More Unnecessary Imaginary Worlds - Part 1: The Institute for Clinical and Economic Review's Evidence Report on Janus Kinase (JAK) Inhibitors in Rheumatoid Arthritis

Paul C Langley, PhD
College of Pharmacy University of Minnesota

Abstract

Previous commentaries in the Formulary Evaluation section of INNOVATIONS in Pharmacy have pointed to the lack of credibility in modeled claims for cost-effectiveness and associated recommendations for pricing by the Institute for Clinical and Economic Review (ICER). The principal objection to ICER reports has been that their modeled claims fail the standards of normal science: they are best seen as pseudoscience. The purpose of this latest commentary is to consider the recently released ICER evidence report for Janus Kinase (JAK) Inhibitors. As ICER continues, in the case of JAK Inhibitors, to apply its modeled cost utility framework with consequent recommendations for pricing adjustments, these recommendations also lack credibility. In contrast with previous ICER evidence reports, the present report adopts only a 12-month timeframe, one due, in large part, to ICER being unable to justify assumptions to drive its construction of imaginary worlds beyond 12 months. This commentary emphasizes again, why the ICER methodology fails to meet the standards of normal science. Claims made by ICER for the competing JAK Inhibitor therapies lack credibility, are impossible to evaluate, let alone replicate across treatment settings. Even so, it is important to examine a number of key elements in the ICER invention of the 12-month JAK Inhibitor imaginary world. While this does not imply any degree of acceptance of the ICER methodology, one element that merits particular attention is the failure of the ICER modeling to meet logically defensible measurement standards in its application of generic health related quality of life (HRQoL) ordinal metrics to create its QALY claims. The failure to meet the required standards of fundamental measurement means that the cost-per-QALY claims are invalid. This raises the issue of the application of Rasch Measurement Theory (RMT) in instrument development and the potential role of patient centric outcome (PCO) instruments that represent the patient voice in value claims. The case made here is that the ICER approach should be abandoned as an unnecessary distraction. If we are to meet standards for the discovery of new facts in therapy response then our focus must be on proposing credible, evaluable and replicable claims within disease states. Instruments, such as the Rheumatoid Arthritis Quality of Life (RAQoL) questionnaire that build on the common construct that QoL is the extent to which human needs are fulfilled should be the basis for value claims. HRQoL Instruments that are clinically focused and reflect the value calculus of providers and not patients in measuring response by symptoms and activity limitations are irrelevant.   This puts to one side the belief that incremental cost-per-QALY models, the construction of imaginary worlds are, in any sense, a 'gold standard'; a meme embraced by the health technology assessment profession. Claims for incremental cost per QALY outcomes and recommendations for pricing and access driven by willingness to pay thresholds are irrelevant to formulary decisions.

Keywords: Rheumatoid Arthritis (RA), Janus Kinase (JAK) Inhibitors, ICER pseudoscience, unnecessary distraction, patient voice, Rasch Measurement Theory (RMT)

Introduction

The construction of assumption driven imaginary worlds to support incremental cost-per-quality adjusted life year (QALY) claims for pricing and access recommendations is the hallmark of the Institute for Clinical and Economic Review's (ICER) business model. ICER's latest evidence report on Janus Kinase (JAK) inhibitors in rheumatoid arthritis (RA) follows this model. The JAK inhibitors under review are upadacitinib (RINVOQ, AbbVie), tofacitinib (Xeljanz, Pfizer) and baricitinib (Olumiant, Lilly). In each case the modeled JAK inhibitor and adalimumab were compared as add on therapies to conventional DMARD therapy in the targeted immune modulator (TIM) treatment arms. First released on 26 September 2019, the ICER evidence report for JAJK inhibitors was subsequently withdrawn with a revised report released on 11 October 2019, followed by a further evidence report for review by the ICER convened California Technology Assessment Forum on 26 November 2019 [1 2 3]. The final evidence report was released on 9 January, 2020 [4]. At the same time, this commentary also considers a complementary model framework, the IVI-RA which is an on-line open source imaginary model that has been proposed to evaluate therapy sequences in similar RA populations to those considered by ICER [5 6 7].

The present commentary is concerned with the final evidence report and the modeled economic evaluation. The purpose of this commentary is to point out that the ICER model and consequent recommendations for pricing and access to JAK inhibitors fail to meet the standards of normal science. .A similar conclusion applies to the IVI-RA model. They are irrelevant to formulary decisions.

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

DOI: https://doi.org/10.24926/iip.v11i1.2402

Previous commentaries have pointed out, including a previous commentary on targeted immune modifiers (TIMs) in RA that if an imaginary incremental cost per QALY model is constructed, then any number of similar models, with the same fatal flaws, can be constructed [8] [9] [10] [11] [12]. These commentaries made the case that applying the ICER methodology is an intellectual and analytical dead-end; none of the claims made for comparative cost-effectiveness are credible, evaluable and replicable. As such, formulary committees would have no idea whether ICER recommendations were right or wrong; they would never know and were never expected to know.

The arguments against the ICER evidence report for JAK inhibitors is an exemplar of the irrelevance of a reference case methodology to support recommendations for pricing and access for any pharmaceutical product or devices. Certainly, the reference case methodology is seen as the 'state of the art' in health technology assessment which supports the construction of imaginary, simulated models projecting over the lifetime of a hypothetical patient cohort to generate incremental cost-per-QALY claims. These claims are set against willingness to pay thresholds to convince an audience, who are typically non-technical, to take at face value recommendations for product pricing and access based on a hypothetical world. It is acknowledged by technology assessment groups that these are artificial (yet 'realistic') but that their redeeming feature, apparently, is that they generate 'approximate information' for decision makers; or, more precisely, 'imaginary' information (or disinformation)[13].

The IVI-RA model suffers from the same lack of scientific merit at the ICER JAK inhibitor model. It is, once again, best seen as pseudoscience; it fails the demarcation test. It opens up the prospect for a multiverse of more complex imaginary worlds to support the apparent need by decision makers for more hypothetical and 'approximate information' on an unknown future.

There are more fundamental flaws. ICER claims for product value do not reflect the interests of patients in RA. The reference case assumes that a generic measure of health related quality of life (HRQoL) is appropriate to evaluating the benefits of competing therapies. This belief has been challenged repeatedly over the past 40 years by first, the needs based framework for evaluating quality of life (QoL) and, more recently, since the late 1990s, supplementing the needs approach with the application of Rash Measurement Theory (RMT) to ensure unidimensionality with interval scoring [14] [15].

As detailed below, a generic multi-attribute instrument such as the EQ-5D-3L, the backbone of the ICER cost-per-QALY imaginary worlds, fails standards for fundamental measurement. Rather than providing a unidimensional metric that supports RMT, it generates an ordinal manifest score that precludes basic arithmetic operations. This is seen, for example, if wanted to assess effect size. It would be a logically invalid measure [16] [17]. Despite the demonstrated disease-area specific superiority of needs base QoL instruments, ICER persists in modeling the EQ-5D-3L which, at best is a limited health related quality of life (HRQoL) instrument without a clearly defined construct. As argued here, HRQoL instruments, despite their inability to provide more than manifest ordinal scores, are not relevant if our objective is to assess the QoL impact of therapy interventions. They are clinically focused, representing the interest of physicians and not patients. They take no account of the needs fulfillment of patients with RA and the extent to which competing therapies impact the lives and needs of the patient [18]. This points to the relevance, not of a generic instrument but ones that are disease specific and patient centric [19] [20]. If we want to assess patient needs, then we don't need QALYs.

Unfortunately, as we also argue below in the case of RA, ICER has nowhere to go. If the case is made, which it has been in the literature for some 20 years and more, that ordinal measures have to be abandoned, then the ICER cost-per-QALY reference case collapses. It follows that the ICER business model also collapses; apart from the nonsense of attempting to construction imaginary lifetime reference case worlds. It is, perhaps, surprising that ICER launched itself as NICE-lite (National Institute for Health and Clinical Excellence) when the focus in QoL is away from HRQoL and towards needs based, RMT consistent, disease specific outcomes instruments. Indeed, as we also note, the debates over alternative value assessment frameworks come down to a failure to take explicit account of needs-based assessments instead of trying to 'bolt-on' other criteria to a core HRQoL model.

The ICER reference case model, to include the IVI-RA, is an unnecessary distraction in health system decision making. If formulary committees and insurers are considering factoring in the ICER recommendations as 'approximate information' to support pricing and access, they should put such claims to one side. Any formulary decision must be evidence based where the evaluation techniques meet the standards of normal science. Decisions should not be based on imaginary worlds, as attractive and 'probably realistic' they may be to the believer. Admitting that the focus is not on testing hypotheses for product impact and cost-effectiveness but on providing imaginary 'approximate information' is not an admissible defense.

The Intelligent Design Meme

Unfortunately, ICER is not alone in its promotion of intelligent design in health technology assessment. This is the position of the leading health technology assessment group in the US, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Creating imaginary worlds to support imaginary claims is, of course, an easy way out. It puts to one side the discovery of new facts in favor of fabrication. The need to fabricate is defended on the grounds that we lack the required information for hypothesis testing which then forces

DOI: https://doi.org/10.24926/iip.v11i1.2402

us to the reference case imaginary worlds; this is an absurd position. Rather than taking refuge in imaginary worlds our focus must be on the discovery of new facts to support formulary decisions that meet both the needs of patients as well as those of the health system.

