MAIMON WORKING PAPERS

Working Paper No. 4 March 2020

Anticipating Imaginary Claims: The ICER Modeling Analysis Plan for Targeted Immune Modulators in Ulcerative Colitis

Paul C Langley, PhD
Adjunct Professor, College of Pharmacy, University of Minnesota

Abstract
In early February 2020 the Institute for Clinical and Economic Review (ICER) released its Modeling Analysis Plan for the evaluation of Targeted Immune Modulators in Ulcerative Colitis. The value assessment framework proposed was the ICER reference case standard for creating a lifetime imaginary world, assumption driven, to generate recommendations for pricing and access. While the draft evidence report is not scheduled until early April, it is important for those engaged as stakeholders with the ICER (the three manufacturers are Pfizer, Janssen and AbbVie) to recognize the limitations implicit in the IER approach. If they wish to engage with ICER then it should be from a position that argues that the ICER reference case value assessment fails to meet the standards of normal science. In consequence, any recommendations for pricing and access should be rejected.

Keywords: imaginary worlds, ulcerative colitis, pseudoscience, nonsense claims, nonsense recommendations

---

Introduction

One of the more intriguing activities of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) over the past 20 or more years has been the promotion of cost-effectiveness claims based on the construction of imaginary worlds. The key concept is 'approximate information'. Practitioners in health technology assessment, notably in situations where a new therapy has received marketing approval but where data other than those generated by Phase 2 and Phase 3 clinical trials are limited, are recommended to build a reference case imaginary world. This is an exercise, not in hypothesis testing, but in the manufacture of 'approximate' imaginary information.

The reference case is a set of imaginary modeling standards, established by agencies such as the National Institute for Health and Care Excellence (NICE) in the UK and, as the self-appointed lead technology assessment group in the US, the Institute for Clinical and Economic Review (ICER) to create an model simulation extending 10, 20 or 30 years in the future. The purpose of the simulation, essentially a fantasy exercise, is to propose an unknown future reality in which the response of a hypothetical population to selected therapy interventions can be simulated.

The purpose of this commentary is to consider the proposed modeling analysis plan, the imaginary world blueprint, for ICER's value assessment targeted immune modulators (TIMs) for moderate to severe ulcerative colitis (UC). A total of eight products are proposed for assessment within this imaginary simulated framework: adalimumab (Humira: AbbVie); golimumab (Simponi; Janssen Biotech); infliximab (Remicade: Janssen Biotech); infliximab-dyyb (Inflectra: Pfizer); infliximab-abda (Renflexis; Merck); tofacitinib (Xeljanz: Pfizer); ustekinumab (Stelara: Janssen Biotech); and vedolizumab (Entyvio: Takeda). In respect of the last product only the IV formulation will be considered. The interventions will be compared to each other and to conventional treatment defined as induction with corticosteroids followed by azathioprine or mercapotopurine.

The Proposed Imaginary World

In common with all ICER evidence reviews, the comparative assessment of these TIMs for UC, is followed by recommendations for pricing and access. These recommendations rest upon the construction of one of potentially many imaginary lifetime simulations. The selected imaginary world follows the ICER value assessment framework a decision analytic model is proposed with a hypothetical cohort of patients with moderate to severe UC treated with TIMs. The model takes a lifetime time horizon with a treatment cycle length of 8 weeks. The results of randomized controlled trials (RCTs) and network meta analyses will support assumptions regarding the likelihood of achieving induction and maintaining response without remission or response with remission. Patients will cycle through two TIMs before discontinuing. Health state utilities will be from a meta-analysis as the basis for creating quality adjusted life years (QALYs) by disease stage. Health state costs, again assumed, will capture hospitalization, emergency department visits and outpatient visits based on published claims analysis.

