MAIMON WORKING PAPERS No. 1 JANUARY 2022

HOODWINKED: THE BLINKERED PERILS OF IMAGINARY LIFETIME MODELING IN TYPE 2 DIABETES

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

Abstract

Value claims for competing products in Type 2 diabetes, as well as any other disease state, must respect the standards of normal science. All value claims must be credible, evaluable and replicable, respecting the axioms of fundamental measurement as single attribute claims with ratio or interval measurement properties. These requirements have been ignored in the decision modelling in the Mt Hood diabetes challenge network and the biannual Mt Hood Challenge meetings, as well as the IQVIA-CDM model framework. At present, there are some 31 imaginary diabetes models, including earlier versions, that have been proposed as the basis for type 2 diabetes cost-effectiveness claims for an unknown future. None meet standards for normal science and fundamental measurement. Efforts over the past 20 years to reconcile the various model structures, decision modeling, simulation modeling and cohort modelling outcomes have attracted attention, with few appreciating the inherent futility of such an exercise. While the tenacity of the model builders must be admired, the problem is that none of these models meet the required standards for viable value claims that can be evaluated by protocol in the short term and reported to formulary committees. Formulary committees are asked to base decisions on invented and non-evaluable claims. These modeling efforts have been a complete waste of time. They must be rejected. The purpose of this commentary is to make the case for abandoning these models. A failure that reflects not only a failure to meet the standards of science rather than non-science (metaphysics and pseudoscience) but the problem of induction where claims from the past cannot support claims on the future, however they are dressed up as imaginary modeled claims for an unknown future.

Keywords: imaginary claims, Mt Hood Challenge failure, IQVIA-CDM failure, I-QALY nonsense

INTRODUCTION

If health technology assessment and value claims for pharmaceutical products are to be taken seriously, then the claims must meet the standards for normal science, in particular the

axioms of fundamental measurement. If not, the analytical framework will fail the demarcation test between science and non-science and are seen properly as metaphysics or pseudoscience (or bunk) [1]. There are four standards that must be met:

- **All value claims must refer to single attributes that meet the demarcation standards for normal science: they must be credible, evaluable and replicable**
- **All value claims must be consistent with the limitations imposed by the axioms of fundamental evidence: they must meet interval or ratio measurement standards**
- **All value claim assumptions must respect the logical requirement that assumptions from the past cannot support assumptions or claims on the future**
- **All value claims must be supported by a protocol that details how they will be assessed and reported in a meaningful timeframe[2]**

If value claims, notably those driven by lifetime decision models (including simulation and cohort models) cannot meet these standards (which they will not) then they should be rejected. Failing the demarcation test between science and non-science; the process of conjecture and refutation embracing sophisticated falsificationism, they are non-science, categorized as metaphysics or pseudoscience [3]. Unfortunately, the last 35 years in health technology assessment has failed to recognize these standards. In type 2 diabetes the Mt Hood Challenge models, a collection of 31 at last count (including earlier model versions) have also all failed to meet these standards. With some 20 years of Mt Hood Challenge meetings to compare these various models, no one has attempted to consider the standards of normal science and fundamental measurement as setting limits to the invention of comparative claims, model transparency and the claims for health related quality of life (QALYs); culminating in imaginary blanket claims for product cost-effectiveness.

The application of economic modeling tools is to support decision making, in this case for diabetes products, in a chronic and progressive disease, is consistent with the approximate information meme that underpins health technology assessment; hypothesis testing is put to one side in favor of inventing non-evaluable evidence that might appeal to a formulary committee [4]. Given limited data, typically at product launch, economic modeling is claimed to provide *an extrapolation of trial data to time horizons sufficient to capture the full costs and benefits of interventions (often lifetime)* [5]. Even when proposed as the basis for meeting evidence gaps, and now some 20 years later, this approximate information framework is not only in defiance of the standards of normal science, but ensures a commitment to the development of a multitude of competing models through the Mt Hood Challenge, that are essentially worthless as economic evaluations. We have no idea if the modelled claims are right, if they and wrong and we will never know, and were never intended to know. A

position in contrast to a commitment to meeting evidence gaps through a structured research program combining clinical trials and observational studies to support the discovery of new, yet provisional facts.