Unfortunately, the intelligent design health technology assessment meme, as a unit of cultural transmission, is well established. It will take a concerted effort, not only in therapy assessments for target RA populations, to abandon this reference case dogma, but to put resources into ensuring in both randomized clinical trial (RCT) protocols and claim assessment that needs based instruments are front and center.

A final point: 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 imaginary constructed models. This is underscored in the case of RA by the need to construct utility manifest scores, by crosswalking from clinical markers. As few clinical trials capture specific quality of life metrics utilizing generic preference-based multi-attribute instruments, ICER is faced with the need to construct utility values to generate QALY estimates. This adds a further element of unreality to the ICER model as there are a substantial number of generic utility value options open that can be emulated as well as a smorgasbord of competing techniques and algorithms available for crosswalking to create utility values and the range of potential crosswalking techniques. In short, while ICER may claim pole position in imaginary evidence constructions in the US, 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 [21] [22] [23]. Fortunately, we do not have to go to the lengths of comparing imaginary claims. We can simply put all imaginary constructs to one side.

Initial Scoping and Response

The original ICER scoping document was characterized by ICER as a condition update to the 2017 ICER report on targeted immune modifiers (TIMs). Public responses from stakeholders was requested by 1 May 2019 and a revised scoping document issued on 9 May, 2019 [24]. The earlier report presented an imaginary modeled assessment of the comparative clinical effectiveness and value of multiple TIMs for moderately to severe active rheumatoid arthritis both as monotherapy and in combination with conventional DMARDS [25]. The 2017 evidence reported was reviewed by the present author in an early commentary [4]. The commentary concluded: *Rather than attempting to inform decision makes through the construction of imaginary worlds, price negotiations should be predicated on evidence that meets the standards of normal science. Unless evaluable and replicable claims are presented by ICER to support recommendations for price discounting, the recommendations should be rejected. This applies not only to the current recommendations for price discounting of TIMs, but to other ICER evidence reviews that have generated non-evaluable claims.*

Initial stakeholder response to the ICER scoping document raised a number of concerns regarding issues to be addressed and the measurement of QoL; whether the focus should be on the clinical concept of health related quality of life (HRQoL) or the wider concept of quality of life (QoL)[**26**]. The Arthritis Foundation, to give one example, noted in its submission that in response to the final evidence report for the 2017 assessment they were concerned that: *'the study analysis was narrow and did not include a representative sample of people with RA, and therefore was not relevant to all people with RA; the conclusions reached were based on inadequate performance measures; the reliance on QALYs was inappropriate for this disease population; and there was an absence of real-world evidence for this disease population; and patient experience in the final analysis'* [27]. As an exemplar, the Arthritis Foundation response points to a number of key issues: (i) the real world use of ICER reviews where '…many concerns remain about the core methodologies and their applicability to chronic disease states, particularly RA'; (ii) the application of point estimate averages of the value of treatments (that) become potentially very harmful, especially if these results are taken at face value and applied as umbrella statements on relative value across the entire population; (iii) avoiding heterogeneity where the misuse of a single cost-effectiveness ratio 'as a proxy for value for all potential patients … is reduced access to a therapy for individuals for which that therapy would provide significant value if delivered in a timely fashion' and (iv) the importance of a longitudinal focus in assessing the impact of RA.

Narrowing the Focus: JAK Inhibitors

In June 2019, ICER refocused its RA review and issued an updated scoping document [28]. ICER decided to narrow the focus of its RA review on JAK inhibitors for two reasons: (i) the approval of baricitinib since the 2017 report for moderate-to-severe RA and (ii) the expected approval in August 2019 of upadacitinib following the FDA acceptance for priority review.

This revised review was to assess the comparative clinical effectiveness and value of JAK inhibitors for moderately to severely active RA, both as monotherapy and in combination with conventional DMARDS. The proposal was to update previous assessments with new trial data for JAK inhibitors together with a review of the clinical and economic evidence for infliximab-dvvb (Inflectra, Pfizer).

The JAK Evidence Reports

In the first evidence report presented by ICER (26 September, 2019) for the JKAK inhibitors, the model took a base-case lifetime perspective. Patients from a hypothetical initial cohort remained in the model until death. All patients could transition to death from all causes and from RA-related mortality The selected model generated constructed claims for outcomes to

DOI: https://doi.org/10.24926/iip.v11i1.2402

include imaginary lifetime costs, life years (LYs), quality adjusted life years (QALYs) and equal value of life years gained (evLYG). The incremental cost per evLYG was proposed to complement the cost per QALY calculations and provide policymakers with a broader view of cost effectiveness. The model was constructed in hēRo3℠, with some components of the model, such as survival distributions, developed in RStudio (Version 1.1.463). The hēRo3 model is a Web-based, health economic modeling platform that supports the development of imaginary technology assessment worlds with both Markov cohort and partitioned survival models (Policy Analysis Inc., Brookline, MA).

The September evidence report was rescinded within a few days and a revised evidence report released on 11 October 2019). ICER maintained, that following a review that suggested 'some of the assumptions and calculations' might be re-evaluated in the modeling apparently 'to align … with how patients transition between these therapies in the real world'. The key changes in assumption were (i) to model how those who did not respond adequately to first line therapy would transition to a basket of targeted products and not palliative care and (ii) that cost-per-QALY claims were evaluated over one year and not a lifetime. This shorter time frame was selected as base-case because of uncertainties over the clinical differentiation of the target JAK therapies over time (i.e., insufficient data to support assumptions). The point to note is that the evidence base had not changed. ICER continued to point out in the overview of its assessment of long-term cost effectiveness that while the modification of their initial objective to assess the relative value of JAK inhibitors versus adalimumab for treatment after failure by a conventional DMARD was still the focus, they were unable, even with assumptions designed to create an imaginary world, to model a direct comparison of tofacitinib to adalimumab due to inadequate data in the TIM-naïve or TIM-experienced population. The same limitation applied in a comparison of upadacitinib to adalimumab in the TIM-experienced population. In the case of baricitinib in patients who failed TNF-inhibitors, modeling attempts to compare it to adalimumab in the TIM-experienced population was considered impossible due to a lack of comparable data.

However, as ICER notes as a rider to their evidence reports that when 'new' data emerge they may revisit the imaginary model and produce a new imaginary model report with value judgments. It is not clear how manufacturers, insurers and formulary committees would respond if they see the ICER model as one of many future imaginary iterations, none of which meet accepted standards for normal science. The result would be (i) a possible multiverse of base-line models and (ii) a budding of future models with a multiverse of imaginary 'modified' models created from each initial base-line model, each claimed to provide necessary 'approximate information'; a daunting prospect for decision makers trying to separate out one from many worlds.

Meeting the Standards of Normal Science

The requirement for testable hypotheses in the evaluation and provisional acceptance of claims made for 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. Indeed, as early as the 16th century Leonardo da Vinci (1452 – 1519) in notes that appeared posthumously in 1540 for his *Treatise on Painting* (published in 1641) clearly anticipated the standards for the scientific method which were widely embraced a century later in rejecting thought experiments that fail the test of experience. 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) [29].

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 [30] [31] . Hypotheses or claims must be capable of falsification; indeed, they should be framed in such a way that makes falsification likely. Life becomes more interesting if claims are falsified because this forces us to reconsider our models and the assumptions built into those models. This leads to the obvious point that claims or models should not be judged on the realism or reasonableness of assumptions or on whether the model 'represents' for a public advocacy research group such as ICER their perception of a future, yet unknown, reality.

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 [32]. 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. Indeed, paradoxically, while the development of pharmaceutical products and the evidence standards required by the Food and Drug Administration (FDA) for product evaluation and marketing approval is driven by adherence to the scientific method, once a product is launched and claims made for cost-effectiveness and, in the case of ICER, modeled pricing and access recommendations, the scientific method is put to one side. Pseudoscience succeeds science.

The rejection of a research program that meets the standards of normal science by groups such as ICER is best exemplified by the latest version of the Canadian health technology guidelines where it is stated: *Economic evaluations are designed to inform decisions. As such, they are distinct from conventional research activities, which are designed to test hypotheses* [33]. While this

DOI: https://doi.org/10.24926/iip.v11i1.2402

position puts modeled health technology assessment in the category of pseudoscience, it is also what may be described as a relativist position. Rather than subscribing to the position that the standards of normal science are the only standards to apply in health care decisions and value claims, the relativist believes that all perspectives are equally valid. Health care decisions are to be understood sociologically. No one body of evidence is superior to another. Results of a lifetime modeled simulation are on an equal basis with those of a pivotal Phase 3 randomized clinical trial. For the relativist, the success of a scientific research program, in this case one built on hypothetical models and simulations, rests not on its ability to generate new knowledge but on its ability to mobilize the support of the community. Basing decisions on models and simulations underpins the consensus view that evidence is constructed, never discovered. Instead of coming to grips with reality, science is about rhetoric, persuasion and authority [29]. Truth is consensus.

