

More Unnecessary Imaginary Worlds – Part 2: The ICER Evidence Report on Modeling Oral Semaglutide for Type 2 Diabetes

Paul C Langley, PhD

Adjunct Professor, College of Pharmacy, University of Minnesota

Abstract

On 9 December 2019, the Institute for Clinical and Economic Review (ICER) released its final evidence report to establish the value of oral semaglutide (Novo Nordisk) for Type 2 diabetes (T2DM). A key element in this report was the development of a lifetime cost effectiveness microsimulation model based on a small sample of NHANES diabetes respondents. The model contrasted oral semaglutide added to current antihyperglycemic treatment for T2DM. The purpose of the model was to estimate outcomes that included life years (LYs) gained, an estimate of equal value life years gained (evLYGs), QALYs gained, clinical events, cost per major adverse cardiovascular events (MACE) avoided and total costs for each intervention over a lifetime time horizon. Previous commentaries in INNOVATIONS in Pharmacy have provided detailed critiques of the ICER modeling framework. While this model differs from previous ICER models, the result is still a framework that constructs a so-called evidence base that fails the demarcation test. It is best described as pseudoscience. The model creates, by assumption, an imaginary world. The claims made for oral semaglutide by ICER should not be taken seriously by health care decision makers. The purpose of this commentary is to point to the limitations of the model with particular reference to the utility metrics employed, the resulting claims for quality adjusted life years (QALYs) and consequent recommendations for price discounting and affordability.

Keywords: imaginary worlds, ICER, pseudoscience, diabetes, oral semaglutide, Rasch modeling

Introduction

The construction of assumption driven imaginary worlds to support incremental cost-per-QALY claims for pricing and access recommendations is the hallmark of the Institute for Clinical and Economic Review's (ICER) business model. ICER's evidence report on oral semaglutide for Type 2 Diabetes (T2DM) (Rybelsus, Novo Nordisk) follows this model. The evidence report, released on 1 November 2019, together with ICER's response to public comments, was followed by the formal ICER evidence presentation and voting on 14 November^{1,2}. The final evidence T2DM report with meeting summary was released on 9 December 2019³.

The primary aim of the ICER model, utilizing an individual, Monte Carlo based-microsimulation was to estimate the lifetime cost effectiveness of oral semaglutide added to current antihyperglycemic treatment for T2DM. The model was chosen because of the complexity of co-morbidities in Type 2 diabetes populations. The model outcomes included life years gained, quality adjusted life years (QALYs), clinical events, cost per major adverse cardiovascular event (MACE) avoided and total costs for each intervention over a lifetime treatment horizon.

The comparisons modeled involved oral semaglutide added to current antihyperglycemic treatment versus: (i) ongoing background antihyperglycemic treatment (e.g., metformin with

or without sulfonylureas), (ii) sitagliptin (Januvia, Merck), (iii) empagliflozin (Jardiance, Boehringer Ingelheim), and (iv) liraglutide (Victoza, Novo Nordisk); comparators (ii), (iii), and (iv) were also added to ongoing antihyperglycemic treatment. The target hypothetical population was adults with T2DM with inadequate glycemic control although receiving antihyperglycemic therapy. The range of modeling options considered and the various proposed endpoints, should not be taken as indicative of the merits of the ICER reference case model. It is an entirely imaginary framework which fails, as will be pointed out in this commentary, to meet the standards of normal science. This failure is made even more egregious, as detailed here, in the absence of any consideration of the need to meet fundamental measurement standards.

Even though The base-case modelled outcomes were considered by ICER to be 'highly uncertain' reflecting both variance in the model assumptions (input parameters and the risk equations) as well as uncertainties from the clinical network meta-analyses, ICER presented a range of outcome claims. Imaginary lifetime mean total cost for oral semaglutide was estimated at \$295,000 with other comparator costs ranging from \$250,000 to \$305,000. Imaginary lifetime QALYs were estimated at 4.09 for oral semaglutide with a range from 3.63 to 3.97 for the comparators. Incremental cost per QALY gained was modeled in the range \$110,000 for oral semaglutide plus background treatment at versus background treatment alone, \$140,000 for sitagliptin plus background treatments and \$480,000 for empagliflozin plus background treatment. Compared to liraglutide plus background treatment oral semaglutide was cost saving.

Corresponding author: Paul C Langley, PhD
Adjunct Professor, College of Pharmacy
University of Minnesota, Minneapolis, MN
Director, Maimon Research LLC; Tucson, AZ
Email: langley@maimonresearch.com
Website: www.maimonresearch.net

The imaginary world of ICER also generated recommendations for oral semaglutide discounting. While, as argued in this commentary, these certainly should not be taken seriously, ICER suggested a value-based benchmark price range that would achieve incremental cost-effectiveness between \$100,000 and \$150,000 per QALY gained. These are discounts between 32% and 36% from list price (WAC),

ICER, in discussing the limitations to its imaginary world modelling, pointed to: (i) the complexity of T2DM management; (II) unrepresentativeness of the NHANES sample for the T2DM subpopulation of interest as the basis for the microsimulation; (iii) lack of data on long-term follow-up for cardiovascular and renal outcomes including adherence; (iii) lack of concordance with clinical guidelines (*‘they have been muted for this modeling exercise’*); and (iv) lack of a *‘single comprehensive source of health-related quality of life inputs’* (discussed in more detail below). Given these limitations a reasonable question is why ICER would *‘endeavor to persevere’* with constructing an imaginary world ⁴. Presumably, this was dictated by the commitment to the ICER business model.

