

A Critical Examination of Simulation Pricing and Access Recommendations for Atopic Dermatitis

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

ABSTRACT

It has been demonstrated conclusively that value and utility preference scores have only ordinal properties. This means, as has been pointed out on numerous occasions, that the quality adjusted life year (QALY) is a mathematically impossible construct. The implications are profound: Some 30 years of health technology assessment is called into question due to a failure to recognize the well-documented limitations imposed by the axioms of fundamental measurement. The purpose of this commentary is provide a critical examination of this practice in recommendations for atopic dermatitis.

Keywords: atopic dermatitis, ICER, pseudoscience, negative preferences, QoLIAD, PIQoL-AD

INTRODUCTION

This commentary focuses on lifetime assumption-driven simulation models claims¹ that are used to support social pricing and denial of care for products entering the US market². It is proposed that this is a futile endeavor³. Its genesis can be traced back to the early 1990s when in order to claims for cost effectiveness for formulary submissions, economists, pharmacists and others in health technology assessment decided that if evidence was not available at product launch it should be invented⁴. Hypothesis testing was rejected (too time consuming) in favor of inventing evidence (the euphemistic phrase was 'approximate information'). It seems odd that you can create lifetime modeled approximate information when there is no reference point to judge the worth of 'approximate'; approximate to what? Outcomes, which are non-evaluable, representing targets for an unknown future? Based on a smorgasbord of assumptions about the future all one had to do was change the assumptions, even reverse engineer, and create a competing set of claims for social pricing and access. Nevertheless this was embraced by the health technology assessment profession as evidenced by the premier textbook for inventing imaginary cost-effectiveness claims⁵. The attractiveness of the approximate imaginary information belief system is undeniable. It is not often that claims are made that can never be empirically evaluated or replicated. Indeed, in the UK, where the National Institute for Health and Care Excellence (NICE) is the ICER lodestar, there are academic institutions where forensic skills developed over many years are employed to assess the validity of the imaginary cost-per-QALY models presented by manufacturers, proposing imaginary amendments or even an alternative imaginary reality. While ICER does not expose its models to this level of imaginative inquisition, the bottom line is that such an approach is in a win-win situation: recommendations are made that are incapable of

empirical evaluation. We don't know if the modeling is right, or wrong; we will never know and we were never intended to know. All we can offer is to change assumptions including the model structure and parameter values and come up with competing cost-per-QALY claims; none of which, in turn, will be empirically evaluable⁶.

IMAGINARY ICER RECOMMENDATIONS

The report examined in this commentary considers six atopic dermatitis therapies: abrocitinib (Cibinquo, Pfizer); tralokinumab (Adtralza, LEO Pharma); baricitinib (Olumiant®, Eli Lilly and Incyte); upadacitinib (Rinvoq®, AbbVie); ruxolitinib (Opzelura, Incyte); and dupilumab (Dupixent®, Regeneron and Sanofi); the objective to propose pricing recommendation and claims for budget impact. Both sets of claims for all products are imaginary, driven by an assumption fueled five year simulation model. The centerpiece to this imaginary presentation imaginary health-benefit price benchmark (HBPB). This, is the highest imaginary price a manufacturer should charge for a treatment. The eugenic implications for access to and denial of care are clear⁷. This highest price is *based on the amount of improvement in overall health (defined by the preference score attributes) patients receive from that treatment, when a higher price would cause disproportionately greater losses in health among other patients in the health system due to rising overall costs of health care and health insurance. In short, it is the top price range at which a health system can reward innovation and better health for patients without doing more harm than good.* The fatal flaw is that the entire exercise is based on a failure to recognize the standards of normal science, notably the axioms of fundamental measurement, and a belief that the imaginary QALY can support health care allocation decisions. Health care resource allocation cannot be based on imaginary constructs and claims which have no pretense to scientific rigor. The HBPB construct is meaningless, implying as it does that some health states, from a community preference for health attributes perspective, are more 'worthy' of support than others. Just as eugenic criteria were pseudoscience, so the HBPB criteria are equally pseudoscience: they fail the demarcation test that separates science from non-science.

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

While not to be taken seriously, ICER's recommended HBPB ranges are as follows: abrocitinib, \$30,600-\$41,800 per year; tralokinumab, \$25,700-\$35,000 per year; baricitinib, \$24,400-\$33,300 per year (which would require a 0-16% discount with current US list price of \$29,000), upadacitinib, \$30,400-\$41,500 per year (which would require a 35-53% discount off the treatment's current US list price of \$64,300) and dupilumab, \$29,000-\$39,500 per year, which would require a 6-31% discount off the treatment's current US list price of \$41,800).