How is this consensus maintained? The ISPOR consensus, embraced by ICER, on health technology assessment has been characterized in previous commentaries as a meme [34]. This is deliberate, as it underpins the interpretation of ICER's continued unqualified acceptance of the reference case as its core business model, as a sociological phenomenon. The ICER reference case which constructs evidence to support pricing and affordability pronouncements, can be characterized as the adoption of a unit of cultural transmission or unit of imitation; as an analog of gene pool propagation 'by leaping from body to body via sperm or eggs' [35]. Human beings are good at imitation. The reference case meme appears to be adept in its infectivity, supported by an organizational infrastructure to defend it against competing views, ensuring survival through supporting propagation, longevity, fecundity (or acceptability) and copying fidelity. The spectacular adoption and propagation of this meme in seen in the health technology literature over the past 35 years with literally thousands of imaginary world technology assessments. Characteristically, few present claims that might be evaluated in, for example, the short term. Rather we are asked to believe in entirely imaginary constructs to drive formulary decisions. ISPOR is quite clear in its support for the reference case imaginary health technology meme: *Leaders in the field of economic evaluation in health care have long recommended that analysts <u>seeking to inform</u> resource allocation decisions <u>approximate the value</u> of interventions in terms of incremental cost per QALY gained* (emphasis added) [13]. It's not clear, with the imaginary constructs, how we might distinguish 'approximate information' from 'approximate disinformation'. Is there some behavioral asymmetry in 'losses' 'vs gains' that would have to be factored into such an assessment by a formulary committee?

Models and Assumptions

It is accepted that knowledge is provisional and permanently so. This stems from the obvious point that we can at no stage prove that what we 'know' is true. Attempting to believe or justify our belief in a theory is logically impossible. What we can do, by empirical assessment, is to try and demonstrate our preference for one theory over another (and apply it to the best of our knowledge).

Constructing imaginary worlds which were never intended to generate potentially falsifiable claims cannot, therefore, be defended by an appeal to the 'truth' of their assumptions. If a health technology assessment claim is built upon a series of assumptions, a reasonable question is to ask what is the status of the various assumptions. Are they to be viewed as 'reasonable or 'realistic' metrics for an unknown future reality? Have they been selected from the literature because they seem appropriate? Are they the 'best available' from limited data? In the case of JAK inhibitors, the one-year model reflects the absence of data on long-term comparisons. ICER cannot defend a longer timeframe because it could not justify the assumptions required.

More to the point, there is a belief that the fact that the selected assumptions are based, where feasible, on an empirical study validates the choice of assumption (e.g., network meta-analysis). For example, if the model is intended to incorporate utilities that have been reported in one or two studies (usually as few as that) for progression and time spent in the stages of a disease, then there is an immediate methodological issue. To claim that an assumption is valid is to revisit Hume's induction problem (David Hume 1711-1776): an appeal to facts to support a scientific statement. Unfortunately, as Hume pointed out, no number of singular observations can logically entail an unrestricted general statement. Certainly, there may be comfort in reporting that 'so far' the claim that all swans are white has not been contradicted (until that Qantas vacation in Western Australia) so that one fully expects the next swan to be white. But as Hume pointed out, this is a fact of psychology and does not entail any general statement. From a utility perspective, the fact that one hundred papers have agreed (within limited bounds) generic utilities from the same instrument for a target population in a disease state stage is immaterial. We cannot secure this assumption: 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*' [36]. Claims, for the relevance of a constructed imaginary world built on the assumption that the model elements have been validated by observation is simply nonsensical.

Despite ICER's continued embrace, logical positivism is dead. It died some 80 years ago. All knowledge is provisional. Popper's contribution was to make clear that Hume's problem with induction can be resolved. We cannot prove the truth of a theory, or justify our belief in a theory or attendant assumptions, since this is to attempt the logically impossible. We can only justify our preference for a theory by continued evaluation and replication of claims. Constructing imaginary worlds, even if the justification is that they are 'for information' is, to use Bentham's (Jeremy Bentham 1748-1832) memorable

DOI: https://doi.org/10.24926/iip.v11i1.2402

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.

The ICER Reference Case

Central to the ICER construction of imaginary value claims and the potential for many worlds is the reference case. Standards for model building, the construction of imaginary worlds, are clearly stated with the preference for imaginary frameworks that take a long-term or lifetime perspective. Value propositions are to be in imaginary cost-per QALY terms. Once an imaginary cost-per-QALY estimate (or estimates under different scenarios) has been constructed, the acceptability of a proposed product price is then assessed against cost-per-QALY willingness to pay thresholds (typically $50,000, $100,000 and $150,000 per QALY with exceptions for higher cut-offs for rare diseases). Whether a product ‘adds value’ is then determined in terms of its impact on an imaginary estimated lifetime modeled QALYs set against a proposed lifetime product cost where both are driven by constructed evidence.

Over the past 30 years, literally thousands of imaginary modeled claims have been presented in the literature, including leading health technology assessment journals. Annual reviews of the status of cost-effectiveness or modeled claims in the three journals, *Pharmacoeconomics*, *Value in Health* and the *Journal of Medical Economics* found that the majority of models presented non-evaluable claims (typically lifetime cost-per-QALY) [37] [38] [39] [40] [41]. Where models were funded by a manufacturer a high proportion supported, in their modeled cost-per-QALY assessment, the manufacturer’s product. All too many of the papers were essentially marketing exercises [42] [43].

Despite the argument that the ICER reference case lacks credibility; that it is pseudoscience rather than science, there is little doubt that ICER will persevere. After all, if ISPOR persists in supporting, after a recent extensive membership review, the construction of imaginary incremental lifetime cost per QALY models to support formulary decisions (together with a gaggle of imaginary scenario offshoots of a base case model), ICER can continue to argue that it represents the ‘state of the art’ meme in presenting evidence reports from imaginary constructs.

Manifest Score: Utilities and QALYs

While apparently overlooked by ICER, ISPOR has published good practice guidelines for model builders to identify, review and apply health state utilities in cost-effectiveness models. Minimum reporting standards are proposed to judge the appropriateness of the utility metric selected. This good practice report was released in March 2019, which would have given ICER ample time to undertake such a review [44]. As it stands, in the brief account given by ICER of the basis for its choice of health status utility metric for the JAK modeling EQ-5D metrics, there is no mention of going beyond the material used in the earlier 2017 RA report.

As an aside and, perhaps unsurprisingly, the ISPOR health state utilities good practice report did not raise the issue of fundamental measurement standards. Utilities or ordinal manifest scores were taken at face value and, for all intents and purposes, there was an implicit assumption that they had cardinal scales. Again, truth is apparently consensus.

Too little attention is given in critiques by stakeholders and others of the ‘evidence’ to support model assumptions. In the latest evidence report, the key assumptions are listed in Table 4.4 together with their rationale. While it is possible to apply sensitivity analyses to overcome specific uncertainties in assumptions (e.g., magnitude of HAQ rebound), in a number of cases ICER admits to an absence of robust published evidence (e.g., mapping of DAS28 to HAQ for the modeled treatment strategies; discontinuation rates) with a fallback on ‘expert opinion’. Even so, the point remains that the model is essentially a series of assumptions. As the revised modeling results for the JAK inhibitors makes clear, changing assumptions can change recommendations so that the prospect is for a series of modeled revisions as the ‘evidence base’ and ICER’s choice of assumption changes. At the same time, given the malleability of assumptions, it is of interest to note the ‘false’ precision with which the modeled results are presented. In the case of upadacitinib and adalimumab, ICER notes that ‘we found no difference in Lys (life-years) gained up to the fourth decimal place’ (pg. 55).

Previous commentaries in this series have raised the concern that an unqualified use of the term QALY may give decision makers the impression that there is a common imaginary QALY standard that has been agreed to in health technology assessment [45]. This is far from the case. ICER uses the term, almost indiscriminately, without qualifying its claims that the utility metric driving the QALY estimate is based on an often arbitrary choice of measure, typically involving a specific definition of HRQoL (defined by choice of health symptoms or dimensions and response level) rather than a broader concept of quality of life (QoL). If the intent is to mandate a specific generic utility metric, as recommended in the ICER reference case, then for US preference measures there are a number of options: EQ-5D-3L, EQ-5D-5L, SF-36, SF-12, SF-6D. Confusion can arise when, as in the RA evidence report, and in the earlier evidence report on TIMs, the ICER refers to the EQ-5D, without qualifying whether it is the 3-level or 5-level variant. Reviewing source documents referenced by ICER points to the 3-level variant. NICE in the UK is still struggling with the use of the 3-level as opposed to the 5L-level. While the 5-level is considered appropriate as a descriptive profile, there are ongoing concerns with the valuation of the 5-level [46].

