MAIMON WORKING PAPERS No. 7 FEBRUARY 2022

A NEW START IN FORMULARY SUBMISSIONS: REJECTING IMAGINARY MODELED CLAIMS FOR RARE DISEASES

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

Abstract

Rare diseases present significant challenge to formulary committees where data at product launch are limited while modeled claims are submitted attempting to justify pricing and access where the model lacks any attempt to meet the standards of normal science. Unfortunately, the commitment to assumption driven lifetime simulations to support imaginary claims for cost-effectiveness is a meme or belief of those involved in health technology assessment. The manifest deficiencies of this analytical framework are well known yet, as evidenced by the recent release of the CHEERS 22 guidance for creating imaginary modeled claims, demonstrates the continued acceptance of this belief. The application of an unacceptable analytical framework present unsought for challenges for manufacturers and other that have pursued new therapies for rare diseases, only to face imaginary non-evaluable recommendations for substantial price discounting. For potential investors in rare disease therapies, for which there is a clear and longstanding unmet need, this is a roadblock. If investors pull back, then patients and caregivers will suffer. This is a ridiculous situation, irrespective of whether or not health systems are in a position to pay for new therapies, when the entire exercise fails scientific standards and meets all criteria for pseudoscience and metaphysics. Clearly, we need a new framework, a New Start, to evaluate value claims for new rare disease therapies that can stand up to scrutiny. The framework proposed here is not unique to rare diseases. It represents a difference in degree rather than in kind as the New Start framework applies across the board to therapies in chronic disease. What sets rare disease apart is the high restricted evidence base that accompanies new rare disease therapies. We have to resolve the question of evidence but not through the invention of assumption driven model claims that clearly fail standards for scientific integrity.

INTRODUCTION

The principal flaw, and it is fatal, for health technology assessment of pharmaceutical products and devices is the insistence, by organizations such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Institute for Clinical and Economic Review (ICER) together with leaders in the field, is to support inventing claims for cost-effectiveness [1] [2]. This has been amply demonstrated over the past 30 years with the commitment, to a lesser extent in the US, but embraced enthusiastically globally, to the construction of assumption driven simulation models, typically looking forward over decades to track the natural course of a disease. The National Institute for Health and Care Excellence (NICE) is the archetypal contributor with the requirement for reference case modeling to

produce non-evaluable claims for cost-effectiveness [3]. Most recently, the release of the CHEERS 22 guidance has attempted to reinforce this commitment to the pre-eminence of imaginary claims; this is to be achieved by asking analysts to conform to an imaginary claim framework for submitting non-evaluable cost-effectiveness claims to leading journals [4]. Two points are worth noting: (i) this submission framework has been endorsed by 15 journals and (ii) there is no mention of the evidence claims that might be relevant to formulary committees. It is assumed, apparently, that speculative, non-empirically evaluable comparative claims are a desired input to formulary decisions.

Challenges facing analysts in the evaluation of rare diseases and the benefits of interventions are documented in a 2018 ISPOR Report of the Rare Diseases Special Interest Group [5]. Although the ISPOR report supports the application of QALYs in assumption driven simulated claims, the issues of (i) disease recognition and diagnosis; (ii) the evaluation of treatment effect; and (iii) patient recruitment for clinical research (including misdiagnosis) are addressed. These are substantive issues from the perspective of clinical trials, but what is missing is the equally substantive, if not more important issue, of creating and presenting empirically evaluable value claims for the rare disease therapy. The ISPOR report makes the point that at product launch and the early stages of product life cycle, data are often limited; in many cases restricted to a handful of Phase 2 and 3 clinical trials. At this juncture, what is missing in the ISPOR report is any commitment to meeting the standards of normal science in the process of technology assessment. If meeting standards for credibility, empirical evaluation and replication are considered central to any health technology assessment then the obvious question is: what is the proposed framework for health technology assessment and what are its evidentiary requirements? While the ISPOR report acknowledges the potential difficulties with the absence of 'validated' outcomes instruments, the analysis reverts to the need to meet the requirements of the ISPOR belief in the creation of assumption driven simulation models, based on the mathematically impossible QALY, incremental cost-per-QALY claims and the application of cost-per-QALY thresholds. This is the fundamental error.