The attraction of applying generic preference scores to support pricing and access recommendations are that the instrument is less than sensitive to therapy impact due to the limited range of symptoms covered, which may be of little relevance to the disease area and target patient. This is seen clearly in the model where imaginary QALY equivalents over the five year time horizon range from imaginary 2.98 QALYs for standard of care (topicals) to 3.59 for Abrocitinib (the imaginary range for all comparators is 3.23 to 3.59 QALYs). Incremental QALY gains over the standard of care (topicals) range from 0.26 in the case of Baricitinib to 0.61 for Abrocitinib, with even smaller increments for Baricitinib and Upadacitinib compared to Dupilumab at 0.12 and 0.03 QALYs respectively. Translating these imaginary incremental QALYs into time gains in a five year time horizon, comparing Abrocitinib to standard of care gives incremental 0.61 QALYs or 7.32 months while if Dupilumab is the comparator the gain is 0.12 QALYs or 1.44 months (42.3 days) over five years. Incremental cost-per-QALY claims and the application are driven almost entirely by hypothetical or imaginary costs.

The QALY (or imaginary I-QALY) is a mathematically impossible construct⁸. It relies on the false belief that preference scores have ratio properties. Indeed, preference scores, although they have negative values, actually viewed as ratio scores in disguise. None of these claims are empirically evaluable and, based on ordinal scores, entirely imaginary. The distance between scores such as preference and QALY estimates are unknown. The rule of thumb is: if you want to minimize imaginary therapeutic gains expressed as imaginary QALYs then use an ordinal generic preference score. Needless to say the choice of competing ordinal preference scores with alternative manipulations will produce different results.

The ICER report says nothing about whether the need of patients is addressed; The model is driven by community preferences for a bundle of clinical symptoms and response levels defining a generic health state which may have little to do with health states relevant to atopic dermatitis populations. The notion of perfect health (preference score equal unity) is entirely contrived.

Claims for the future based on claims from the past suffer from the "problem of induction". The pervading assumption is the ability to use past claims to support or justify assumptions about an unknown future; forgetting that claims from the past

cannot support claims on the future. Justification is a psychological defense.

Unfortunately, all too many decision makers take ICER's assumptions and recommendations as if they were holy writ; whether this is just a negotiating tactic or a more concerning failure to appreciate the standards of normal science (which underpin drug development) is an open question. The downside, of course, is that patients and caregivers can be denied therapy. A recent commentary has described this as eugenics by the back door: if preferences are based, as they are, on the views of a community sample on the value of health states then we face the issue of 'worth' in the allocation of health care. We can use these preferences and ICER's modeling of QALY increments and costs to refuse care to the 'less worthy', restricting it to the more 'worthy'. More pernicious, is the presence of negative preference scores, or 'state worse than death'. The eugenic association again is obvious, but more to the point is that if there are negative scores bundled with positive ordinal scores to generate an average preference score (which is mathematically impossible as these are ordinal scores) then for these health state cost-per-QALY estimates will be inflated with smaller average preference scores but the same costs, and claims for price discounting and access more disadvantageous for that target patient population.

IMPOSSIBLE REPLICATION

Any attempt to replicate the ICER model simulation is virtually impossible given the lack of supporting information. Consider the preferences (utilities) employed in the model. Clarification on the inappropriate use of preference scores requires more information than that provided in the evidence report. Unfortunately, we have no idea as to what these scores actually are for mild, moderate and severe stages of AD. They are blanked out. All we have is assumption-driven claims for the pricing and recommendations for atopic dermatitis therapies that were '*weighted by a single set of health state utility values from pooled manufacturer data to derive quality-adjusted life-years (QALYs)*'. Seeking further clarification on these utility scores the process is described in the report:

We derived health state utilities for the non-responder and responder states by pooling utility estimates from manufacturer submitted data. We estimated weighted average utility values for each health state, combining estimates from all treatments with data available by health state. We considered therapy-specific health state utility values to capture benefit beyond EASI score, however the available evidence did not support differential utility scores by treatment (p. 42).