This commentary has a number of objectives:

- to point to the lack of a mandate in the US for ICER modeling;
- to make the case, following from previous commentaries, that the ICER lifetime reference case model lacks scientific credibility and should be seen as pseudoscience;
- to make the point that, even if one accepts ICER imaginary reference case models, the ICER model is only one of a possible multiverse of models where each purport to provide approximate information (or disinformation) for formulary decisions;
- to consider the failure of ICER to recognize the importance of meeting the standards of fundamental measurement in modeling evaluable claims for therapy options;
- the misuse of utility scores in creating QALYS by assuming they have interval measurement properties while they are in fact ordinal manifest scores
- to emphasize the importance of the Rasch model framework as the only approach to creating outcomes instruments that meets the standards of fundamental measurement;
- to consider, as a case study the limitations of the widely accepted Audit of Diabetes Dependent QoL (ADDQoL) instrument in its failure to meet fundamental measurement standards;
- to abandon QALYS as a meaningful metric for evaluating value claims
- to consider whether manufacturers and other stakeholders should continue to engage with ICER; and

- to consider options for evaluating the introduction of oral semaglutide that focus on, not imaginary constructs, but on empirically evaluable claims.

This commentary follows from a Part 1 commentary on the ICER final evidence report for JAK inhibitors ⁵. As a number of the arguments present are identical to those in the previous commentary, references are made rather than repeating the same arguments.

In achieving these objectives, it should be emphasized that deconstructing an ICER incremental cost-per-QALY model does not imply support for the imaginary world meme ⁶. While the construction of imaginary worlds has been the mainstay of health technology assessment claims over the past 30 or more years, ICER’s plea that it is applying *‘state of the art; techniques’* is both misleading and irrelevant. Constructing imaginary evidence, as considered in more detail below, is not how science advances, particularly when the claims for QALYs are nonsensical. They fail the axioms of fundamental measurement. They should be abandoned. Rather, as in drug development, technology assessment should be a process of discovery, not a process of fabricating one imaginary construct after another with various non-evaluable claims and value judgments. ICER, as noted in previous commentaries is an irrelevant and unnecessary distraction; unfortunately, against all common sense ICER, will be determined to persevere in marketing imaginary models and recommendations.

Mandate: ICER is not NICE

Previous commentaries have made the point that ICER is not an agency such as the National Institute for Clinical Excellence (NICE) in the UK or the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia. It is not even NICE-lite which is apparent comparing ICER’s perceived mission and the range of activities allocated to NICE. Most significantly, ICER has no legislative role. Its role as health technology assessment arbiter is assumed not anointed. Its recommendations have no regulatory impost; unless agencies are prepared to take its recommendations at face value rather than deconstructing claims which have no scientific merit.

Certainly, ICER mimics NICE and the PBAC (and others) in following a reference case or imaginary world paradigm. It has taken on board the health technology assessment meme, considered by those who believe in the construction of lifetime imaginary worlds to create *‘approximate information’* or *‘approximate disinformation’*, as the state of the art or dogma in health technology assessment. The difference is that NICE, the PBAC and others *‘require’* the construction of imaginary worlds. Manufacturers submit their imaginary world, it is challenged by academic assessment groups with decades of experience in dissecting these fairyland constructs, and an agreement on an imaginary world and its value recommendations made. ICER does not do this: it creates its

own imaginary world. Certainly, stakeholders can review but in the last resort, ICER wins out.

If the arguments presented here are accepted, then ICER may continue, but under a cloud. How can manufacturers and patient interest groups continue to pay any attention to ICER when its 'recommendations' are a fabrication. ICER may continue to create patently imaginary constructs which suffer from fatal methodological and measurement flaws and have the temerity, after a rather odd voting scene, argue for price discounts and access limitations. It is, perhaps time to move on.

The Standards of Normal Science

Again, previous commentaries have pointed to NICE standing aside from the standards of normal science; the focus on evidence and the discovery of new facts. Instead, ICER, like NICE, recycles existing 'facts' (assumptions) and creates an imaginary world to support value claims. This is, as noted in previous commentaries, nonsensical. If a new product or device is to be admitted to formulary and a price negotiated, we don't need fabricated imaginary value claims, but evidence for the impact of the product on a target treating population. At launch, we may only have clinical data. This may be sufficient to propose cost-outcome claims, but these must be credible, evaluable and replicable. If these data points are not to hand then, rather than creating an imaginary modeled world, the focus should be on claims assessment and feedback to a formulary committee. We need to provide the framework for new discovery; the process of hypothesis testing, conjecture and refutation, that has been in place since the scientific revolution of the 17th century. Not, it must be emphasized, a retreat to a medieval world where the search for new facts is discouraged; subsumed in an acceptance of imaginary and unsupported claims for therapeutic impact and value.

