 'the old eugenic assumption, transformed into a post-war clinical argument is now returned as social science fact' [1]. Why do we always assume that any assessment of quality of life allocative mechanism always starts with the notion of 'disease burden?

Where does this leave us? Arguably in a better place which may focus on the need fulfillment of patients and caregivers defined by their health experience as members of a target patient group. We have to move beyond the belief that the importance about a person is not disability and the ability to increase societal 'worth' by addressing that disability but a more embracing notion of their need and whether that need is met. Therapies should not be judged on their 'worth' in possibly returning patients to an acceptable life quality non-eugenic adverse status. The question is: what is and who judges (not the Schutzstaffel) the acceptability (and reporting by the patient and caregiver) of a positive life quality and the likelihood of its achievement.

Contemplating the implications of measuring the burden of disease can lead us to some dark places. How do we value, from societies' perspective, the impact of competing therapies on the constructed 'worth' of the individual? Is it reasonable to resurrect the specter of 'eugenics' as the cornerstone of measures to allocate, deny and even supplement limited health resources in therapy interventions? A poorly constructed measurement instrument, one that fails, on a number of criteria, the standards of normal science, is no place to start. It is an analytical dead end; it shares the label pseudoscience with eugenics and its more extreme applications.

To be frank: there is no acceptable macro-resource allocation metric that can guide resource allocation in healthcare. It may be messy and less earth shattering, but the ICER framework must be rejected outright. Instead, we should look the need of patients and caregivers within disease states. This may be defined in quality of life terms, but only to the extent that competing therapy interventions can impact positively life's quality in target patient populations. We can provide suitable metrics, but only as inputs to negotiations between the respective parties.

The implications are clear cut: there is no basis for saying a product is cost-effective. There is no single criterion for cost effectiveness. Cost per QALY thresholds are mathematically impossible as well as failing the standards of normal science. Ersatz claims for cost-effectiveness can only be made if parties to a negotiation for a new product can agree on the attributes of interest, their ratio measurement properties and how they might be combined as ratio measures into a composite score. The same obstacles face attempts to emulate the eugenics fiasco in attempting to define an agreed composite metric for 'worth', expressing it as a criteria to support, at the margin, a reallocation of healthcare resources from denial though supplementation to maximize the societal 'benefits' of health expenditures.

REFERENCES

[1] Koch T. Life quality vs. the 'quality of life': assumptions underlying the prospective quality of life instruments in health care planning. *Soc Sci Med*. 2000; 51:419-27

[2] Galton F. Hereditary Genius. London: Macmillan, 1869

[3] 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

[4] 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

[5] Langley P. Suppling with the Devil: Belief and the Imaginary World of Multiple Myeloma Therapies Invented by the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(3): No 6
https://pubs.lib.umn.edu/index.php/innovations/article/view/4215/2917

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

[7] Tufts University. Center for the Evaluation of Value in Risk in Health. CEA Registry.
https://cevr.tuftsmedicalcenter.org/databases/cea-registry

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

[9] Hume D. An Enquiry Concerning Human Understanding. 1748

[10] 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

[11] Greene M. Until the End of Time. Knopf: New York, 2020

[12] 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

[13] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6 https://pubs.lib.umn.edu/index.php/innovations/article/view/3798