No further details are given; all that is provided is a list of atopic dermatitis trials. This is unfortunate because if the protocols for the various AD trials are reviewed (Clinicaltrials.gov: ECZTRA 1&2; MEASURE UP 1 & 2; AD UP; and SOLO 1&2) there is no evidence from the list of secondary outcomes for each of these

of any health related quality of life or just quality of life instrument that is designed to generate either direct or indirect preference scores. At best, we have the ordinal Dermatology Life Quality Index (DLQI) in two trials (ECZTRA 1 & 2 and SOLO 1 & 2) which simply provides an aggregate of 10 4-level Likert scales (scores 0 – 30). The DLQI creates ordinal scores in attempting to aggregate ordinal Likert values. Other than that we have no idea how utility values were created for a ratio scale with a true zero and a range of 0 = death to 1 = perfect health; an impossible undertaking. When an inquiry for clarification was made, no reply was received. We must presume, as these are all secondary endpoints for the various protocols that they were all powered to create a 'composite' utility scale.

It should be emphasized that if these various inputs from manufacturers are patient reported outcomes (PROs) with ordinal scores, then the calculations are mathematically impossible (with a further concern that they lumped together utilities from different instruments). Ordinal scales can only support non-parametric assessments. Ordinal utility estimates cannot be pooled and weighted to create average utility measures; this is mathematically impossible. This is a major concern that should be addressed with regard Hume's Problem of Induction, the rationally unfounded premise that the future will resemble the past [David Hume, 1711 - 1776] ⁹.

This can further be pointed out in the review of atopic dermatitis and subsequent modelling for Dupilumab in moderate to severe atopic dermatitis in regard to the EQ-5D-3L utility values (source Sanofi data on file) ¹⁰. For patients with moderate disease (IGA), the utilities ranged from 0.684 (baseline) to EASI 50 0.892, EASI 75 0.895 and EASI 90) 0.907 while for severe disease (IGA4) the baseline was 0.536 to EASI 75 0.535, EASI 75 0.090 and EASI 90 0.911. The results presented failed to note that the EQ-5D-3L has only ordinal properties which nullifies the analysis; as well as being inconsistent in its scales with the EQ-5D-5L.

What is puzzling is the lack of any acknowledgement that there are a range of preference scores for AD available from the literature. There appears to have been no attempt to undertake a systematic review of the QoL (HRQoL) literature in atopic dermatitis as recommended by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Even so, attempting to undertake a systematic review of preference scores with ordinal properties would actually have been a waste of time under the assumptions of this commentary.

HEALTH STATES WORSE THAN DEATH

The question of the distribution of health states defined by ordinal preference scores is not addressed in the report: the prevalence and distribution of health states worse than death are ignored. Again, this brings in the eugenics overtones which subscribes to the belief in allocating resources by health state and QALY. The fact that all generic preference instruments support negative preference scores is well established; they all

subscribe to the potential for eugenic criteria. While often glossed over, the presence of negative scores has important implications. The most obvious is that if a preference algorithm can produce negative preference health state values then it cannot claim to have a true zero. It cannot support the standard arithmetic operations (e.g., multiplication) and hence cannot support QALYs; QALYs are, as noted, mathematically impossible. A more important issue is how are we to interpret a negative preference score based on community sampling? Is health care to be withdrawn or denied?

As a first step, any application of ordinal preference scores should include a ranking of patients to indicate the proportion of patients with negative scores; attempting to claim QALY increments when a significant proportion of respondents report health states worse than death would give a misleading impression of imaginary benefits. In addition, there is the possibility that if an average preference score is presented to support QALY estimates, the presence of negative preferences could, as noted, 'deflate' the preferences and hence QALY values (although, as also noted, with ordinal scales averages cannot be computed; it's mathematically impossible). Are we to imply that the preference averages include an allowance for withdrawal or denial of health care to these 'negative' souls? As it stands, however, with ordinal preference scores this would be an invalid exercise.

Unfortunately, there is no resolution to this issue as there is no access to the distribution of preference scores that support the utility data points abstracted from the literature. It is not as though negative health states are unlikely to occur. In the case of the latest US valuation of EQ-5D-5L health states (5 symptoms, 5 response levels) we find that of the 3125 possible health state values, 625 (20%) take negative scores (range - 0.573 – 1.0) ¹¹. Indeed, one study using the EQ-5D-5L has reported negative preference scores in topic dermatitis ranging from -0.003 to -0.53 ¹².