Assumptions

The ICER claim to fame is the ability to construct or fabricate an imaginary world that sets the stage for value impact over 10, 20 or 30 years in the future. The point is, and this goes back to Hume's (David Hume 1711 – 1776) induction problem; we cannot ask clients in health care to believe in models constructed on the belief that prior assumptions will hold into the future. It is logically indefensible. Certainly, models can involve assumptions but the difference is that the worth of the assumptions can be assessed if the model is designed to generate evaluable, credible and replicable claims.

Add to this the assumption that utility number lines have interval properties. Rejecting this assumption, for which there is no basis in instrument development, means that QALYs, the aggregation of QALYS over a modeled lifetime and consequent incremental cost-per-QALY claims have no meaning.

Utilities and QALYs

As detailed below, the utility metrics that ICER and other health technology model builders use to construct QALYs and cost-

per-QALY claims, fail to meet the axioms of fundamental measurement theory, including the axioms of conjoint simultaneous measurement where the construct or object to be measured is psychological or non-physical. Conjoint simultaneous measurement, and its modified version in Rasch measurement theory, provide the tools for detecting, if they exist, measurement structures that meet the criteria for fundamental measurement. Unfortunately, few patient reported outcomes (PRO) measures meet these standards. This includes both generic multi-attribute metrics such as EQ-5D utilities as well as disease specific instruments such as the ADDQoL

The obvious point that ICER and ISPOR overlook is that if an instrument is required then it has to be developed to meet fundamental measurement standards⁷. The key standards are for interval scales: invariance of comparisons and sufficiency⁸. Unless an instrument is designed from the ground-up to meet these standards for a single latent construct then any claims for interval scoring, unidimensionality, are wrong. The contribution of the Rasch model is to provide the framework for creating such instruments. This is not an opinion; it is a question of logic in mathematics. If you want to measure temperature, a latent single construct, then you create a thermometer. This has been the standard in the physical science since the work of Galileo in the early 17th century. So far, health technology assessment has ignored these requirements, opting instead for a discipline which extols the virtues of building 'approximate information' imaginary worlds.

Unfortunately, neither ICER nor ISPOR got the memo. We find the absurd situation where the most widely used textbook in health technology assessment simply assumes, in order to create QALYs, that the EQ-5D-3I/5L instruments have interval properties⁹. They do not. The reason? The EQ-5D-3L was not designed to have interval properties. The thought never occurred to the developers. It is no good saying that the EQ-5D-3L/5L have 'sort of' unidimensional characteristics. The fact is they do not and should be abandoned, along with the QALY notion that depends on them. Necessity, the need to defend the QALY measure in lifetime imaginary incremental cost-per-QALYT worlds, is not the mother of invention.

Although it is now some 60 years since the seminal contributions of conjoint simultaneous measurement and the Rasch model were proposed, with widespread acceptance in the social sciences, the essential contribution of Rasch modeling has not been recognized in health technology assessment; ICER has certainly not recognized it. ISPOR apparently feigns ignorance even though the question of measurement as a fundamental requirement in both the physical and social sciences is key to any assessment of change and hypothesis testing. Failing to recognize this, as Tennant et al point out in their contribution to a special section in the ISPOR house journal *Value in Health* (2004): *As long as primitive counts and raw scores are routinely mistaken for measures by*

*our colleagues in social, educational and health research, there is no hope of their professional activities ever developing into a reliable useful science*¹⁰

The outcome is the commitment to the creation of imaginary world resting on the ‘belief’ in assumptions for unidimensionality. Unfortunately, the notion of constructing QALYs from utility measures, lacks any scientific status. It is pseudoscience. The reason is quite simple: the utility metric lacks interval and ratio scale properties. It cannot be used to support arithmetical operations (addition, subtraction, multiplication, division). At best it can be considered an ordinal measure in calibrating response to the symptoms it captures as its contribution to the HRQoL family of ordinal measures. That is, it can support claims for modal and median values. Nothing more.

Any (ordinal) utility will do

Even if, for the purpose of argument we were to ‘accept’ the role of an ordinal utility measure as a ‘proxy’ for an interval measure, ICER faces further problems. Although a review of previous ICER evidence reports might have suggested that ICER has, in practice, adopted the EQ-5D-3L system as the preferred metric, this is put to one side in the evidence report for oral semaglutide. Apparently, at a pinch, any multi-attribute ordinal system will suffice: *The utility values for events modeled from the risk equations were drawn from two sources due to a lack of a single comprehensive source of health-related quality of life inputs. It is also important to point out that the two sources used different preference-weighted measures (EQ-5D and HUI3), and these two instruments are known to produce slightly different utility estimates* (emphases added)³.

For an organization which sees itself in pole position for reference case health technology assessment modeling in the US, as the arbiter for state of the art standards in the modeling of imaginary worlds, this is a most unfortunate statement. There are no references given for this claim, specifically for references supporting this claim for the target T2DM target population which ICER is attempting to model. Indeed, if ICER is to make unsupported claims for utility ‘equivalence’ then it should have provided a systematic review of utility metrics in the target T2DM population as proposed by ISPOR as ‘good practice’. Only then, for those who believe in the construction of imaginary worlds on ordinal utilities, could this assumption have been justified.