A point to note, however, is not just the limited number of responses open to patients within the health dimensions captured in the generic measure, but the limited ambit of those measures. Are these measures appropriate if the intent is to represent health related quality of life (HRQoL) in RA? The EQ-

DOI: https://doi.org/10.24926/iip.v11i1.2402

5D, for example, is based on five broad health dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Patients respond: no problems, some problems and major problems. The SF-6D, by contrast, has six health dimensions (with levels of response in parentheses): physical functioning (6), role limitations (4), social functioning (5), pain (6), mental health (5) and vitality (4).

Without going into details of each of the various preference-based multi-attribute health status systems it is worth noting that the decision (if you insist on ordinal manifest scores) does matter, as the systems are far from identical. They differ in their coverage of health dimensions, in the defined levels, the description of these levels, the severity of the most severe level, the populations surveyed, the instruments used to determine the preference scoring and the theoretical approach for modeling the preference data into a scoring formula [47]. The same patients can have quite different scores depending on choice of instrument. But, of course our intent is to provide only 'approximate information'. To which might be added: none meet the required axioms of fundamental measurement.

Interval Scales

A further reason for describing the ICER reference case as an analytical dead end, rather than a path to discovery, is the failure to appreciate the importance of meeting the general axioms of measurement: invariance of comparisons and sufficiency [48]. If these axioms are met then the result is an interval or cardinal scale. It is worth remembering Steven's key observation in his seminal 1946 paper: 'With the interval scale we come to a form that is quantitative in the ordinary sense of the word' [49] .

Unfortunately, interval scores are not commonly found in the construction of imaginary worlds. Generic multi-attribute measures generate ordinal scores; the majority of patient reported (PRO) instruments generate ordinal scales. Indeed, it is not clear as to the extent, if any, that those developing generic multi-attribute instruments and clinically focused HRQoL instruments considered a latent construct and how this might guide instrument development to meet interval scoring or even ratio standards.

General measurement theory recognizes four scales: nominal, ordinal, cardinal or interval and ratio. With ordinal scales the ranked response within an item 'order' is known but the differences between the responses is not. We may define a mode or median but not a mean. Cardinal or interval scales are where the order is known as well as the difference between responses on the scale. Interval scales open up a range of statistical options: means and standard deviations can be computed, change and effect sizes can be presented. However, with cardinal scales there is no 'true' zero. Ratios can't be computed; there is no multiplication or division. It is only with a ratio scale, one with a true zero, that these are possible. As the QALY is a ratio, then a reasonable question is whether the 'utility' measure (the denominator) should not only have cardinal properties but also the property of a true zero.

Unless ICER can demonstrate that its choice of utility score and consequent QALY claims meet the axioms of fundamental measurement, then over and above the imaginary nature of the model, the cost-per-QALY claim lacks credibility in terms of fundamental measurement. Decision makers might accept the notion of 'approximate information'; it is unlikely they would accept measures within the 'approximate information package' that lacked mathematical credibility.

Patient Centric Outcomes

PRO instruments such as the EQ-5D-3L support the more narrow and clinical concept of HRQoL. Typically, HRQoL measures do not capture the patient voice. They capture, as in the case of the EQ-5D, the interest of the treating physician in assigning health symptoms or dimensions with ordinal response levels within dimension. This is combined with time spent in health state clinical stages to create QALYs. Patient preferences are absent; the preferences are those of a community sample where only a handful of respondents will have had any experience of specific disease states, their natural history and impact. The QALY measure is, of course, meaningless.

If we consider the perspective of the patient, a more appropriate framework is to put HRQoL to one side and consider the QoL of the patient. The question we might consider is whether the patient's needs are fulfilled. The needs model, developed in the 1990s*, hypothesizes that the value of individual lives is dependent on the extent to which their human needs are fulfilled. Value is low when few needs are met* [14 18]*.* Certainly, the impact of treatments within a disease state can be captured by health symptoms and response defined within a HRQoL model, but these are only sufficient as an index of the clinical or operational impact of interventions as reported by the patient or perceived by the treating physician for a selected symptom set. The functional status of a patient is impacted by more than symptom resolution. There is a range of possible other influences such as social support, family assistance finances, and education. HRQoL raw scores may 'improve' while failing to meet the needs of the patient [18]. At the same time, providing closer social support may improve needs fulfillment but may not register as a HRQoL improvement.

If we are to develop a patient-centric outome (PCO) needs-based measure, then we have to begin with the patient. This is achieved by a 'bottom-up' commitment to *qualitative interviews with patients to ask how the patient's life has been affected by the disease in question, and probe how limitations of functioning affect the interviewees* [14 18]. This is the first step to providing, within a Rasch modeling framework, statements regarding needs fulfilment, issues common to the target patient group and ultimately the final item set.

DOI: https://doi.org/10.24926/iip.v11i1.2402

If a manufacturer is to make a needs-based case for its product it is not enough to focus on a 'top down' HRQoL instrument. Assessments by third parties such as ICER to assess the 'value' of the intervention miss the point. It is the ability to measure response defined in needs fulfillment terns, not attempts to draw inferences from a partial and possibly misleading, clinically focused HRQoL measure that is claimed to support the value case. This does not mean that it is not possible to construct a generic needs measure, such as the Nottingham Health Profile developed in the late 1970s [50]. Rather, our focus should be on disease specific patient centric outcome (PCO) measures if we wish to evaluate therapy response claims .

Rasch Measurement Theory

The first step in the development of an instrument, a PCO rather than a PRO instrument, is agreement on a unifying construct or attribute set. A latent trait or construct is unobservable; what our primary task should be is the construction of a PCO that meets Rasch standards. Given the construct, responses to locally independent items are observable manifestations of those internal but unobservable attributes. If our construct is needs-based QoL, as the measure of benefit to the patient of a new therapy, then the presence and amount of the latent trait has to be operationalized, or inferred, from item responses. To illustrate this point, Bond and Cox give the example of temperature [17] . As a latent construct, temperature cannot be observed directly, we can only assess the effects of temperature changes on certain classes of objects (e.g., the volume of mercury in a carefully calibrated and constructed space). The thermometer is, importantly, unidimensional and we read off temperature by a calibrated scale that has cardinal or interval invariance properties. This is what we must achieve in the assessment of QoL as a needs construct in therapy impact; not an ordinal HRQoL scale that lacks a coherent construct. It is important to note also that while the term 'measure' is often applied to the Rasch construct, the score is actually counts of discrete observations. The raw score is the sufficient statistic for that person in quantifying QoL.

Fundamental measurement is central to the Rasch model. As first conceptualized, the object was to apply the standards of measurement common to the physical sciences to construct fundamental measures: interval level measurement. The Rasch model is confirmatory; it requires the data to fit the model [17]. If this is achieved, then we can claim that the instrument conforms to a scale with invariant, interval measurement properties generating a single score. Our role, therefore, is to identify a set of items that satisfy the Rasch model. Retention and exclusion of items is a key issue, together with the application of the tools of classical test theory (CTT) to assess face and content validity together with assessments of the reliability and validity of the final item set.

While, as noted, Rasch fits the data to the model; conventional modeling fits the model to the data. In the case of EQ-5D utility scores, these are generated by an algorithm that is designed to ensure preference weighted responses to the HRQoL items to fit a utility scale (range 0 = death; 1 = perfect health). This is required to create QALYs (dividing time spent by the utility of that interval). Users implicitly assume in the reference case that the EQ-5D and other generic measures (and for the majority of disease specific patient reported outcomes (PRO) instruments) that the scale has cardinal rather than ordinal properties. This means a response from 0.4 to 0.5 is the same interval as 0.7 to 0.8). This, unfortunately, is not the case: the instruments utilize an ordinal calibration. This occurs because the issue of unidimensionality in instrument development has been ignored. Certainly, the instrument may claim to meet the standards required in classical test theory (CTT) but they fail to provide valid means and standard deviations, requiring non-parametric statistical analysis for evaluation. More formally, we require a mechanism for translating manifest to latent scores: this is achieved by Rasch analysis *that delivers conjoint measurement when the data fit the model* [16 51].

The key features, therefore, of Rasch analysis, which unifies measurement issues required for interval scaling is where a set of questionnaire items are to be summed to provide a total score. This involves testing for:

- *Internal construct validity of a scale for unidimensionality which is required for a raw summed score*
- *Item invariance for interval level scaling*
- *Appropriate category ordering*
- *Differential item functioning* [19 20]

Unidimensionality and Rating Scales

'A basic assumption of rating scales is that their items measure a common underlying construct. If this can be demonstrated then the scale is unidimensional; if not the scale is invalid. This is a critical standard for modeling. Hypotheses must be presented as unidimensional. We cannot assume unidimensionality. Even in the construction of an imaginary ICER-type world, unidimensionality has to be demonstrated not assumed. If value claims are made, then the measures must allow change in scores to be interpretable. The issue with HRQoL 'measures' is that they were not developed from an underlying QoL construct. Their composition (choice of items) reflects the decision of clinicians, with preference determined by the community rather than the patient.