Creating imaginary worlds to represent an unknown future is no basis for meeting the standards of normal science to establish credible and evaluable value claims for competing pharmaceutical products. The purpose of this commentary is to assess critically the contribution, if any, that has been made by the Mt Hood Challenge meetings in health technology assessments in supporting imaginary value claims for therapies in type 2 diabetes. The required standards for non-Imaginary value claims are considered from the following perspectives:

- **The standards of normal science**
- **The axioms of fundamental measurement**
- **Impossible preference claims**
- **The impossible quality adjusted life year (QALY)**
- **The problem of induction**
- **Impossible claims for cost-effectiveness**

GUIDELINES FOR COMPUTER MODELLING OF DIABETES

The American Diabetes Association Consensus Panel for computer modelling guidelines argued for computer modeling to 'provide more informed answers to questions that have not been, or will not be answered by clinical trials' as decision making aids; inventing comparative claims to support formulary decisions. For the consensus panel transparency was the key to model acceptance [6]. Model developers should provide a complete description of the model structure, inputs, equations, algorithms, assumptions and data sources to assist in interpreting the various claims and allowing reproduction. The highest level of validation for the model was that 'it accurately predicted the results of studies that were not used to build the model'.

The Mt Hood Challenge biannual meetings were initiated in 2000. Their purpose is to provide a forum for diabetes modeling groups to compare and contrast models, methods and data in *the context of simulating standardized treatment scenarios to …. improve performance and input transparency* [7] . Following from the ISPOR-SMDM Modelling Good Practices Task Force for imaginary modeled claims, the emphasis in diabetes comparative modeling is seen as a decision aid to improve the predictive accuracy of risk equations as well as translating biomarker risk factors into economically relevant outcomes such as event rates, life

expectancy, QALYs and costs [8] . More recent efforts have focused on improving the transparency of the various imaginary diabetes models applying the Mount Hood Diabetes Challenge Reference Case and the challenge instructions for standardizing structural and input assumptions, together with Mount Hood Diabetic Challenge Network, yet make no mention of the standards of normal science, including the axioms of fundamental measurement [9] . Indeed, there appears to be no effort to evaluate the measurement status of the various model inputs, apart from preference scores which we know are ordinal, there is no assessment of whether or not other input have ratio or interval properties. The relevance of establishing coherent yet imaginary value claims, apart from attempting to mimic clinical trial results, and putting these in a form that allows empirical evaluation seems not to have been considered. Unfortunately, while it may be possible, it is not clear why the challenge members bother to 'approximate' trial results; perhaps a more useful activity would be to design trials to fill evidence gaps for the discovery of new yet provisional facts based on evaluable value claims. Yet it is argued that for those who believe in the importance of imaginary modelled claims, transparency of input data reporting apparently offers increased confidence and credibility for the acceptance of claims in formulary decisions irrespective of the model chosen As the claims for economic outcomes are entirely imaginary it is difficult to see what 'improved' transparency actually offers.

It should not be thought that these various models are an outlier. The Mt Hood Challenge modeling framework is entirely consistent with the mainstream health technology assessment approach to building imaginary lifetime models to support formulary decisions. A rejection of the standards of normal science that sets health technology assessment in the unique position of being the only social science that rejects the standards of normal science and elementary logic. If this may seem surprising, reference can be made to the principal textbook in the economic evaluation of health technologies which is, in fact, a primer for constructing imaginary worlds and imaginary claims , based in large part on imaginary or I-QALYs [10].

STANDARDS OF NORMAL SCIENCE

The Mt Hood Challenge activities and the 31 diabetes models that a have been considered at these various meetings all have one common feature: they fail to meet the standards of normal science. They are imaginary, assumption driven constructs that fail the demarcation test between science and non-science. Unfortunately, the question of demarcation is not one that has been addressed in these (to date) nine biannual challenge meetings. There appears no concept of the required modeling standards to produce empirically evaluable value claims for the economic benefits of competing products or the contribution of a structured research

program to resolve evidence gaps in the process of discovering new, yet provisional facts to support claims in type 2 diabetes. Instead the various model groups have adhered to the meme that has captured health technology assessment over the past 30 years with its commitment to inventing evidence to support formulary decisions in complete disregard of the standards of normal science and fundamental measurement. Instead they have embraced the health technology assessment meme that truth is consensus *which systematically excludes from consideration the very feature that makes scientific arguments distinctive: their appeal to superior evidence* [11].

The reason for committing to this meme is obvious; for leaders in the field of health economics (echoed by groups within the diabetes community), evidence available at product launch, notably for a chronic long term and complex disease such as type2 diabetes, is limited. The options are to invest in a program of clinical and observational studies or to invent evidence. The latter option was selected as shown by the Mt Hood Challenge meetings and the commitment to long term model building. In retrospect this was the worse decision that could have been made in its defiance of the standards of normal science, notably the axioms of fundamental measurement.