The purpose of this brief commentary is to argue for abandoning the ISPOR approximate information modeling in the evaluation of rare disease therapies; the mistaken application of assumption driven simulations to create an evidence base and support blanket claims for product cost-effectiveness. Rare diseases deserve more than the application of an analytical framework which, as demonstrated here, is a dead end which not only fails the standards of normal science but ignores the axioms of fundamental evidence. While it is difficult to underestimate the potential impact of these challenges and the limited evidence base that typically accompanies rare disease therapy claims, the purpose here is more focused on the health care system. The question is straightforward: How are claims for rare disease interventions to be empirically evaluated?

APPROXIMATE INFORMATION

The standard, in what has been described as the health technology assessment meme, is to create an assumption driven lifetime simulation to invent non-evaluable outcome and cost-effectiveness claims. This is the global standard, widely endorsed by assessment agencies, academic groups and journal

editors. Unfortunately, it fails the standards of normal science. The claims produced, by design, are not credible, evaluable or replicable. On the criteria proposed to demarcate science from non-science, the construction of imaginary claims falls clearly under the category of non-science or pseudoscience and metaphysics [6].

The decision to create evidence, driven in large part by the limited evidence typically available at product launch was deliberate; it was a lot easier (and quicker) to create an assumption driven model to support non-evaluable claims for cost-effectiveness than to propose a program aimed to discover new, yet provisional, facts about therapy impact in diseases [7]. To this extent, rare diseases are no different from any run-of-the-mill chronic disease. Many of the challenges articulated above point to a difference in degree not in kind between those facing analysts in rare diseases compared to more prevalent disease states.

The decision by leaders in the field of technology assessment had an unfortunate consequence: 30 years of wasted effort in health technology assessment through chasing the will o'the wisp of imaginary claims; a meme or belief system that is now endorsed by thousands of analysists. A belief system that eschews discovery in favor of continually re-inventing yet more imaginary claims; in other words a barren, analytical dead-end. Whether this belief system will be overcome is a moot point. It is well entrenched with a lot to lose by leaders in the field.

A DIGRESSION ON MEASUREMENT

In both the physical and more mature social sciences measurement is the key to the discovery of new, yet provisional, facts to evaluate the benefits of new therapies for both rare disease and, more generally, chronic disease that impact wider populations. Unless the instruments that are applied have been designed to have the appropriate measurement standards, then we can say little is anything about the merits of competing therapies or a new therapy designed to replace or supplement the existing standard of care.

Following the formalization by Stevens and others in the 1930s and 1940s, scales or levels of evidence used in statistical analyses are classified as nominal, ordinal, interval or ratio [8] . Each scale has one or more of the following properties: (i) identity where each value has a unique meaning (nominal scale); (ii) magnitude where values on the scale have an ordered relationship with each other but the distance between each is unknown (ordinal scale); (iii) invariance of comparison where scale units are equal in an ordered relationship with an arbitrary zero (interval scale) and (iv) a true zero (or a universal constant) where no value on the scale can take negative scores (ratio scale). Nominal and ordinal scales only support nonparametric statistics. Interval scales can support addition and subtraction while ratio scales support the additional operations of multiplication and division as they have a true zero. This zero point characteristic means it is meaningful to say that the one object is twice as long as another. Given these limitations, the only acceptable empirically evaluable value claims are those designed for single attributes with interval or ratio properties; this means that we reject composite multiattribute direct and indirect generic preference scales as they lack

dimensional homogeneity and construct validity, with the same applying the virtually all disease specific PRO scales. This may seem an unduly narrow conclusion, but the fact is that it is the only conclusion we can come to given these axioms. This means, to re-emphasize, rejecting the overwhelming majority of both generic and disease specific outcomes instruments. If we accept this conclusion then value claims for rare disease become more manageable and meaningful. But first it is important to make the case for abandoning the ISPOR belief in the creation of assumption driven claims for product cost-effectiveness; a term that is misleading and redundant [9].