If we accept the need for a coherent construct such as needs based QoL then we have a firm base to guide instrument development and the construction of an instrument that has the required unidimensional properties. There is no common ground with either multi-attribute utility scores or the CTT based PRO measures. Claims that, under certain circumstances and with appropriate minor adjustments we can 'unidimensionalize' such instruments are simply beside the point. We need a common and defensible construct that is applicable across specific disease areas in a representative

DOI: https://doi.org/10.24926/iip.v11i1.2402

target population. In other words, for a unidimensional Rasch scale we require a spread of items ordered for difficulty and a distribution of abilities among respondents to respond to those items. With few exceptions, unidimensionality cannot be inferred *ex post*, it has to be demonstrated for item order and item fit as the instrument is developed.

Rasch analysis has been applied widely in developing instruments for assessing therapy response over the past 30 years. Apparently, ICER (and ISPOR) did not receive the memo. Or, to be possibly more reasonable, they took on faith the mandatory requirement by the National Institute for Health and Care Excellence (NICE) in the UK for the EQ-5D -3L in their reference case requirements. However, RMT is now effectively the *standard for creating and evaluating measurement instruments* [20]*.* Given this, ICER has an interesting (and impossible) task to demonstrate both globally and, in this case, for the JAK inhibitor target population that its selected outcome measures, while not patient centric, meet required instrument standards for fundamental measurement and that this 'measure' and not a patient centric needs based QoL instrument is the appropriate outcomes measure..

Once the requirement for RMT in instrument development is recognized, with the construction of unidimensional cardinal scales (not utilities) ICER in the absence of other supporting evidence is in an awkward position. It can either go ahead and continue with its reference case model ('not perfect but others do it') or propose an alternative framework to meet the objectives of its business model. This is a daunting prospect. ICER will not only have to admit that its previous evidence reports that propose modeled claims for pricing and access, including those for RA, are mistaken (and misleading) but that they represent a methodology that fails to meet the standards of normal science. Put simply: incremental cost per QALY claims and thresholds are redundant. This, of course, is unlikely to happen

Standing by Generic Utilities

If the decision by ICER is to carry on in the sure and certain hope that the many worlds claim will escape close scrutiny in the absence of accepted measurement standards for unidimensionality and interval scoring, then ICER has to take on board ongoing debates for utility scores and the choice of an acceptable instrument. It is incumbent upon ICER to undertake a systematic review of the application of the measures in RA, with particular focus on the target JAK population, to identify relevant studies. It is not acceptable, according the ISPOR good practices to rely, as ICER does, on only one or two studies. Rather, the recommendation is for a synthesis of published health state utilities for a given health state. If this is impossible due to a limited number of studies, the model builder might be well advised not to press forward.

As well as issuing good practice guidelines for health state utilities in imaginary cost-effectiveness models, ISPOR has also released good practice guidelines for mapping from non-preference based outcome measures for those attempting to model health state generic utilities (January 2017) [52]. Again, with sufficient time, it would have been incumbent on ICER to apply these guidelines to support the choice of mapping algorithm, its quality and relevance. This, unfortunately, is not the case. The RA mapping function is presented with little justification for its application, relying on a limited and somewhat dated evidence base. The mapping technique used in the ICER report for JAK inhibitors is identical to that used in the earlier evidence report on Targeted Immune Modulators (Table D6) [53]. The comments made in respect of the former apply equally to the latter, which should give pause as to the merits or otherwise of the earlier ICER recommendations for pricing and product access.

In the absence of a systematic review of health status utility scores and mapping functions for target patient RA populations appropriate to the introduction of JAK inhibitors, it is not clear why ICER persists with the EQ-5D-3L in modeling imaginary worlds. The EQ-5D-3L and EQ-5D-5L are, to all intents and purposes, separate systems. ICER should make quite clear that it is not the EQ-5D-5L utility values that are being estimated but utility values for the EQ-5D-3L. ICER should not think it can use these metrics indiscriminately in the same model nor, as it has done with oral semaglutide use EQ-5D-3L and HUIMk3 utilities in the same model with the rather disingenuous excuse that they give similar scores [54].

ICER, in fact, is in a similar bind to NICE in the UK in the choice of mapping algorithm and the presence of two versions of the EQ-5D. If there is to be a level imaginary playing field for comparing imaginary products value claims over time in a defined target population within a disease state such as RA, or for comparisons between disease states then, ICER must adopt not only a standard mapping algorithm (e.g., mixed method) to create a preference-based score for modeling, but it must also mandate the preference-based outcomes instrument that will be utilized for all evidence report modeling. If these requirements are not met then ICER will face criticism that its willingness to pay thresholds must be re-calibrated for the mapping technique utilized and the choice of target utility values. If two utility metrics are used in the same model then to present the same price discounting recommendation with different cost-per-QALY willingness of pay thresholds calibrated to the two metrics is problematic. The issue becomes even more problematic when different models employ different frameworks for time spent in different disease states and then different estimates of costs.

If ICER decided to mandate the EQ-5D-5L utility values for the reference model to construct imaginary worlds, then it would have to re-calibrate previous studies in order to ensure that the evidence based for QALY claims was consistent. In this respect, two recent studies are worth noting. The first of these studies evaluates the relationship between the EQ-5D-3L and the EQ-

DOI: https://doi.org/10.24926/iip.v11i1.2402

5D-5L to consider what effect this has on cost-effectiveness claims in RA [55]. Applying best-fitting models the authors found that the mapping coefficients for the modelling and the latent factors were significantly different. Overall, the 5L shifts ordinal utility values towards full health and compresses them to a smaller ordinal range so that improvements in quality of life are valued (ranked) less. This results in substantially different estimates of cost effectiveness. In the second evaluation, the focus is on the impact of moving from the EQ-5D-3L to the EQ-5D-5L in NICE appraisals, estimating 5L utility vales from 21 comparisons of interventions [56]. The authors concluded that the two measures of utility value lead to substantially different estimates of incremental QALYs and cost-effectiveness (without any discussion of fundamental measurement axioms).

Of particular interest is a recent study modelling the implications of a switch from the EQ-5D-3L to the EQ-5D-5L not least because it is focused on evaluating drug therapies in rheumatoid arthritis [57] . As the authors point out, the switch to the EQ-5D-5L version of the original instrument was prompted by concerns over the sensitivity of the instrument and floor/ceiling effects; but not to its fundamental measurement properties. Although mapping or crosswalking between instruments (inputs to outputs) is popular, the authors caution against conditional modeling (e.g. a regression model) mapping from one measure to the other even when they are captured in the same data set. Reasons given are (i) utility scores have highly irregular distributions and mapping methods often fit poorly; (ii) use of a single utility score loses the additional information in the dimensions of response; and (iii) the direct comparison approach is necessarily specific to the particular scoring system making it difficult to explore sensitivity to scoring system variations. The approach taken by the authors is 'response mapping' which considers the statistical relationship between the 3L and 5L responses but with the utility score brought in at the final analysis stage.

This last assessment of the impact of alternative mapping functions including valuation algorithms for the two versions of the EQ-5D utilizes the US National Data Bank for Rheumatic Diseases (NDB) which captured both versions of the EQ-5D in the January 2011 wave when the NDB switched from the EQ-5D-3L to the EQ-5D-5L versions. Two features of the analysis are important if we consider the ICER reference case. First, comparing the EQ-5D-5L and the EQ-5D-3L when the two are captured in the same data set and second, the options for mapping from the EQ-5D-3L to a 'revised' utility score based on the EQ-5D-5L. In the former case, there are significant differences between the descriptive systems where the two instruments give 'significantly different pictures of the relationship between individual health states and their demographic and clinical determinants in respect of mobility and pain dimensions'. Mapping between health states before applying utility scores can provide a robust framework for converting old 3L evidence to a 5L basis. Direct mapping between utility scores requires a sufficiently flexible model. Even so, as this conflates the effect of the redesigned 5L health description and the revised utility tariff 'it does not offer a natural way of comparing alternative utility tariffs'. Finally, in a re-examination of a trial of combination RA drug therapies switching from the 3L to the 5L can make 'a substantial difference to the conclusions from cost-effectiveness studies'.

To illustrate the impact of moving from theEQ-5D-3L to the EQ-5D-5L the authors consider cost-effectiveness claims based on the 2-year CARDERA trial. Estimating the EQ-5D-3L directly from the trial data for methotrexate (MTX) as monotherapy yields an estimated total QALYs of 1.24 with estimated incremental cost-effectiveness compared to MTX+ciclosporin (CS) of £4,648 and £13, 714 for MTX+prednisolone (PNS). When the EQ-5D-5L is mapped from the EQ-5D-3L trial data utilizing a limited covariate set yields an estimated 1.54 QALYs and an ICER MTX vs. MTX+CS of £6,755. In a second model the covariates are reintroduced but with an assumption of independence across health domains. This yields an estimated 1.437 QALYs with an incremental cost effectiveness ratio of £6,054 for MTX vs. MTX+CS and £15,137 for MTX vs. MTX+PNS.