Indeed, ISPOR has issued practice guidelines that address this question with standards proposed for reporting on the choice of algorithm and utility metric^{11 12}. Again, as these guidelines were available to ICER, it was surely incumbent upon ICER to demonstrate that it had considered and applied these guidelines in its choice of utility metrics.

Direct Medical Costs

If willingness to play thresholds are to be center stage in the ICER world, then care also needs to be taken regarding cost

estimates. If there is no guideline for costs and the projected costs of the reference case timeline then there is considerable flexibility in the cost assumptions supporting the numerator of the incremental cost-per-QALY ratio. This is not clear from the ICER reference case, although ICER admits to ‘difficulties’ in projecting direct medical cost. One result is that no attempt is made to project future drug prices; a somewhat puzzling decision given evidence for manufacturers endorsing long-term pricing policies.

Choose your disease stage

If we continue to believe, for the purpose of argument, in the relevance of the reference case model meme, the next step in the creation of imaginary QALYS is to estimate the time spent in different disease stages to which our EQ-5D or other ordinal utilities are to be attached. Again, the options are wide open. Different models will generate different times spent in disease stages. There is no standard in the health technology meme to guide model assumptions. Neither is there any standard to guide how serious adverse events with disease stages are to be modeled. It is, to use a tired cliché, ‘modeling open season’.

If there are competing models that generate different estimates of time spent, and then even with the same utility metric, the number of estimated QALYS, as the product of the two measures, will differ. ICER needs, therefore, to quite clear as to the basis for utility and time spent by disease state estimates is ‘by assumption’.

Cost per QALY

Multiplying time spent in a disease state by an ordinal utility metric is, of course, nonsensical. This means that, in addition to the choice of model framework and assumptions to fabricate claims for time spent in disease stages over the lifetime of a hypothetical patient cohort, the next step in the creation of lifetime costs per QALY also makes no sense. Certainly, we could assume that the ordinal utility is a proxy for an interval measure; after all it would be just one more assumption. Unfortunately, the ICER model is only one of a possible multiverse of models, each generating competing imaginary claims. .

The Model Multiverse

The issue with creating imaginary worlds is that there is no limit to the number of imaginary worlds that model builders can create. The ICER reference case is quite flexible; any ICER model can be challenged. As the imaginary cost-per-QALY paradigm is central to the reference case, any number of competing imaginary cost-per-QALY claims can be created. Matched to threshold values opens the hatches to a prospective tsunami of competing imaginary claims. Unlike the UK where NICE will contract with imaginary world referees, typically academic centers, to act as assessors, ICER acts as its own judge. The ICER modelled world, with its foundation in ordinal utilities and QALYS that are simply manifest scores, takes center stage.

Without deconstruction, its claims for cost-effectiveness are taken at face value.

Given ICERs commitment to the imaginary QALY, it is worth noting that a previous commentary, following a review of the practical impact of modeled cost-per-QALY claims and the unlikely event that they would ever be followed up as credible and evaluable hypotheses, concluded that: *In retrospect, it is doubtful, that the great expectations for QALYs could ever be realized outside of reference case imaginary worlds, or the willingness of decision makers to suspend belief in the standards of normal science, and accept lifetime cost-per-QALY claims as decision criteria. Unless, therefore, a case can be made for short-term and evaluable QALY claims, there seems little scope for QALYS, and associated cost-per-QALY claims, as inputs to formulary decision making. Perhaps, as Pip d says to Estella, it has been 'a vain hope and an idle pursuit' ¹³. After over 30 years perhaps we can put QALYs to one side and return to clinically and quality specific endpoints in comparative claims for pharmaceutical products in disease and therapeutic areas ¹⁴.*

Threshold Willingness to Pay

The degree of flexibility in choice of assumption in fabricating lifetime reference case models has obvious yet important implication for ICER value claims. Given the options in utility metric, choice of mapping function, estimates of time spent stage of disease, modeling of adverse events within disease stages and modeled direct medical costs, any claim for price discounting or affordability alerts, as in the case of oral semaglutide, will be a function of the model assumptions and model specification. Apart from the fact that cost-per-QALY claims lack credibility given their ordinal basis, for those who wish to continue they have to face the issue that any threshold application is dependent on the model structure and assumptions. Different fabricated imaginary worlds will produce different pricing and affordability recommendations; the ICER model is just one of many. Threshold value claims which are, unfortunately, taken at face value by the media actually make little sense.

Claims for price discounting will vary depending on the components of the incremental cost-per-QALY ratio. Competing models, if the same threshold cut-off is applied recommendations for price discounting will vary. In fact, claims for price discounting will be unique to the imaginary construct. Vary the model assumptions and the discounting recommendation will change. A \$50,000 per QALY threshold for one model for a product will result in different value judgements for price discounting with other models because the imaginary cost-per-QALY estimate will differ. With the same threshold, different models will produce different recommended price adjustments. If ICER is to subscribe to some 'universal' application of specific willingness to pay thresholds then these must be defined for the model structure and assumptions. *Ceteris paribus*, if the threshold is applied for \$50,000 per QALY then the utility metric, the EQ-5D-3L for

example, must be used exclusively, with a specified mapping function to generate the utility metric. If there is a literature search to locate utility metrics then it must identify utility metrics that are comparable if mapping functions are used. Competing models will yield different pricing conclusions. ICER should make this quite clear in media releases.