ORDINAL PREFERENCES AND IMPOSSIBLE QALYs

It has been known for some time, yet studiously ignored, that the QALY is an impossible mathematical construct [10]. The reason is obvious: direct and indirect multiattribute preference scores (e.g., EQ-5D-3L/5L) and the majority of disease specific PROs have ordinal properties; they are ordinal scores. As such they cannot support the arithmetic operation of multiplication. This is required because you are multiplying time spent in a disease state by a preference score. Apart from the fact that no one thought about the required fundamental measurement standards when these various PRO instruments were developed, if you want a preference score to have multiplicative properties then it must be a bounded ratio scale; that is, a true zero capped at unity. The preferences are then proportions. Unfortunately, the preference scores used are capped at unity with dis-utilities as decrements, with an open ended lower bound determined by the various score algorithms. Again, unfortunately, there is no true zero as the various preference algorithms produce negative values (states worse than death). The implications of this have been sidelined with ICER, for example, convinced that its belief in the ordinal preference score is that it is really a mystical ratio scale in disguise which is, ICER believes, commonly held by health economists. The more reasonable conclusion is that few analysts in health technology assessment have actually thought about the axioms of fundamental measurement and the constraints they impose on instrument development and interpretation.

STATES WORSE THAN DEATH

The fact that the composite scoring algorithms that support ordinal preference scores can generate negative values or states worse than death has been recognized since the algorithms were first applied; the response has been to ignore this unfortunate characteristic or, more bluntly, sweep it under the carpet. In the case of the EQ-5D-3L, for example, the most widely applied composite preference score, the algorithm determines scoring range is from 1 = perfect health to -0.58 (with death = 0). In the case of rare diseases where the EQ-5D-3l is applied there are two questions of interest: (i) what is the distribution of ranked values for a given target patient population and (ii) what is the impact of negative values (if present) on the overall 'average' EQ-5D-3L score. The average is, of course, disallowed as the score is ordinal (and disallowed also because it is dimensionally heterogeneous), but this is the form in which it is usually presented, with equally disallowed measures of dispersion (e.g., standard deviations, range). Interpreting a positive 'average' preference score which includes negative values is difficult to interpret; particularly as the average is meaningless.

It should not be though that negative 'average' ordinal preference scores are relatively infrequent. The best example of the pervasiveness of these negative scores is from the Tufts Medical Center Cost-Effectiveness Analysis (CEA) database. This database was initiated 46 years ago and comprises extracts from studies (now over 8000) that have present cost-utility analyses. Apart from summarizing preference or utility scores from the various multiattribute instruments, the data base includes a range of impossible mathematical measures to include QALYs, cost-per-QALY claim and incremental cost-per-QALY claims. There are now some 36,000 preference scores for health states; obviously a go-to database for constructing imaginary modeled claims. Unfortunately, no one apparently recognized that these preference scores are composite ordinal 'averages' and that the entire exercise is essentially a waste of time (and mathematically disallowed); except, presumably, for users who believe ordinal preference scores are actually ratio scales in disguise. This belief is challenged by the fact that, from the 100 health state 'average' preference scores on the Tufts CEA website, some 47% present with apparently negative values. The range of composite ordinal negative health states is from -0.01 to -0.55; the range for positive weights is from zero to 0.93. These ranges are questionable because they reflect the algorithm used. The various multiattribute instruments are not comparable; they reflect different health dimensions and levels of response that are collapsed into the scoring algorithms, as well as the basis for preference scoring together with differences in the theoretical assumptions underpinning the modeling.

Given the likelihood that the application of direct or indirect multiattribute preference instruments in rare diseases, the multiple violations of fundamental measure standards is a salutary lesion for any residual belief in assumption driven simulated model claims to support pricing and access recommendations in rare diseases; a residual belief that is further reinforced by the application of cost-utility thresholds.