A Multiverse of Imaginary Worlds

Although ICER recognizes that the modeled claims from the 2017 evidence TIM report should not be compared to those for the 2019 JAK inhibitors evidence report, this is symptomatic of a wider problem. With ISPOR and its acolyte ICER's embrace of the imaginary technology assessment meme, any numbers of models can, and have been developed within disease areas ostensibly to address the same question of lifetime claims for cost-effectiveness for specific therapies and the apparent value that these confer on the target treating population. Understandably, this creates confusion not only in respect of the diversity of modeled outcomes that utilize the same metric but the diversity of outcome metrics that are claimed to represent the 'true' measure of value. All would or could claim that they subscribe to the technology assessment meme and that they recognize the reference case framework. While it might not be clear whose value (physician, patient, insurer, health system) a model represents, the model builder can press forward in the sure and certain hope that the claims made will escape any scrutiny. The claims are 'for approximate information only' and are not intended, as detailed above, to meet standards for empirical credibility, evaluation and replication in treating environments. Claims will not be deconstructed; they will be taken at face value.

Onward to Imaginary Treatment Pathways

For those committed to the ISPOR/ICER heath technology meme, a logical next step is to consider treatment pathways within an expanded imaginary framework. As these pathways, involving both sequential and polypharmacy are complex, a characteristic of chronic diseases where RA is an exemplar, creating a framework to indulge the multiplicity of options could be seen as a 'useful' next step. Although it is early days before we witness a 'torrent' of treatment pathway imaginary

DOI: https://doi.org/10.24926/iip.v11i1.2402

models, a simulation model IVI-RA has been developed to assess the value in RA of treatment sequence strategies [5 6 7].

The IVI-RA model, which is an open source on-line resource, attempts to capture, through the creation of user-defined assumptions and constructions of multiple imaginary worlds, the effect of patient histories on therapy impact and the variation of those outcomes across patients. In its present incarnation the imaginary IVI-RA open source model is intended to provide an assessment of the value of sequential treatment strategies for those with moderate to severe RA who have not responded to conventional disease modifying antirheumatic drugs (cDMARDs) and who are naïve to biologic DMARDS (bDMARDS) or JAK/STAT (signal transducers) and inhibitors.

The model is a discrete time simulation with a 6-month cycle containing some 384 possible model structures and assumptions 'informed by previously published models' (i.e., previously constructed imaginary worlds). Presumably, not all of these model structures would be presented at the same time for a health system for review. The principal purpose of the model, which adopts an analog to the ICER reference case imaginary lifetime modeling framework, is to assess the impact on the IVI-RA assumed value criteria (the EQ-5D-3L) of the impact of parameter uncertainty, structural uncertainty and model perspective. Fundamental measurement does not appear to be an issue. The view taken is entirely consistent with the health technology meme: it is to inform decisions in order to ensure (or assist?) in efficient health care decisions.

The principal rationale for the IVI-RA is to overcome issues of model access and lack of transparency through an open source iterative collaborative approach to building imaginary worlds in RA. It is proposed to value alternative treatment pathways over the lifetime of a target cohort of RA patients allowing both cost-effectiveness analysis claims as well as multiple criteria decision analysis (MCDA). For each treatment sequence selected, the imaginary IVI-RA framework will, for selected time horizons generate the following outcomes: progression of disease severity (HAQ scores); time to treatment discontinuation; remaining life expectancy; QALYs; health sector costs and productivity losses. Issues of credibility of claims, evaluation and replication are apparently not relevant.

The authors envisage that the IVI-RA model will be continually updated as new studies provide sufficient data, agreed to by an expert panel, to change the assumptions driving the simulation model. This is seen as a necessary step in the evolution of imaginary worlds to counter the once-upon-a time feature of the overwhelming majority of published imaginary worlds where the model is held by the author with little if any chance of being updated. The open source framework for the IVI-RA model allows this process of 'continuing revolution' with presumably a continuing succession of updated models generating a budding-off cascade of 'new' scenarios to be presented to decision makers *ad infinitum*. The underlying structure of the model will, presumably remain the same with QALYs re-generated from new studies embodying the EQ-5D-3L generic instrument to re-create lifetime incremental 'manifest score' cost-per-QALY claims. The patient voice is ignored.

It is recognized, not surprisingly, that the IVI-RA can generate multiple competing model structures with a given set of assumptions, with assumption modifications such as mortality having large impacts on QALYs and incremental QALY differences (and presumably) on pricing recommendations if threshold willingness to pay criteria are introduced). Clearly, there needs to be a winnowing process to identify which of the imaginary constructs is expected to be most 'realistic' in projecting the hypothetical RA population' treatment pathways forward for 10, 20 or 30 years. One suggestion is some form of 'averaging' over the competing imaginary worlds to simplify information claims. These issues are expected to be addressed (i) through research, debate and collaboration and (ii) model averaging techniques to 'properly capture structural uncertainty'. The authors also propose the option of expected value of perfect information techniques to identify the most sensitive parameters. It is puzzling that so much attention should be given to what is after all an imaginary construct that fails the standards of normal science; a model framework that sets the initial condition for a multiverse of imaginary worlds (c.f., quantum decoherence, many worlds and the absence of waveform collapse).

Without revisiting earlier assessments of the ICER reference case, it is clear that the IVI-RA model suffers from exactly the same fatal flaws. It is, one again, pseudoscience. It is to be hoped that resources are not being devoted to emulating this framework in other disease states.

Needs Fulfilment in Rheumatoid Arthritis

If the references cited in the final evidence report for JAK inhibitors are an indication, ICER is either not interested or unaware of the amount of instrument creation, following RMT standards, which has occurred in RA over the past 25 years. Perhaps they again missed the memo. This may be because patient centric measures are of no interest or that ICER is simply unaware of the importance of measurement in therapy impact claims. The closest ICER comes to recognizing patient needs in RA is to review briefly the recent Radawski et al survey of unmet medical need in RA and to report on discussions with patients and patient organizations. [58]. As ICER notes from these surveys, only a minority of patients reported satisfaction with their treatment with ICER noting that 'much work remains to be done on quantitative, patient centered measures of treatment success' (pg. 10);'. Perhaps, once again ICER (and patient organizations) have failed to receive the memo that disease specific needs fulfillment QoL proposals had been in press for the past 40 or more years. Indeed, it is worth noting that much of the debate over the appropriateness of QALYS for particular patient groups (e.g., older persons, those with disabilities, persons with rare diseases) can be put to one side once the

DOI: https://doi.org/10.24926/iip.v11i1.2402

QALY is abandoned. The issues can be accommodated with a needs-fulfillment disease specific PCO.

The first of these instruments is the Rheumatoid Arthritis QoL (RAQoL) instrument developed in 1997 [59]. A further four rheumatology instruments were published over 10 years ago. The motivation behind the RAQoL was that with the increasing number of multinational clinical trials in RA, a measure was necessary, in a range of language versions (presently 33), that 'is derived from the experiences of RA patients'. The theoretical basis for the model is that, from a needs perspective 'life gains its quality from the ability and the capacity of the individual to satisfy his or her needs'.

The RAQoL was the first patient centric quality of life questionnaire in RA. It is also distinct from other questionnaires to include physical contact as a dimension together with capturing activities of daily living, social interaction/function, mood and recreation and pastimes. The RAQoL has 30 items with yes/no response [60]. The first five ordered items are:

- I have to go to bed earlier than I would like
- I'm afraid of people touching me
- It's difficult to find comfortable shoes that I like
- I avoid crowds because of my condition
- I have difficulty dressing

As a recent review notes, the RAQoL has excellent psychometric properties and measuring 'the impact of RA and its treatment *from the patients perspective* (emphasis added), [makes] it suitable for determining the value patients gain from interventions' [61].

The RAQoL is of particular interest as it is only one of the Galen instruments in RA that are designed to capture the patient voice. Since 2000, Galen Research PCO instruments have all been developed using the RMT standards. The other instruments develop by Galen Research in rheumatology [62] are:

- Ankylosing Spondylitis Quality of Life Measure (ASQoL) [63]
- Osteoarthritis Quality of Life Measure (QAQoL)[64]
- Psoriatic Arthritis Quality of Life Measure (PsAQoL)[65]
- Systemic Lupus Erythematosus Quality of Life Measure (LQoL)[66]

Importantly, the RAQoL has been used as an exemplar for the application of RMT to developing patient-centric models (again some 13 years ago) [67] . Also, given the timeframe for the ICER final evidence report it might be worth noting that the McKenna et al seminal papers on patient value and the needs fulfillment construct for RMT were first published in 2004, in the ISPOR journal *Value in Health*, with a further set published in the *Journal of Medical Economics* in 2018/19.

Response to Public Comments

Before asking for public comment from stakeholders in RA, it might be useful for ICER to point out that their reference case methodology fails to meet standards for credible, evaluable and replicable claims that capture the value of competing therapies for patients. More to the point, ICER should also point out that their incremental cost-per-QALY projections, while 'state of the art' fails the required measurement properties for interval scores even within the imaginary reference case framework.