Should the sub-group of negative preference patients be separately identified in modelling? Are they considered less 'worthy'? Should ICER avoid deflating the preference scores due to the presence of 'states worse than death'? One approach would be to use, assuming ICER has preference distributions, non-parametric ranking comparisons as one criterion for evaluating benefits before and after therapy. This is most unlikely to occur. It would actually be a waste of time as the ordinal scores, based on a bundle of health states defined by a handful of symptoms or attributes, lack dimensional homogeneity, unidimensionality and construct validity ¹³. The ordinal preference scores themselves, although there are many different ordinal ones to choose from, all lack the standards required to assess response to therapy: for one simple reason, they try to capture multiple attributes at one time rather than following the standards of the physical sciences and mainstream social sciences in focusing on one attribute at a

time¹⁴. Responses can then be evaluated across a spectrum of required attributes, each reported separately.

Without exception, studies claiming to evaluate the QoL or HRQoL in atopic dermatitis fail to appreciate that all the disease specific instruments such as the Dermatology Life Quality index (DLQI) and the Children Dermatology Life Quality Index (CDLQI) that claim to capture aspects of quality of life in atopic dermatitis together with the generic preference instruments, including comparative studies with generic preference measures including the SF-6D, the EQ-5D-5L all fail to meet the required standards of fundamental measurement^{15 16 17}.

It is worth noting that although the question of fundamental measurement was not addressed, a review of classical measurement properties concluded that only the Quality of Life Index for Atopic Dermatitis (QoLIAD) and DLQI merited further evaluation in atopic dermatitis¹⁸. The result is that, despite considerable attention given to QoL (and HRQoL) in the last 20 or more years, there are, with two exceptions [QoLIAD and the Parents Index of Quality of Life in Atopic Dermatitis (PIQoL-AD)] no acceptable measures of QoL in atopic dermatitis^{19 20}. This is not unusual in chronic disease states.

QoLIAD AND PIQoL-AD: INCONVENIENT TRUTHS

One aspect of the current ICER report on atopic dermatitis is the disregard of published and peer reviewed studies that point to a response assessment that meets the standards of normal science. The respective material was cited in an evidence commentary to ICER, but failed to materialize in the final report. As noted above, two QoL measures for atopic dermatitis meet the required measurement standards. These are the QoLIAD and the PIQoL-AD need fulfillment instruments. The QoLIAD has been revised and used to create interval scores in atopic dermatitis trials, most recently Dupilumab in moderate to severe atopic dermatitis²¹.

Given the focus on measurement theory, it should be noted that the QoLIAD and PIQoL-AD instruments apply Rasch Measurement Theory to create interval response scores consistent with these requirements. The study found that compared to mean QoLIAD scores at baseline, dupilumab significantly improved the QoLIAD score at 12 weeks of treatment against placebo. These scores were significantly correlated with changes in efficacy outcomes including EASI, 5-

dimensional pruritis, pruritis NRS, total SCORAD and SCORAD VAS scores for sleep.

Since the QoLIAD and PIQoL-AD were developed over 10 years ago, a recent innovation has demonstrated that it is possible to transform the respective scores produced by these instruments to a bounded ratio scale, the need fulfillment or N-QoL²². This allows us to evaluate the extent to which patient and caregiver need is met and the impact of competing therapies on need fulfillment. As a ratio scale, the N-QoL can also create quality adjusted life years; although it is certainly not recommended that these should be used for invented lifetime simulation models.

CONCLUSIONS

The standard ICER response to criticism is that the case for the I-QALY in assumption-driven simulations to create imaginary recommendations for pricing and access is that everyone else does it. The belief is that the ordinal generic preference score, even with well-documented negative values and lack of dimensional homogeneity, is truly a ratio measure in disguise; a view to which health technology assessment practitioners should no longer subscribe.

Simulation modeling frameworks are not just an analytical dead end, but a framework for inventing non-evaluable evidence that should never have been attempted in the first place. Emulating agencies such as NICE is not a defense for making claims for pricing and access to pharmaceuticals that are not evaluable. That measurement theory is a key standard for normal science is uncontroversial; it has been recognized for centuries. The application of quality of life to support pricing and access, let alone investment in new products in atopic dermatitis, deserve more scrutiny than this modelling approach.

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

The opinions contained in the paper are those of the author.