IMAGINARY CLAIMS AND IMAGINARY THRESHOLDS

One of the more pointless exercises in the application of modeled imaginary incremental -cost-per-QALY claims is the raising of the threshold cost-per-QALY bar for rare disease therapies; a role that is confirmed in the ISPOR rare disease report [11]. ICER has proposed that in the case of ultra rare diseases (defined by ICER as a prevalence $<3/10^5$) that aside from providing their standard analysis where thresholds are $150,000 per QALY, there should be further threshold of $500,000 per QALY with associated pricing recommendations. This makes little sense because the modeling to support the incremental cost per QALY and the cost-per-QALY threshold fail to meet the required measurement standards; the preferences are ordinal and the QALY mathematically impossible. Proposing a specific threshold for ultra rare diseases might be thought to go some way towards accommodating criticisms of ICER's standard threshold, but in it is just a sleight of hand. It might be believed, but if belief is the issue then it is a belief in an analytical framework that is a dead end. Indeed, as a further effort to support the impossible QALY, ICER proposes, given the absence of data, to crosswalk to the QALY any surrogate quality or health related quality of life measure. As the surrogate measures, which presumably will be disease specific will typically fail to meet the required measurement standards as Likert-based ordinal scales, the mapping (or crosswalking) exercise is again a waste of time. The reason is clear: you cannot map from one ordinal scale to another and even map to an ordinal scale. However ICER may attempt to

respond to criticisms, it cannot overcome the fact is that the models and its business case fail the standards of normal science.

ABANDONING RARE DISEASE MODELS

Given these manifest deficiencies in ICER models and, indeed, models that characterize the approximate information meme as detailed in CHEERS 22. The conclusion that these models should be abandoned is unsurprising. It has, however, a particular resonance with rare diseases where data are typically limited at product launch which means, given the imperative from ICER and other technology assessment groups, to stake out a 'market entry' claim for pricing (and likely restricted access) in the target population for a new therapy, that the contribution of assumptions is substantial. The problem with assumption driven simulation models is, of course, that there is no basis in logic to prefer one set of rare disease assumptions over another. This means, as ICER is aware of this possibility with its launch of ICERAnalytics, that there is a potential multitude of competing models not only for a rare disease but for any other chronic condition [12]. This sets the stage for formulary committees evaluating submissions for rare disease therapies to find the selves bombarded by competing modeled value claims, with no basis for judging the relative merits of any of them.

If a manufacturer is to challenge ICER then the opportunity is there; ICER cannot sustain its business model unless it rejects the standards of normal science, accepts fundamental measurement while rejecting as well the invention of evidence for non-evaluable claims. It is interesting to note that while there is an extended literature on the practical shortcomings of the QALY in its scope and relevance for target patient groups, no one thought that the debate is actually a complete waste of time as the QALY cannot exist; it is mathematically impossible. The implications are dire, at least for the assumption driven incremental-cost-per-QALY model builders. As the QALY cannot exist, the entire exercise is a waste of time with the various outcome claims for QALYs and the application of sensitivity and probabilistic sensitivity analysis an irrelevant exercise There are no implications that can be drawn for rare or any other disease; it is just the exercise of a belief system which after 30 years is well entrenched yet irrelevant..

A NEW START: ATTRIBUTES AND PROTOCOLS

The proposed New Start in formulary submissions and product value claims rests on two premises:

(1) All value claims for a product or therapeutic intervention must refer to a single attributes that meet the demarcation standards for normal science: all value claims must be credible, evaluable and replicable;
(2) All value claims must be consistent with the limitations imposed by the axioms of fundamental measurement: they must meet interval or ratio measurement standards

These are requirements that are accepted in the physical and more mature social sciences. They have stood the test of time over the past three centuries and more from the scientific revolution of the 17th

century. In a nutshell, they respect the motto of the Royal Society (1660): *nullius in verba* (take nobody's word for it). In the case of the approximate information models of ISPOR, ICER and CHEERS 22, we have to take (or reject) anybody's word for it in value claims, pricing and access [13]. If the focus had been on these premises some 30 years ago we would not be in the present situation.