Diabetes Impact on QoL

Although this commentary is focused on the ICER model claims for oral semaglutide for patients with T2DM, it is of interest to consider attempts to date to provide QoL, not HRQoL, assessments of the impact of diabetes. The most widely reported instrument, for both T1DM and T2DM is the Audit of Diabetes Dependent QoL (ADDQoL)¹⁵. Initially developed in the 1990s this has gone through a number of modifications without changing its basic structure. It is not the intention here to challenge the use of this instrument as a possible profile for review with the patient by their treating physician, but to raise some concerns that users should be aware of in interpreting claims based on the instrument (and spin-offs from the instrument). These concerns are to be seen against the background of the Rasch model and its adherence to the axioms of fundamental measurement. If these are met then, as noted above, the derived latent estimate is an interval scale. Calibrated change, addition and subtraction, require an interval scale. With the Rasch model, we might assume that there exists a true zero which then allows the range of arithmetic operations. The ADDQoL instrument was not developed with fundamental measurement as its lodestar.

Briefly, the ADDQoL takes a patient centric approach to assessing QoL impact, promising an individualized measure of the impact of QoL. There are two overview items: (i) "in general my present QoL is ...and (ii) If I did not have diabetes my QoL would be Both are 7 response Likert scales. There are a total of 13 domain items ..." If I did not have diabetes my Would be ..., These are 'scored' by assigning integer values to each levels (-3 to +3) as arbitrary weights. Summing and averaging over these arbitrary weights attached to an ordinal scale yields an overall 'weighted' ordinal score average with a range from -9 (maximum negative impact of diabetes) to +3 (maximum positive impact of diabetes).

In interpreting these average scores, It is important to remember that an ordinal scale, such as the Likert scale, 'orders' the elements of interest but the difference between the 'orders' or 'levels' described in the scale are unknown. The difference between an assigned 'value' of 2 and 3 for one Likert type response for one QoL item cannot be compared to the difference between a 2 and a 3 for another item because we have no idea of the difference between the numerals 2 and 3 for each item. Values or weights have been assigned on the mistaken belief that they have a common interval value or one that has been assigned by attaching arbitrary integer units. This is 'encouraged' by assigning integer values to response levels; a misleading impression that the scale has interval and ratio

properties with a true zero. Certainly, you can report modes or medians, but nothing else. Arithmetical or statistical operations are out, including assigning weights to the item elements (weighting an unknown difference) and creating an ‘average’. To avoid confusion, it would be more appropriate to assign alpha rather than numeric ‘scales’.

However, if the user wishes to interpret these ordinal weighted average manifest raw scores as interval scores it is their choice. The point to make, if the Rasch model is taken as the gold standard for creating QoL instruments, in this case for diabetes (but perhaps separate Rasch model instruments for T1DM as opposed to T2DM diabetes) then we can assign interval properties, but to a needs based instrument that will be quite different from the ADDQoL. For those who may object to putting the ADDQoL to one side, after all it has been applied for the last 25 years; the case is that the ADDQoL scores have no meaning in terms of the axioms of measurement. If the intent is to generate from non-physical or psychological constructs robust measurement structures then the Rasch model provides the appropriate (and well tested) framework. Fitting data to the Rasch model rather than fitting a model to data provides the basis for disease-specific needs based QoL measures. This is absent in the ADDQoL instrument.

Public Comments and Response

Although only a handful of manufacturers and groups submitted comments to ICER in response to the release of the draft evidence report on 12 September 2019, there were a few common themes in these comments that relate to the methodological issues raised here in respect of ICER models. While there was no direct challenge to ICER on its failure to meet standards for evidence based claims and fundamental measurement, issues raised were:

- Why did ICER persist in going forward with its imaginary modeling on a limited evidence base and its continued use of the phrase ‘significant uncertainties’?
- Why a failure to include anticipated price changes, in particular when sitagliptin goes generic?
- Why continue to assume trial based adherence rates?
- Why introduce utilities from disparate sources instead of waiting for US preference utilities to be reported?

ICER’s responses are not unexpected. ICER insists on going forward to fabricating imaginary worlds even with a limited evidence base for its assumptions:

- *....clinicians, patients, and all stakeholders will need to make decisions about treatment long before data become available that would allow comparisons between oral semaglutide and other options for therapy*

- *Decision makers must make a well-reasoned decision today and our effort is to produce an estimate of value for this product with the available data*
- *Health care payers must make decisions about funding now and cannot hold out for perfect information. We appreciate the concern that uncertainty and heterogeneity exist, and thus our report produces a range of cost-effectiveness estimates and thresholds with confidence ranges to aid decision makers.*
- *....we feel we have adequately highlighted uncertainty.*

In addition:

- *Including assumptions about price changes is not currently the standard in academic or health technology assessment agency cost effectiveness analyses. In part, this is because it is very difficult to predict the pricing landscape years into the future.*
- *Trial-based discontinuation rates are included in the model estimates. However, we have no long-term adherence data for oral semaglutide and thus have assumed equal adherence in subsequent model cycles.*
- *We acknowledge that there could be differences in patient preferences between countries. However, we did not identify any other sources for the disutility associated with injections in people with Diabetes that were specific to the US*
- *Furthermore, we utilized multiple sources of utility estimates because there was not a single source that could provide all of the necessary input values for the model.*

Affordability

Although a therapy may meet ICER’s fabricated willingness to pay thresholds for cost-effectiveness as determined by the imaginary modeled diabetes world, this first hurdle may be surmounted only to be halted at the second hurdle: ICER’s potential budget impact threshold.