If ICER admits that it is utilizing measures taken from the literature on utility metrics in RA that lack logically defensible measurement properties and that its cost-per-QALY claims lack merit, then it needs to defend its choice of metric and model and explain why patient centric measures such as the RAQoL which meet criteria for both relevance to the target patient population and unidimensionality are ignored. It is not sufficient that in its defense of QALYs to say that: *ICER believes that the QALY is highly useful and informative measure of patient outcomes with a broad context and long-standing application. Importantly the QALY reflects patient preferences for health states in a consistent and evidence based manner…*[68]. Apart from the fact that the ICER in its advocacy of a generic QALY does not address the question of whether or not the needs for patients in the target RA patient group are being captured, the ordinal measurement characteristics of the EQ-5D do not in fact reflect 'patient preferences' and certainly not in a 'consistent' manner. What is overlooked is that the patient responses for the EQ-5D-3l are weighted by community preferences; the manifest score is not the utility of patients in the target RA group, but the community valuation of the responses to the limited symptoms captured by physician choice in the EQ-5D-3L which may bear no resemblance to patients' needs.

Of course, it might be argued from a 'health care central planning' perspective that we have to apply community preferences across the board to make value claims for pricing and access in RA. This begs the questions of: (i) whose preferences in the 'community' and (ii) which utility system. Perhaps the community might also be canvassed on the importance of meeting fundamental measurement standards in value assessment.

It is also worth noting ICER's standard response to public record comments on the choice of a one-year model: *As stated in the report, we chose to model these treatments over a one year time horizon due to the uncertainty surrounding the long-term impact of these drugs, the number of subsequent lines of TIMs and if and when patients transition to palliative*. Even with this limited evidence base (and speculative use of assumptions), ICER still feels it is in a position, presumably in its role of arbiter of the 'public interest' to announce: *Our evidence review suggests that upadacitinib is modestly more effective than adalimumab, whereas the evidence cannot demonstrate added effectiveness for tofacitinib, and we found no evidence with*

DOI: https://doi.org/10.24926/iip.v11i1.2402

which to compare baricitinib to adalimumab. Policymakers will need to consider how to judge the value of a new treatment when its direct competitors are not fairly priced to begin with [69]. The notion of 'not fairly priced' being driven by the construction of a one-year imaginary world with utility estimates based on an unsupportable ordinal calibration.

Affordability

Although a therapy may meet ICER's arbitrary willingness to pay thresholds for cost-effectiveness as determined by the imaginary modeled construct, 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 cost 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 put to one side assessed clinical benefits and needs fulfillment 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 imaginary QALYs and estimate the allocation of these 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 imaginary generic QALYs are used when the focus is presumably (again) on the benefits and harms to patients.

Conclusions

If our understanding of therapy response and the development of guidelines to support evidence based decision making in health care is to advance then it has to be through the application of the scientific method to discover new facts; a process, through hypothesis testing, of conjecture and refutation. Science does not advance through constructing medieval imaginary worlds or subscribing to a meme that believes truth is consensus. It is difficult to accept that claims created by reference case models with time horizons of 10, 20 or 30 years that lack scientific merit should be taken seriously.
Nevertheless, ICER perseveres. This is, of course, supported by the absence of any attempt to generate credible and evaluable claims as a basis for a non-imaginary understanding of the contribution of JAK inhibitors to the progression of RA. As noted in previous commentaries, constructing imaginary worlds is an analytical dead end. This stands in contrast to the commitment, in many countries, to the role of RA patient registry platforms. One example is the Swedish BARFOT study and its links to national registries to capture comorbidity and mortality data [70]. If therapies such as the JAK inhibitors are to be evaluated then, rather than constructing imaginary worlds to drive formulary decisions resulting in 'approximate information' recommendations for limited access for patients to novel therapies, we should consider research platforms. These have the potential to provide a meaningful framework for assessing clinical impact as well as options for RA specific HRQoL and QoL needs-assessments with feedback to formulary committees in real time [71]. This is the approach proposed in the Minnesota formulary guidelines published in 2017 [72].

Our resources in time and effort would be better employed in tracking the impact of JAK inhibitors, or any other RA therapy, in treatment practice both to evaluate and replicate pivotal RCT claims but also to extend this to assessment for the impact of JAK inhibitors on PCO measures. Measurement is integral to this. The RAQoL has some 25 years of application in clinical practice. If we are concerned with JAK inhibitors and patient value this should be our reference point, not a smorgasbord of modeled scenarios from an on-line computer game that expressly excludes patient value and fails to meet required measurement standards.

As detailed in this commentary and in previous commentaries on ICER's reference case methodology, the approach taken by ICER in constructing imaginary worlds to support pricing recommendations fails to meet the standards of normal science. The approach should be seen as pseudoscience and any conclusions reached subject to this caveat. The fact remains, and ICER implicitly acknowledges this in its reformulation of its 'house model', there is a potential multiverse of RA models, to include competing models for JAK inhibitors. This is shown in the ICER review of the RA model literature. In practice, you choose your timeframe, the model structure and the assumptions to drive your conclusions.

The potential for a multiverse of models is further enhanced by the need to apply utility values to create cost per QALY claims. In common with other generic utility instruments, the EQ-5D does not meet required cardinal measurement standards. It also, in common with other instruments, puts to one side any claim to be a patient centric instrument. If we are to assess the impact of competing therapies in RA then we need to forget operational symptom and functioning HRQoL instruments which lack a coherent latent construct to those patient centric instruments which have a coherent needs-based QoL construct.

The question that is of most concern is whether or not ICER (and IVI-RA) will take note of the criticisms advanced here. If the end-game is to provide a 'take our word for it' price discounting and

DOI: https://doi.org/10.24926/iip.v11i1.2402

affordability media release, then issues such as the appropriate measure of patient value that meets RMT criteria are beside the point. After all, we can safely assume that few in the audience will delve any deeper than the final recommendations. Indeed, from a global perspective there is a health technology assessment industry devoted to exercising its imagination in cost-effectiveness claims. Whether it is possible to convince advocates that the 'emperor has no clothes' is questionable given their long-term commitment of ISPOR and others to this meme and their ongoing success in enforcing meme transmission fidelity.

In the absence of a commitment by ICER to the standards of normal science, ICER can expect continued criticism of its modeling claims by manufacturers and other interest groups pointing to failure of ICER to meet the standards of normal science. If the focus on imaginary 'approximate information' constructs is seen as the gold standard this merely underscores the point that not only can competing models be built, but that it is always possible to 'reverse engineer' a model to achieve a 'supportive' cost-effectiveness claim. Certainly, the various models can be viewed as providing 'approximate information' as well as 'approximate disinformation'; the problem is the impossible task of distinguishing one from the other.

It is unclear what decision makers are to conclude in the prospective deluge of competing and conflicting claims for preference weights, utility metrics and cost-effectiveness: drowning in information yet thirsting for knowledge. As a recent opinion piece in the New York Times commented: *But the internet is as good a tool to restrict free speech – to flood the zone with so much low-quality information that the marketplace of ideas becomes landfill where it's impossible to separate the good from the garbage* [73].

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

References

[1] ICER. Janus Kinase Inhibitors for Rheumatoid Arthritis: Effectiveness and Value. Draft Evidence Report. September 26, 2019 https://icer-review.org/wp-content/uploads/2019/03/RA_Working_Draft_Report_PUB2.pdf

[2] ICER. Janus Kinase Inhibitors for Rheumatoid Arthritis: Effectiveness and Value. Draft Evidence Report. October 11, 2019 https://icer-review.org/wp-content/uploads/2019/03/ICER_RA_Draft_Evidence_Report_101119.pdf

[3] ICER. Janus Kinase Inhibitors for Rheumatoid Arthritis: Effectiveness and Value. Evidence Report. 26 November 2019 https://icer-review.org/wp-content/uploads/2019/03/ICER_RA_Evidence_Report_112619.pdf

[4] ICER. Janus Kinase Inhibitors and Biosimilars for Rheumatoid Arthritis: Effectiveness and Value. Final Evidence Report and Meeting Summary. 9 January 2020 https://icer-review.org/wp-content/uploads/2019/03/ICER_RA_Final_Evidence_Report_and_Meeting_Summary_010820.pdf

[5] Incerti D, Curtis J, Shafrin J et al. A flexible open-source decision model for value assessment of biologic treatment for rheumatoid arthritis. *Pharmacoeconomics*. 2019;37:829-43

[6] Jansen J, Incerti D, Curtis J. Toward relevant and credible cost-effectiveness analyses for value assessment in the decentralized US health care system. *J Managed Care Special Pharm*. 2019;25:5(3):518-21

[7] Innovation and Value Initiative (IVI). Issue Brief: Aligning vale assessment with treatment in chronic diseases. *Value Blueprints*. 9 December 2019 https://www.thevalueinitiative.org/wp-content/uploads/2019/12/Value-Blueprint_Aligning-Value-Assessment-with-Treatment_FINAL.pdf

[8] Langley PC. Cost-Effectiveness and Formulary Evaluation: Imaginary Worlds and Entresto Claims in Heart Failure. *Inov Pharm*. 2016;7(3): No. 6 https://pubs.lib.umn.edu/index.php/innovations/article/view/449

DOI: https://doi.org/10.24926/iip.v11i1.2402

[9] Langley PC. Multiple Sclerosis and the Comparative Value Disease Modifying Therapy Report of the Institute for Clinical and Economic Review (ICER). *Inov Pharm.* 2017;8(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/492