THE AXIOMS OF FUNDAMENTAL MEASUREMENT

Claims for response to therapy must recognize the axioms of fundamental measurement. 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 [12] . 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). The implications for the ability to utilize a scale to support use of arithmetic operations (and parametric statistical analysis) are clear. Nominal and ordinal scales do not support any mathematical operations; only 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 the one object is twice as long as another. To measure any object on a ratio scale it has to be demonstrated that all criteria for an interval scale have been met with the presence of a true zero. It is impossible to take an ordinal score and translate it to a ratio or interval score. If a ratio scale requirement is dictated by the need to create QALYs then that

has to be designed from the get-go in instrument development. Unfortunately, creating ratio measures for single attribute latent constructs is far from settled, requiring as the first step the creation of an interval or invariance of comparisons score which can report on response to therapy, then translated to a ratio score which is required to create the mathematically impossible QALY.

It is worth noting that for a latent construct such as quality of life (as distinct from the notion of health related quality of life [HRQoL] which is just a cluster of clinical symptoms) it is no easy task to create an instrument with the required measurement properties for disease states such as type 2 diabetes. Simply aggregating over a cluster of attributes for symptoms is impossible; to do so would require each attribute to have ratio properties [13] . In fact, while it is possible to construct an instrument with interval properties, it is only in limited instances that we can build on this to create an instrument with ratio properties. This is seen in the measurement of need fulfillment quality of life for target groups in disease areas. Applying Rasch Measurement Theory (RMT) it is possible to construct the required instrument with interval properties [14]. A recent innovation has allowed such a transformation to a bounded ratio scale. As there are no instruments in diabetes with the required single attribute interval or ratio properties, the potential, if that is the right word, for their application to support imaginary models as a need fulfillment analog of a preference score is absent.

The Mt Hood Challenge appears unaware of these requirements. The result is claims for therapy response that are meaningless. They assume ordinal scores are ratio measures which invalidate attempts to compare the various imaginary model constructs. There is no defense for this position as there were a number of red flags raised from the endorsement of RMT in the 1960s through to the present advising on the need to factor the axioms of fundamental measurement into model claims[15] . It was not until the 1960s with the application of RMT following on the contributions of Rasch and then Luce and Tukey that the importance of designing instruments to have the required measurement properties in the social science was recognized [16] [17].

IMPOSSIBLE PREFERENCE SCALES

The axioms of fundamental measurement require all scales that support statistical operations to have ratio or interval measurement properties. This simple fact appears to have been overlooked by authors of the various diabetes models participating in the Mt Hood Challenges. It should be a requirement of those building imaginary simulation or cohort diabetes models to demonstrate that this is the case for all scales applied in the construction of these models, even though the models fail the standards of normal science. Consider the

application of preference scores from both generic or disease specific instruments. There are a number of direct (standard gamble (SG) and time trade off (TT0)) as well as indirect preference instruments (EQ-5D-3L/5L, HUIMk2/3, SF-36/6D); none of these has any claim to ratio properties; they produce only ordinal scores. There are a number of reasons for this, apart from the fact that no developer of these instruments thought in terms of creating a preference score with ratio properties. The deficiencies are easily tabulated: (i) they lack a true zero (i.e., the preference score algorithms produce negative values); (ii) they lack invariance of comparisons; (iii) they are multiattribute, which means they lack dimensional homogeneity and construct validity. Apart from these, each of these instruments is different in their design and application; the resulting ordinal scores are different even though, mistakenly, attempts are made to crosswalk from one ordinal scale to another. If you wish to compare or challenge competing imaginary worlds, it would be thought advisable that they all incorporated the same ordinal preference score. As it stands none of these instruments are capable of capturing response to therapy as they lack invariance of comparison; at best they support application of non-parametric statistics for ranked observations.

The lack of recognition of the standards of fundamental measurement applies also to the various disease specific instruments that claim to capture quality of life in diabetes. These include the Audit of Diabetes Dependent Quality of Life, the Diabetes Quality of Life instrument, the Appraisal of Diabetes Scale and the Diabetes Health Profile [18]. All lack dimensional homogeneity and construct validity, they rely on aggregating over Likert or similar scales which are in fact ordinal. The resulting overall scale fails as a measure of response to therapy; with the same result with the scores for sub-domains. A recent systematic review of measuring quality of life in diabetes completely overlooks the importance of fundamental measurement and its relevance to claims for therapy response [19].