REFERENCES

- ¹ Atlas SJ, Brouwer E, Fox G, Carlson JJ, Campbell JD, Agboola F, Hansen RN, Pearson SD, Rind DM. JAK Inhibitors and Monoclonal Antibodies for the Treatment of Atopic Dermatitis: Effectiveness and Value; Evidence Report. Institute for Clinical and Economic Review, July 9, 2021. <https://icer.org/assessment/atopic-dermatitis-2021/#timeline>
- ² Langley P. Peter Rabbit is a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review in Health Technology Assessment. *InovPharm*. 2021;12(2):No. 2 <https://pubs.lib.umn.edu/index.php/innovations/article/view/3992/2855>
- ³ Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer reviewed] *F1000Research* 2020, 9:1048 <https://doi.org/10.12688/f1000research.25039.1>
- ⁴ Neumann PJ, Willke R, Garrison LP. A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *Value Health*. 2018;21:119-123
- ⁵ Drummond M, Sculpher M, Claxton K et al. *Methods for the Economic Evaluation of Health Care Programmes*. 4th ed. New York: Oxford University Press, 2015
- ⁶ Langley P. Let a Thousand Models Bloom: ICER Analytics Opens the Floodgates to Cloud Pseudoscience. *InovPharm*. 2021;12(1): No.5 <https://pubs.lib.umn.edu/index.php/innovations/article/view/3606/2668>
- ⁷ Langley P. Abandoning Eugenics and the QALY. *InovPharm*. 2021;12(3):No.20 <https://pubs.lib.umn.edu/index.php/innovations/article/view/4291/2939>
- ⁸ Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7 <https://pubs.lib.umn.edu/index.php/innovations/article/view/3359/2517>
- ⁹ Hume D. *An Essay Concerning Human Understanding*. 1748
- ¹⁰ Zimmermann M, Rind D, Chapman R et al. Economic evaluation of Dupilumab for moderate-to-severe atopic dermatitis: A cost-utility analysis. *J Drugs Derm*. 2018;17(7):750-6
- ¹¹ Pickard A, Law E, Jiang R et al. United States valuation of Eq05D-5L health states using an international protocol. *ValueHealth*, 2019;22(8):931-41
- ¹² Andersen L, Nyeland, Nyberg F. Higher self-reported severity of atopic dermatitis in adults is associated with poorer self-reported health-related quality of life in France, Germany, the UK and the USA. *Br J Dermatology*. 2020;182:1176-83
- ¹³ McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting patient reported outcome Instruments and Interpreting the data they produce *InovPharm*. 2021;12(2):No. 17 <https://pubs.lib.umn.edu/index.php/innovations/article/view/3911/2764>
- ¹⁴ McKenna S, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality. *J Med Econ*. 2020;23(10):1196-1204
- ¹⁵ Chernyshov P, Tomas-Aragones L, Manolache LK et al. Quality of life measurement in atopic dermatitis. Position paper of the European Academy of Dermatology and Venereology (EADV) Task Force on quality of life. *J Eur Acad Dermatology Venereology*. 2017;31:575-593.
- ¹⁶ Chiesa Fuxench Z, Block J, Boguniewicz M et al. Atopic dermatitis in America Study: A cross-sectional study examining the prevalence and disease burden of atopic dermatitis in the US adult population. *J Invest Dermatology*. 2018;139
- ¹⁷ Silverberg J, Gelfand J, Margolis D et al. Health utility scores of atopic dermatitis in US adults. *J Allergy Clin Immunol Pract*. 2018;7(4)

- ¹⁸ Heintz D, Prinsen A, Deckert S et al. Measurement properties of adult quality-of-life measurement instruments for eczema: a systematic review. *Eur J Allergy Clin Immunol*. 2016;71:358-70
- ¹⁹ Whalley D, McKenna SP, Dewar AL et al. A new instrument for assessing quality of life in atopic dermatitis: international development of the Quality of Life Index for Atopic Dermatitis (QoLIAD). *Br J Dermatol*. 2004; 24(1):105-13
- ²⁰ Whalley D, Huels J, McKenna S et al. . The benefit of pimecrolimus (Elidel SDZ ASM 901) on parents' quality of life in the treatment of pediatric atopic dermatitis. *Clin Trial*. 2002;110(6):1133-6
- ²¹ Tsianakas A, Luger TA, Radin A. Dupilumab treatment improves quality of life in adult patients with moderate-to-severe atopic dermatitis: results from a randomized, placebo controlled clinical trial. *Br J Dermatol*. 2018;178(2):406-14
- ²² Langley P, McKenna S. Fundamental Measurement: The Need Fulfillment Quality of Life (N-QOL). *InovPharm*. 2021;12(2):No. 6 <https://pubs.lib.umn.edu/index.php/innovations/article/view/3798/2697>