[10] Langley PC. Imaginary Worlds and the Institute for Clinical and Economic Review (ICER) Evidence Report: Targeted Immune Modulators for Rheumatoid Arthritis. *Inov Pharm.* 2017;8(2): No. 10. https://pubs.lib.umn.edu/index.php/innovations/article/view/515

[11] Langley PC. Rush to Judgement: Imaginary Worlds and Cost-Outcomes Claims for PCSK9 Inhibitors. *Inov Pharm.* 2017;8(2): No. 11 https://pubs.lib.umn.edu/index.php/innovations/article/view/516

[12] Langley PC. Another Imaginary World: The ICER Claims for the Long-Term Cost-Effectiveness and Pricing of Vesicular Monoamine Transporter 2 (VMAT2) Inhibitors in Tardive Dyskinesia. *Inov Pharm*. 2017;8(4): No 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/927

[13] 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;21:119-123

[14] McKenna S, Doward L, Niero M et al. Development of needs-based quality of life instruments. *Value Health*. 2004;7(Supp1):S17-S21

[15] Doward L, McKenna S, Meads D. Effectiveness of needs-based quality of life instruments. *Value Health*. 2004;7(Supp 1):S35-S38

[16] Tennant A, McKenna S, Hagell P. Application of Rasch Analysis in the Development and application of quality of life instruments. *Value Health*. 2004;7(Suppl 1):S22-S26

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

[18] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21(5):464-80

[19] 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

[20] McKenna S, Heaney A, Wilburn J. et al. Measurement of patient-reported outcomes. 2: Are current measures failing us? *J Med Econ.* 2019;22(6):523-30

[21] 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

[22] 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

[23] 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

[24] ICER. RA Update: Revised Scoping Document 9 May, 2019 https://icer-review.org/material/ra-update-revised-scoping-document/

[25] ICER. RA Update: Draft Scoping Document 11 April 2019 https://icer-review.org/material/ra-update-draft-scoping-document/

[26] ICER. RA Update: Public Comment on Draft Scoping Document. https://icer-review.org/material/ra-update-public-comment-on-draft-scoping-document/

[27] Arthritis Foundation, April 30 2019 https://icer-review.org/wp-content/uploads/2019/03/ICER_RA_Update_Draft_Scope_Comments_050919.pdf

DOI: https://doi.org/10.24926/iip.v11i1.2402

[28] ICER. RA Update: Updated Scoping Document. June 2019 https://icer-review.org/material/ra-update-updated-scoping-document/

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

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

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

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

[33] Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the economic evaluation of health technologies: Canada. Ottawa: CADTH, 2017

[34] Langley P. ICER, ISPOR AND QALYs: A Tale of Imaginary Worlds. *Inov Pharm*. 2019;10(4):article 10 https://pubs.lib.umn.edu/index.php/innovations/article/view/2266

[35] Dawkins R. The Selfish Gene (30th Anniversary Ed). Oxford: University Press, 2006

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

[37] Langley PC. Imaginary worlds: Modeled claims for cost-effectiveness published in PharmacoEconomics January 2015 to December 2015. *Inov Pharm*. 2016;7(2): Article 9. https://pubs.lib.umn.edu/index.php/innovations/article/view/432

[38] Langley PC, Rhee TG. Imaginary worlds: A systematic review of the status of modeled cost-effectiveness claims published in the Journal of Medical Economics from January 2015 to December 2015. *Inov Pharm*. 2016;7(2): Article 16. https://pubs.lib.umn.edu/index.php/innovations/article/view/439

[39]Langley PC, Rhee TG. Imaginary worlds: The status of modeled economic evaluation claims published in *Value in Health* January 2015 to December 2015. *Inov Pharm*. 2016;7(2): Article 18. https://pubs.lib.umn.edu/index.php/innovations/article/view/441

[40] Langley PC, Rhee TG. Imaginary worlds: A systematic review of the status of modeled cost-effectiveness claims published in the Journal of Medical Economics from January 2015 to December 2015. *Inov Pharm*. 2016;7(2): Article 16. https://pubs.lib.umn.edu/index.php/innovations/article/view/538

[41] Langley PC, Rhee T. The Imaginary Worlds of ISPOR: Modeled Cost-Effectiveness Claims Published in Value in Health from January 2016 to December 2016. *Inov Pharm.* 2017;8(2): Article 14 https://pubs.lib.umn.edu/index.php/innovations/article/view/538

[42] Langley PC. Validation of modeled pharmacoeconomic claims in formulary submissions. *J Med Econ*. 2015;18(12):993-99

[43] Langley PC. Modeling imaginary worlds: Version 4 of the AMCP Format for Formulary Submissions. *Inov Pharm*. 2016;7(2): No. 11 https://pubs.lib.umn.edu/index.php/innovations/article/view/434

[44] Brazier J, Ara R, Azzabi J 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

[45] 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

[46] NICE. Position statement on use of EQ-5D-5L valuation set for England (updated November 2019) https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l

[47] Drummond M, Sculpher M, Torrance et al. Methods for the Economic Evaluation of Health Care Programmes 3rd Ed. Oxford University Press, 2005.

[48] Grimby G, Tennant AQ, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? (Editorial). *J Rehab Med*. 2012;44:97-98

[49] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:667-680

[50] Hunt S, McKenna S, McEwen J et al. A quantitative approach to perceived health status: a validation study. *J Epidem Comm Health*. 1980;34:281-86

[51] Luce R, Tukey J. Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psychol*. 1964;1:1-27

[52] Wailoo A, Hernández-Alava M, Manca A et al. Mapping to Estimate Health-State Utility from Non-Preference-based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. *Value Health*. 2017;20:18-27

[53] ICER. Evidence Report: Targeted Immune Modulators for Rheumatoid Arthritis. April 2017 https://icer-review.org/wp-content/uploads/2016/08/NE_CEPAC_RA_Evidence_Report_FINAL_040717.pdf

[54] 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

[55] Hernández-Alava M, Wailoo A, Grimm S et al. EQ-5D-5L versus EQ-5D-3L; The impact on cost-effectiveness in the United Kingdom. *Value Health*. 2018;21(1):49-56

[56] Pennington B, Hernández-Alava M, Pudney S et al. The impact of moving from EQ-5D-3L to -5L in NICE technology appraisals. *Pharmacoeconomics*. 2019;37(1):75-84

[57] Hernández-Alava M, Pudney S. Econometric modelling of multiple self-reports of health states: the switch from EQ-5D-3L to EQ-5D-5L in evaluating drug therapies for rheumatoid arthritis. *J Health Econ*. 2017;55:139-52

[58] Radawski C, Genovese M, Hauber B et al. Patient perceptions of unmet medical need in rheumatoid arthritis A cross-sectional survey in the USA. *Rheumatol Ther*, 2019;6:461-71

[59] De Jong Z, van der Heijde D, McKenna S et al. The reliability and construct validity of the RAQoL: A rheumatoid arthritis-specific quality of life instrument. *Br J Rheumatology*. 1997;36:878-83

[60] Galen Research, Manchester UK http://www.galen-research.com/content/measures/RAQoL_UK_-_First_page_sample.pdf

[61] Heaney A, Stepanous J, Rouse M et al. A review of the psychometric properties and use of the Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL) in clinical research. *Curr Rheumatol Rev*. 2017;13(3):197-205

[62] Galen Research, Manchester UK http://www.galen-research.com/measures-database/

[63] Doward LC, Spoorenberg A, Cook SA, et al. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Ann Rheum Dis.* 2003;62:20–6.

[64] Keenan A, McKenna S, Doward L et al. Development and validation of a needs-based quality of life instrument for osteoarthritis. *Arthritis Rheum*. 2008;59(6):841-8

[65] McKenna S, Doward L, Whalley D et al. Development of the PsAQoL: a quality of life instrument specific to psoriatic arthritis. *Ann Rheum Dis*. 2004;63(2):162-9

[66] Doward L, McKenna S, Whalley D et al. The development of the L-QoL: a quality of life instrument specific to systemic lupus erythematosus. *Ann Rheum Dis*. 2009;68(2):196-200

[67] Tennant A, Conaghan P. The Rasch Measurement Model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper. *Arthritis Rheumatism*. 2007;57(8):1358-62

[68] ICER. Janus Kinase Inhibitors for Rheumatoid Arthritis: Effectiveness and Value - Response to Public Comments on Draft Evidence Report. 26 November 2019 https://icer-review.org/wp-content/uploads/2019/03/ICER_RA_Response_to_Public_Comments_112619.pdf

[69] ICER Press Release 26 November 2019 https://icer-review.org/announcements/jak_inhibitor_evidence_report/

[70] Hafström I, Ajeganova S, Andersson M et al. A Swedish register based, long-term inception cohort study of patients with rheumatoid arthritis – results of clinical relevance. *Open Access Rheumatology Res Reviews*. 2019;11:207-217

[71] Malm K, Bergman S, Andersson M, et al. Quality of life in patients with established rheumatoid arthritis: A phenomenographic study. *SAGE Open Medicine*. 2017;5:1-8

[72] 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

[73] Warzel C. Pierre and the Paradox of Anonymity. New York Times. 30 October 2019