The proposed new start in formulary submissions makes the case that all core value claims for a product should be specific to the target patient population with single attribute claims having evaluable, unidimensional ratio or interval properties. There are no other options. This sets the stage for a progressive evaluation of claims as evidence is presented to assess these claims; to conform to the standards of conjecture and refutation in the physical and mature social sciences.

The value claims can be considered under three categories: (i) value claims that have clinical ratio properties; (ii) value claims that are PROs but with interval (and if possible) ratio properties that capture the latent construct; and (iii) resource utilization claims that include product uptake, compliance and adherence, again with ratio properties. PROs present a challenge as there are only a few disease specific measures that meet Rasch Measurement Standards, with even fewer focused on a rare disease [14]. These sibgle attribute value claims are need-fulfillment measures which capture response to therapy but as interval measures cannot support multiplication and division. Recently, it been demonstrated that for selected interval measures, it is possible to apply a transformation that creates a bounded ratio scale in the range 0 (a true zero) to unity [15]. This allows response to therapy to be evaluated in terms of need-fulfillment; the impact of competing therapies on the needs of the target population, patients and caregivers,and the benefit of that therapy in meeting unmet need. This is a critical step forward in therapy value claims.

The link between value claims and evidence to assess that claim is the protocol; this is not a recent innovation as it was proposed some 18 years ago. It is the missing link in value claims. Manufacturers should be required to submit a protocol to support each value claim that constitutes the formulary submission. This applies equally to all chronic disease states, including rare diseases, as well as if required to short-term or acute interventions.

Rare diseases present many of the same challenges that are faced for more prevalent chronic disease states and, to a lesser extent acute disease states where claims are extrapolated in the short term from RCTS. A principal issue with rare diseases is the limited access to data (or its absence) to evaluate claims in the real world together with the virtual absence of PROs to capture claims for latent constructs such as needs fulfillment quality of life. Protocols should be aligned with a delivery timetable for reporting individual value claims. The manufacturer should propose a timetable and seek agreement from the formulary committee for its implementation. This is not an unusual request as there are many examples of contractual agreements to support product roll out and the sharing of cost savings against clinical targets. This proposal is more comprehensive as it relates to a range of what a manufacturer in agreement with the formulary committee might designate as core value claims to support ongoing provisional pricing and access decisions. Implicit is this choice of core values is the possible commitment to supporting ongoing disease area and therapeutic reviews.

Agreeing to a protocol to support what might be described as core value claim, presupposes that the health system, in anticipation of a submission for a designated target patient population in a rare disease state, has identified the value claims it considers important to support pricing and access negotiations. The committee should put to one side third party models which produce non-evaluable claims, employment standards which fail to meet those of normal science. The focus is on core value claims and the timeframe over which they can be evaluated. In many cases, the value claims will be outside the remit of phase 2 and 3 clinical trials or, if included as outcome claims. The committee should, therefore, given the unique nature of rare diseases, prepare a statement defining the required standards and the relevant core value claims. The request to a manufacturer to submit a formulary submission (required in the US) should detail the value claims for the target therapy indication for clinical end points, PROs and resource allocation (including compliance and therapy uptake). Ideally, the manufacturer will have raised this question as part of product development, seeking if possible feedback from potential client health care systems.

Feedback to establish core value claims should also be sought from patients, caregivers, physicians and patient advocacy groups. This is of particular interest in the questions of quality of life, defined in this case by the patient voice in need fulfillment, where this applies both to patients and caregivers, particularly for pediatric patients. Therapy impact is relevant both for the patient and the caregiver.