It should be emphasized that the ICER value assessment framework (VAF) is concerned solely with the construction of imaginary worlds; modeled simulacra that ICER believe are a 'reasonable approximation' to an unknown future reality for disease areas and products selected for review. A recent

commentary in INNOVATIONS in PHARMACY reviewed the latest ICER VAF to be applied over the period 2020 to 2023 [1]. The commentary concluded that the VAF failed to meet the standards of normal science; it was considered pseudoscience. The principal reason for the ICER VAF failing the demarcation criteria between science and pseudoscience (or pure bunk) was its rejection of modeled claims that allow empirical evaluation. At the same time, the ICER VAF fails the standards for fundamental measurement [2]. It applies utilities which are manifest scores; they have ordinal rather than interval scale properties. This means that the consequent QALY and cost-per QALY estimates are meaningless. The consequences are, for products such as TIMs for UC, that conclusions regarding estimated value and proposals for lifetime cost-per-QALY, with consequent recommendations for price discounting are unsupportable. They are an unnecessary illusory distraction, irrespective of the degree of precision presented.

The purpose of this commentary is to build on the analyses and arguments presented in previous commentaries, notably the review of the ICER 2020-2023 VAF, to make the case for rejecting the proposed marketing analysis plan; not for an alternative imaginary world but for rejecting the proposed analysis as failing the standards of normal science. It is important to make clear what these standards are and why the ICER VAF is a non-starter if we are to understand the benefits and costs of competing therapies in UC.

The Standards of Normal Science

The requirement for testable hypotheses in the evaluation and provisional acceptance of claims made for pharmaceutical products and devices is unexceptional. Since the 17th century, it has been accepted that if a research agenda is to advance, if there is to be an accretion of knowledge, there has to be a process of discovering new facts. By the 1660s, the scientific method, following the seminal contributions of Bacon, Galileo, Huygens and Boyle, had been clearly articulated by associations such as the Academia del Cimento in Florence (1657) and the Royal Society in England (founded 1660; Royal Charter 1662) with their respective mottos *Provando e Riprovando* (prove and again prove) and *nullius in verba* (take no man's word for it) [3].

By the early 20th century, standards for empirical assessment were put on a sound methodological basis by Popper (Sir Karl Popper 1902-1994) in his advocacy of a process of 'conjecture and refutation [4] [5]. Hypotheses or claims must be capable of falsification; indeed, they should be framed in such a way that makes falsification likely.

Although Popper's view on what demarcates science (e.g., natural selection) from pseudoscience (e.g., intelligent design) is now seen as an oversimplification involving more than just the criteria of falsification, the demarcation problem remains [6]. Certainly, there are different ways of doing science but what all scientific inquiry has in common is the 'construction of empirically verifiable theories and hypotheses'. Empirical testability is the 'one major characteristic distinguishing science from pseudoscience'; theories must be tested against data. We can only justify our preference for a theory by continued evaluation and replication of claims. This applies in cystic fibrosis just as it does in other therapies. Constructing imaginary worlds, even if the justification is that they are 'for information' is, to use Bentham's (Jeremy Bentham 1748-1832) memorable phrase 'nonsense on stilts'. If there is a belief, as subscribed to by ICER, in the sure and certain hope of constructing imaginary worlds, to drive formulary and pricing decisions, then it needs to be made clear that this is a belief that lacks scientific merit.

If we consider proposed imaginary scenarios to make claims for social costs, productivity costs and indirect costs of disability, these are similarly non-starters. To be taken seriously these outcomes should be put in empirically evaluable terms. But of course they will not; yet more examples for the fantasy world driving fanciful claims.

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. In the cystic fibrosis model of the therapies offered by Vertex Pharmaceuticals, the number of assumptions made to support the microsimulations across the four patients groups is truly awesome; some come from the literature, others are pure guesswork. Unfortunately, even if an assumption driving the imaginary value assessment framework is defended by appealing to the literature (including pivotal clinical trials) the effort is wasted.

The point, and this goes back to Hume's (David Hume 1711 – 1776) induction problem, is that 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: it cannot be '*established by logical argument, since from the fact that all past futures have resembled past pasts, it does not follow that all future futures will resemble future pasts*' [7].