In May 2019, ICER determined that the annual budget impact threshold for each individual new molecular entity would be \$819 million. If projected annual US spending on a specific drug exceeds this threshold then ICER will determine the maximum number of eligible patients who would be able to receive the therapy, at multiple possible pricing points (lower than the price deemed effective in the first hurdle analysis) without exceeding the threshold.

The final evidence report concludes that the JAK inhibitors budget impact falls within this arbitrary ceiling. Whether anyone should take this back-of-the-envelope rationing alert seriously is a moot point. To recommend a ceiling for patient access to meet a notional budget threshold is to put to one side assessed clinical benefits for the individual patient, and

whether this merits additional funds being allocated, as well as potentially creating waiting lists for access. It is all well and good to recommend prior authorization but without recommended criteria for approval/refusal, it is a hollow recommendation. After all, it would be presumably possible to translate the aggregate budget limit into QALYs and estimate the allocation of QALYs to each molecular entity and estimate the number of patients allowed to utilize the therapy! Unfortunately, this would raise the question again of why ordinal and generic QALYs are used when the focus is presumably (again) on the benefits and harms to patients.

As noted above, ICER recommends a price reduction (from WAC) of between 32% and 36% for oral semaglutide. This is equivalent, following ICERs lifetime imaginary construct, to an annual price of \$5,983 to \$6,396 for the two thresholds. While the precision of these conclusions should not disguise the underlying lack of scientific merit in the ICER case, this implies limiting patient access ('trigger policy actions' – undefined) to between 13% and 18% of eligible patients as a switching therapy at threshold prices. As an add-on, therapy at the threshold price the percentage of patients who could be treated without crossing the budget impact threshold is between 6.9% and 8.1%. These claims should not, of course, be taken seriously.

Expert Panel Voting

The last step in the ICER model gestation period is to submit to an expert panel of ICER appointees a series of questions regarding the model and its value claims. The panel is asked to vote on the questions. Their responses are tallied and published in the final evidence report.

As this commentary is focused on the reference case model, only one question (of the 7 asked) is relevant. This is (Q7) where the question is whether the voting panel believes that oral semaglutide has a low, intermediate or high long-term value for money at current pricing versus ongoing background therapy alone at current pricing. As noted in a previous commentary, asking this question presupposes that each respondent on the panel is familiar with the construction of imaginary worlds, in this case microsimulation, and the requirements for fundamental measurement. If not, as in the case of health system decision makers, the respondents are asked to take ICER's claims at face value. The final evidence report provides the panels considered opinions: low 4 votes; intermediate 6 votes and high 2 votes.

Conclusions: an unfortunate truth

ICER has not demonstrated that, in the case of T2DM, that the EQ-5D health dimensions and response levels are items that are relevant in assessing the value T2DM patients attach to their experience of oral semaglutide. There is no indication from the ICER evidence report that this was even considered. Utilities are just 'dropped in' with no thought to their relevance, their strengths or their limitations. From a patient value perspective,

therefore, the imposed use of the EQ-5D is of little interest. If formulary committees are interested in their value to patients, as opposed to clinicians, then the choice of instrument is misplaced. The information provided is not only 'approximate' (whatever that means) but also irrelevant.

The question of how to reflect the patient voice through the development of patient-centric measures has been recognized for over 40 years. Although ICER seems oblivious to these, it is worth pointing out that the patient-centric general health question, the Nottingham Health Profile, was developed in the mid-to-late 1970s and has been used in a significant number of studies and in different language versions¹⁶. In addition, there are now more than 30 disease specific patient-centric instruments developed by Galen Research, Manchester, UK based on the requirements of RMT.

There is an extensive literature on how a needs-based instrument should be developed^{17 18}. The critical elements are: (i) the role of unstructured qualitative interviews with patients to identify item selection for the measure covering all aspects of the patient's life; (ii) ensuring that the item selection process is independent of specific interventions; (iii) identifying those outcomes most relevant to patients; (iv) creating, through Rasch analysis, a unidimensional scale supporting a total score; and (v) creating a scale that is reliable valid, easy to administer, avoids complex language, complex responses and is able to detect change.

Clearly, a patient-centric instrument, which has yet to be developed for T2DM, would be a marked improvement in capturing the patient voice' in T2DM over the EQ-5D. It would also put to one side the reference case model in focusing on the response in real time, through a Phase 4 RCT or a platform of T2DM patients.