The catalysts for developing rare disease evidence bases are the proposed protocols to support claims assessment. These are probably the strongest incentive to consider a registry or, if one exists, the review and updating of a registry. Given the demands of formulary committees for feedback in a meaningful time frame, the registry is the most promising contender for not only post-market entry claims assessment but the basis for ongoing disease area and therapeutic class reviews. The focus must be on the discovery of new, yet provision, facts to support ongoing pricing and access negotiations as well as feedback to patients, caregivers, physicians and health systems. Predicating claims assessment on a protocol driven proposal to establish a registry would be a major step forward; instead the risk is a succession of imaginary assumption driven claims justified by their conformity to the CHEERS 22 guidance with, essentially, zero value.

CONCLUSIONS

It is not the purpose of the proposed New Start in formulary submissions to support assumption driven imaginary claims for overall cost-effectiveness; this is an analytical dead end. Rather the purpose is to provide evidence to evaluate an agreed profile of value claims. This is achieved by linking claims to an assessment by a protocol that details how the value claim is to be assessed and reported to the formulary committee or other health system decision maker in an agreed timetable.

The contributions of ISPOR, ICER and academic groups has led to thousands believing in the pre-eminence of inventing evidence by assumption for over 30 years. The belief holds sttrong due to its transmission fidelity. In common with other enduring belief systems, believers and those influenced by them, have been chasing the will o'the wisp of an analytical dead end: imaginary claims. Rare diseases

are particularly susceptible to these blandishments; the absence of evidence outside clinical trials, the difficulty or impossibility of securing data to support claims evaluation in the immediate or very short run have made claims for rare disease impacts, notably for an unknown future, an easy target for modeled simulation purveyors. If technology assessment had committed at the beginning to creating data sets to support claims evaluation, particularly in rare diseases, we would, after 30 years be in a more advantageous position.

Abandoning the approximate imaginary information meme is long overdue; It rejects the role of science in the discovery of new, yet provisional facts, in favor of non-evaluable impossible claims. We are now in a position to put this behind us. Certainly, protocols may be difficult to implement and underwrite by manufacturers, but in terms of chronic disease states, and not just rare diseases, the New Start is a major step forward. Rare diseases will raise challenging issues due to, in many cases, a real world data vacuum, but the health system and the formulary committee will be in the box seat. The critical issue is one of reporting on the assessment of value claims. Protocols should propose and agree a timeframe for reporting claims assessments to a formulary committee or other nominated decision makers with, importantly, a future commitment to an ongoing support of disease area and therapeutic reviews. This applies equally to all chronic disease states and not just rare diseases. The latter have attracted attention for, in all likelihood, the fact that they don't sit easily in the approximate information modeling assessment which has led to ad hoc adjustments of cost-per-QALY thresholds which seems, once again, pointless given the irrelevance of the analytical framework.

REFERENCES

[1] Langley P. Peter Rabbit is not a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(2): No 22

[2] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *Inov Pharm*. 2020;11(1):No. 12

[3]Langley PC. Sunlit uplands: the genius of the NICE reference case. *InovPharm*. 2016;7(2): No.12.

[4] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 22) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth*. 2022;25(1):3-9

[5] Nestler-Parr S, Korchagina D, Toumi M et al. Challenges in Research and Health Technology Assessment of Rare Disease Technologies: Report of the ISPOR Rare Disease Special Interest Group. *ValueHealth*. 2018;493-500

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

[7] Neumann P, Willke R, Garrison L. A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *Value Health*. 2018;21:119-123

[8] Stevens S. On the theory of scales of measurement. *Science*. 1946;103: 677-80

[9] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes (3rd Ed.).New York: Oxford University Press, 2015

[10] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No. 7

[11] Institute for Clinical and Economic Review. Modifications to the ICER value assessment framework for treatments for ultra-rare diseases. Final version. November 2017

[12] Langley P. Let a Thousand Models Bloom: ICER Analytics Opens the Floodgates to Cloud Pseudoscience. *Inov Pharm*. 2021;12(1):No. 5

[13] Feyerabend P. Against Method. New York: Schocken, 1975

[14] Bond T, Fox C. Applying the Rasch Model: Fundamental Measurement in the Human Sciences 3rd Ed. New York Routledge, 2015

[15] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6