Proximus Notitia

The claim and acceptance by those supporting ISPOR and ICER is that the role of health technology assessment is to generate approximate information rather than testing hypotheses is absurd. We have no idea, nor do health system decision makers, of the meaning of the term. Are we to consider the ICER claims as approximating some unknown and unknowable ultimate 'truth'? Or does each model, from a multiverse of possible models based on some 'approximation' to the ICER VAF, reveal itself as the 'best' approximation to an absolute truth for that particular model? How are we to distinguish, in the absence of 'evaluable claims; 'approximate information' from 'approximate disinformation'? It is a ridiculous assertion.

Utilities

Quality adjusted life years (QALYs) can only survive if the measure is credible, evaluable and replicable. The QALY constructed by ICER in the cystic fibrosis model meets none of these criteria. The concept of a QALY is not new; it goes back some 40 plus years with the notion of combining time spent in a disease state with some multiplicative 'score' on a required interval scale of 0 to 1 (death to perfect health), Combining the two, multiplying time by utility is assumed to produce a QALY. In the ICER imaginary UC world these are combined to produce QALYs for the modeled life span.

The proposed modeled imaginary world for UC is to be populated by utilities drawn from a systematic review of the utilities reported by patients with Crohn's disease and ulcerative colitis [8]. The study by Malinowski and Kawalec comprises a systematic literature review, a meta-analysis of reported utilities in Crohn's disease and UC, followed by meta-regression analysis. Unfortunately, the authors fail to recognize that if you wish to undertake statistical operations on utility metrics than, individually, each of those metrics (EQ-5D, HUI, etc.) must meet fundamental measurement properties. As this is not demonstrated, the entire analysis is a waste of time. In fact, the generic utilities employed are manifest scores; ordinal scales where the only permissible operations are to present medians and modes. This point has been made at length in a previous commentary [1].

Apart from demonstrating the rather obvious point that different instrument generate different utility scores for the same underling stage of disease, together with the further obvious conclusion that the more severe a disease stage the lower the utility, the analysis has little to offer. There is no evidence cited that the measures have anything other than ordinal measurement properties; nor is there any suggestion that the instruments in their development were designed to meet fundamental measurement properties. Even if we made the assumption that the various measures had a required measurement property, attempting an aggregate score across such diverse measures, makes no sense.

It is not clear, at this stage, what utility measures the ICER model for UC will adopt. One option, holding to the belief in the obvious fundamental measurement standards of generic utilities, that ICER will choose the utility metric generated by the meta-analysis. These range from an 'aggregate' manifest score of 0.8726 (note the precision) for persons in remission with UC to a low of 0.6992 for those with active UC. Among those with active UC, the illusory manifest scores range from 0.7059 for severe UC to 0.7834 for those with mild UC. As these are ordinal scales we have no idea of what the difference between the mild to moderate, moderate to moderate/severe or moderate to severe difference actually means. They are just approximations to a rank ordering with no ratio (or even interval) scale properties.

If ICER decides to hold to the EQ-5D-3L as the metric then, although there are two studies cited there will have to be a systematic review, following ISOOR good practice standards, of UC studies [9]. This review will presumably capture studies that have appeared in the last few years. This is, of course, a futile exercise as the EQ-5D-3L utilities that are uncovered will still fail to meet the required standards of fundamental measurement.

Utilities and QALYs

Creating QALYs from these manifest scores is clearly nonsensical; creating a QALY from an ordinal scale is a mathematically meaningless operation. The resulting QALYs have no meaning. Clearly, attempting to build on these 'numbers' to generate lifetime incremental cost-per-QALY claims is requires a willingness to embrace the 'faith' and 'mysteries' of the ISPOR meme. The implications are obvious: any attempt to generate a QALY, let alone an incremental cost-per-QALY model, makes no sense.

A point that is often overlooked (and the Malinowski and Kawalec fail to make this distinction) is that the generic utility measure typically embodies community not patient preferences. In the case of the EQ-5D, the most widely used measure, utility is not the patient's assessment of value. It is the communities assessment; their valuation (community preference weights) for the response levels to each of the five symptoms reported by patients. The patient does not report a utility score as stated by Malinowski and Kawalec. It is a third party valuation from people the overwhelming majority of whom will have had no experience of UC. In consequence, the value that a patient in a disease state attaches to their health state is ignored.