ICER is apparently not concerned with recognizing the importance of fundamental measurement in its claims for utility scores. With its preferred EQ-5D-3L multi-attribute utility metric, ICER is committed to the use of ordinal scales. This undercuts the entire ICER value claim from the cost per QALY imaginary or fantasy world. Clearly, any model can generate estimates of future imagined time spent in various disease stages, with added adverse events as plus utility brownie points, with utility scores either from the literature or 'constructed' from expert panels or the modeler's imagination. Given the weights attached to time spent, an aggregate estimate of lifetime QALYs is created,. Unfortunately, multiplying units of time spent by disease stage (presumably cardinal scores) by ordinal utilities for that experience of time spent makes no sense. Extending this to cost-per-QALY ratios makes no sense either. QALY based claims for oral semaglutide, if this argument is accepted, fail to have any meaning. They are an artifact that fails standards for fundamental measurement. QALYs, as a key element in the health technology assessment meme, are resilient. They will continue to be promoted, just as

relics and indulgencies were after 1517. A belief is not easily overturned. The health technology assessment meme, as detailed in previous commentaries is resistant; it has significant transcription fidelity and it is a strongly held dogma. Its champions, such as ISPOR, will not want to hear that the emperor has no clothes. After all, why have manufacturers been supporting consultants to build imaginary worlds for the past 30 or more years? This is understandable where the single payer health agency mandates constructing imaginary reference case world; it is another in the US where there may be a greater understanding of the role of fundamental measurement and the need for credible and evaluable comparative product claims.

It is one thing to encourage ICER and health care decision makers to drop claims based on an incremental ordinal and nonsensical cost-per-QALY paradigm, it is another to establish invariant measurement standards for claims assessment. The Part 1 study and other previous commentaries have made the case that, if we are concerned with therapy impact assessment in target populations, then the assessment of response should be specific to the QoL of that target population. We have to return to basics; the sunk capital of technology assessment over 30 years will be an obstacle. There will also be some degree of embarrassment.

Understandably, ICER wants to defend its business model and its self-appointed task of issuing evidence reports to, as far as possible, coincide with product marketing approval. This apparently trumps, to use the current vernacular, questions of the strength or otherwise of the evidentiary base to support clinical pronouncements, modeled imaginary cost-per-QALY claims and constructed affordability alerts. Even though ICER believes it has taken 'adequate' account of uncertainty and has flagged where 'significant uncertainties' exist, the point remains that ICER has constructed a lifetime model whose claims cannot be evaluated and were never intended to be evaluated. Attaching scenario uncertainty claims to a model that lacks any scientific credibility is clearly a waste of time. The 'well-reasoned' ICER model is seen, by ICER, as essential (or at least an input) to formulary decisions in its construction of an unknown future. The future is a mystery, yet ICER has a belief that its vision, its belief in the reference case meme, is sufficiently 'approximate' yet unknowable information, for formulary decisions. The audience must share ICER's faith¹⁹.

Unfortunately, the technology assessment burden that ICER has taken upon itself is neither viable nor sustainable. Certainly, ICER will persevere; its business case rests upon its self-appointed role as the arbiter of health technology claims. If the ICER reference is the NICE model, then there is no reason it should be adopted in the US. This is even more problematic, given that the NICE reference case also fails to meet the standards of normal science.

Next Steps: Rejecting Imaginary Worlds and Responding to ICER

Manufacturers and other stakeholders in a disease or therapy area, who are considering engaging with ICER in its process of developing imaginary worlds for its evidence reports, have two options. First, they can take the position that the ICER business model fails to meet the standards of normal science, inform ICER accordingly and reject any engagement or, second, they can put to one side concerns over the pseudoscientific nature of the ICER model and attempt to engage with ICER to make the model 'more' relevant to decision makers. Adopting the second option can be seen as endorsing the ICER technology assessment meme. The manufacturer may then be placed, if the ICER recommendations are for significant price discounts or involve affordability alerts, in the position of challenging the ICER model, its structure and assumptions. This may involve resources to explain to health decision makers the limitations of the ICER approach. The result being the manufacturer investing in such an exercise, challenging the ICER imaginary world versus competing imaginary worlds. Irrespective of the disease area or the therapies chosen for review by the ICER team, the model selected for the respective evidence report is merely one in a potential multiverse of constructed models. While ICER may claim pole position in imaginary evidence constructions, the evidence report model of ICER can be easily challenged by other imaginary constructs given the options open to change assumptions and the construction of competing models within the same reference case paradigm^{20 21 22}.

If science is about discovery, then the focus on 'state of the art' imaginary worlds is both misplaced and misleading. If we are concerned with building on the results of RCTs to assess the impact in treatment practice of new and competing therapy interventions, then this has to be both evidence driven and reportable to formulary committees in a meaningful time frame. It is absurd to substitute imaginary worlds for a program designed to add to our knowledge of therapy interventions.

In late 2017, the Program in Social and Administrative Pharmacy at the University of Minnesota, published its second edition of proposed formulary submission guidelines^{23 24}. The guidelines proposed that if a manufacturer made a submission for formulary submission then this should be accompanied by a protocol or proposal for how the therapeutic claims were to be assessed. These claims could be expressed in purely clinical terms or as cost-effectiveness claims, to include HRQoL

The recommended next step, after putting ICER to one side, is to focus on the issue of evaluating patients' needs in T2DM through developing a disease specific instrument to evaluate response to therapy. Not the measurement of functions and symptoms, but a need-based measure. This will set the stage for a discovery-focused research program capturing the patient voice; a program that meets the standards of normal science.

Conflicts of Interest PCL is an Advisory Board Member and Consultant to the Institute for Patient Access and Affordability, a program of Patients Rising.

References

- ¹ ICER. Evidence Report – Oral Semaglutide for Type 2 Diabetes. 1 November 2019. https://icer-review.org/wp-content/uploads/2019/09/ICER_Diabetes_Evidence-Report_110119.pdf
- ² ICER. Diabetes: Response to Public Comments. https://icer-review.org/wp-content/uploads/2019/11/ICER_Diabetes_Draft_Report_Response_to_Comments_110119.pdf
- ³ ICER. Oral semaglutide for Type 2 Diabetes: Effectiveness and value. Final Evidence Report. 9 December 2019 https://icer-review.org/wp-content/uploads/2019/04/ICER_Diabetes_Final-Evidence-Report_120919.pdf
- ⁴ Chief Dan George: Lone Wattie in *The Outlaw Josie Wales* (1976).
- ⁵ Langley PC. More unnecessary imaginary worlds – Part 1: The Institute for Clinical and Economic Review’s report on Janus Kinase (JAK) Inhibitors in rheumatoid arthritis. *Inov Pharm.* 2020;11(1): No. 2
- ⁶ Neumann P, Willke R, Garrison L. A health economics approach to US value assessment frameworks – Introduction: An ISPOR Special Task Force Report (1). *Value Health.* 2018;21:119-25
- ⁷ Bond T, Fox C. Applying the Rasch Model. 3rd Ed. New York: Routledge, 2015
- ⁸ Grimby G, Tennant A, Tesio L. The use of raw scores from interval scales: Time to end malpractice (Ed.). *J Rehab Med.* 2012;44:97-98
- ⁹ Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes. 4th Ed. New York: Oxford University Press, 2015
- ¹⁰ Tennant A, McKenna S, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health.* 2004;7(1 Suppl):S22-S26
- ¹¹ Wailoo A, Hernandez-Alava M, Manca A et al. Mapping to estimate health-state utility from non-preference-based outcomes measures; An ISPOR Good Practices for Outcomes Research Task Force. *Value Health.* 2017;20(1):18-27
- ¹² Brazier, Ara R, Azzabi I et al. Identification, review, and use of health state utilities in cost-effectiveness models: An ISPOR Good Practices for Outcomes Research Task Force Report. *Value Health.* 2019;22:267-75
- ¹³ Dickens C. Great Expectations. London, 1861
- ¹⁴ Langley PC. Great Expectations: Cost-utility models as decision criteria. *Inov Pharm.* 2016;7(2); Article 14 <https://pubs.lib.umn.edu/index.php/innovations/article/view/437>
- ¹⁵ Bradley C, Todd C, Gorton E et al. The Development of an individualized questionnaire measure of perceived impact of diabetes on quality of life: The ADDQoL. *Qual Life Res.* 1999. 8(1/2): 79-91
- ¹⁶ Hunt S, McEwan J, McKenna S. Measuring health status: a new tool for clinicians and epidemiologists. *J Royal College Gen Practitioners.* 1985;35:185-88
- ¹⁷ McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes 1: The search for the Holy Grail. *J Med Econ.* 2019;22(6):516-522

- ¹⁸ McKenna S, Heaney A, Wilburn J. Measurement of patient-reported outcomes. 2: Are current measures failing us? *J Med Econ.* 2019;22(6):523-30
- ¹⁹ Apocryphal: Oliver Cromwell (1599-1658) instructions to the painter Peter Lely (1618-80)
- ²⁰ Langley PC. Resolving Lingering Problems or Continued Support for Pseudoscience? The ICER Value Assessment Update. *Inov Pharm.* 2017;8(4): No 7 <https://pubs.lib.umn.edu/index.php/innovations/article/view/933>
- ²¹ Langley PC. Transparency, Imaginary Worlds and ICER Value Assessments. *Inov Pharm.* 2017;8(4): No 11 <https://pubs.lib.umn.edu/index.php/innovations/article/view/926>
- ²² Langley PC. Alternative Facts and the ICER Proposed Policy on Access to Imaginary Pharmacoeconomic Worlds. *Inov Pharm.* 2018;9(2): No. 10 <https://pubs.lib.umn.edu/index.php/innovations/article/view/1300>
- ²³ Langley PC. Guidelines for Formulary Evaluation [Proposed]. Program in Social and Administrative Pharmacy. College of Pharmacy. University of Minnesota. Version 2.0. December 2016. https://www.pharmacy.umn.edu/sites/pharmacy.umn.edu/files/minnesota_guidelines_december_2016.pdf
- ²⁴ Langley PC. Nullius in Verba: Version 2.0 of the University of Minnesota, School of Social and Administrative Pharmacy Program, Proposed Guidelines for Formulary Evaluation. *Inov Pharm.* 2016;7(4): Article 16 <https://pubs.lib.umn.edu/index.php/innovations/article/view/478>