If ICER is interested in how patients with disease 'value' the stage of disease then rather giving a misleading impression that the generic metric reflects the patients' utility, a disease specific measure should be proposed; one that meets the standards of fundamental measurement. An obvious candidate here is a needs-fulfillment measure which captures a single latent QoL construct, meeting Rasch measurement standards [10].

However, if we restrict our use of utility scores to their 'manifest' or ordinal; status, the analysis comes to some not surprising conclusions in respect of UC:

- The more severe the disease the lower the utility score reported by patients (incorrect; responses valued by community preferences)
- Utility varies between different disease states (as valued by the community) driven primarily by state of the disease and its severity (lowest for severe disease and highest for remission)

Any subsequent analysis of the generic manifest score utilities should be ignored. An ordinal scale allows only medians and modes to be presented; any other operation (e.g., differences between utility scores, bootstrapping, meta-regression, confidence intervals) is meaningless as the scale has neither ratio (true zero) nor even interval properties.

One reason for presenting this assessment of the UC modelling analysis plan is simply to point out that assuming these generic utilities to have the required measurement properties for QALY claims is a waste of time. This means that the proposed ICER lifetime incremental cost-utility model is also redundant. Ideally, we can dispense with QALYs and focus on the needs of patients under competing therapy interventions. ICER's only option would be to assign arbitrary utilities to these four UC disease stages, assigning by assumption measurement properties. In other words, just one more assumption among many to drive imaginary world recommendations.

Conclusion: Imaginary vs. Real World Evidence

Abandoning imaginary generic utilities to create imaginary worlds, opens up a more fruitful line of enquiry: abandoning ICER's commitment creating evidence from imaginary and nonsensical simulacra to a focus on real world evidence. ICER is not interested in hypothesis testing and the discovery of new facts; rather, existing facts are recycled as assumptions to create imaginary worlds. The proposed modeling analysis plan for UC, which should be smothered at birth, is a classic example of the recycling of assumptions and the creation of nonsensical recommendations.

If a manufacturer is to abandon the imaginary worlds of ICER in favor of real world evidence in the tradition of normal science then the first step is to commit to product claims that are credible, evaluable and replicable. The second step is to propose how those claims are to be evaluated. This opens up the opportunities to both assess evidence from 'big data' but, and more importantly for the quality, consistency and tracking of outcomes, a commitment to a product registry. This can be developed from scratch or the required evidence modules piggy-backed onto an existing registry. The third step is to review how the QoL of patients, their response to therapy, is to be assessed. This opens the door to patient-centric instruments with fundamental measurement properties to report on the needs of patients and how competing therapies meet those needs. Finally, the manufacturer has the task to wean decision makers off the ICER press release. Few decision makers appear to want to address fundamental issues of scientific method in challenging the ICER model. This will not be easy. Unless it is accomplished we face a future where decisions continue to be based on imaginary evidence.

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] Langley PC. Nonsense on Stilts – Part 1: The ICER 2020-2023 Value Assessment Framework for Constructing Imaginary Worlds. *InovPharm*. 2020;11(1):No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444

[2] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice. *J Rehabil Med*. 2012.44:97-98

[3] Wootton D. The Invention of Science: A new history of the scientific revolution. New York: Harper Collins, 2015.

[4] Popper KR., The logic of scientific discovery. New York: Harper, 1959.

[5] Lakatos I, Musgrave A (eds.). Criticism and the growth of knowledge. Cambridge: University Press, 1970.

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

[7] Magee B. Popper. London; Fontana, 1973

[8] Malinowski K, Kawalec P. Health utility of patients with Crohn's disease and ulcerative colitis: a systematic review and meta-analysis. *Expt Rev Pharmacoecon Outcomes Res*. 2016;16(4):441-53

[9] Brazier J, Ara R, Azzabi 1 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

[10] Bond T, Fox C. Applying the Rasch Model. 3rd Ed. New York: Routledge, 2015