The necessity of designing an instrument to have the required measurement properties with the focus on a single attribute latent construct, is emphasized by Bond and Cox in their advocacy of RMT [14]. If we are to escape the morass of value claims that fail to meet the requirement of fundamental measurement, one avenue is to consider the application of RMT to developing value claims that meet the required evidence standards for response to therapy. It cannot be assumed, *ex post facto*, that a given scale has interval, invariance of comparison, or ratio properties. In traditional statistical theory (TST) and item response theory (IRT) the observed data have primacy; results are exploratory and descriptive of those data. Rasch models are, on the other hand, confirmatory and predictive; a confirmatory model requires the data to fit the model where, following the principles of conjoint measurement, they are sufficiently realized to claim the results are a measurement scale with

interval measurement properties while detecting measurement structures in non-physical attributes [14].

The Rasch model is designed to analyze categorical data where the likelihood of a positive response is a function of the trade-off between item difficulty and the respondent's abilities or proficiency as locations on a continuous latent variable (e.g., need fulfillment quality of life). The object in RMT is to develop an index of response, from ordinal scales, that has interval level properties. In certain circumstances this index can be translated to a bounded ratio scale. In the case of the recently proposed need fulfillment quality of life measure (N-QoL), this meets all required RMT standards as well as allowing the analyst, for the N-QoL class of attributes for specific disease areas, to develop a bounded ratio scale [20]. This allows direct comparison of this dimension of benefit as well as allowing, if required, the construction quality adjusted time spent in disease states. It, effectively, eliminates the I-QALY.

Whichever preference score a Mt Hood Challenge model employs (and the choice is not obvious), the result is the same: it is an ordinal scale which cannot capture response to therapy or support anything other than nonparametric statistics. There appears to be no discussion over the choice of preference scale for the various Mt Hood Challenge models nor any appreciation of their ordinal properties. The implications of this are quite apparent: any attempt to use preference scales and ultimately QALYs to model economic outcomes is impossible.

THE IMPOSSIBLE QALY (I-QALY)

Claims based on the I-QALY, the application of a preference score for a disease stage multiplying estimated model time spent in that disease state, are a key part of modeling economic outcomes in imaginary lifetime Mt Challenge models. If a preference score is ordinal, then it cannot support arithmetic operation, in this case multiplication. Advocates of utilizing a I-QALY as a generic measure of the benefits of therapy interventions, would require the I-QALY to have ratio measurement properties. Unfortunately, this is impossible given that all preference scores have only ordinal properties; hence the tag impossible or I-QALY [21] . An I-QALY cannot support even imaginary claims for response to therapy. This is a major oversight. As the various preference scores are ordinal scales, they cannot be used to create I-QALYs, even those stretching 25 or more years into an unknown future. Basing resource allocation policy and therapy decisions on the various modeled I-QALY claims would be, to say the least, inadvisable. Any Mt Hood Challenge model that basis its claims on QALYs is redundant. This means the continued focus in the Mt Hood Challenge on claims for cost-

effectiveness is a wasted effort; the cost estimates are equally without merit because they are non-evaluable for the time spans covered.

The modelling endeavors for type 2 diabetes are not unique; for 30 or more years there have been thousands of imaginary lifetime modelled claims published all of whom are equally redundant. Cost-per-QALY claims are similarly disallowed. Imaginary cost-per-QALY claims will not support means and confidence intervals, least of all a cost-effectiveness plane presentation and attempts at probabilistic sensitivity analysis, as well as fanciful claims for net monetary benefit. Modelling to create imaginary non-evaluable claims stretching decades into the future, however we try to dress it up with simulation and cohort modelling, protected by sensitivity assessments, is not science. ISPOR might claim a role for advocating approximate information instead of a research program supported by hypothesis testing, but we have no idea of what the modeled information is supposed to be approximate to, possibly an unknown future reality based on assumptions stretching over decades.

THE PROBLEM OF INDUCTION

There is a curious belief among model builders, including the various diabetes models, that assumptions about the future can be categorized by their 'degree' of realism based upon past observations. Presumably, differences in assumptions reflect the belief among the model builders that one set of assumptions is more representative of an unknown future than another, even though the future is unknown. If assumptions that support the various diabetes models are selected on this basis then this reflects a failure to recognize the problem of induction, first raised by David Hume [1711-1775] in 1748 [22]. The problem is easily stated: claims from the past cannot support claims on the future. As Magee makes clear: *The whole of our science assumes the regularity of nature – assumes the future will be like the past in all those respects in which natural laws are taken to operate – yet there is no way the in which this assumption can be secured. It cannot be established by observation since we cannot observe future events. And 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*. Certainly the model builder may 'prefer' one assumption over another; but this reflects the psychology of the model builder, irrespective of the size of the 'bucket' of prior observations 'supporting' an assumption or claim on a hypothetical future; nothing else.

In lifetime simulation modeling where there is no logical claim for the 'superiority' of one assumption over another the result is, inevitably, an embarrassment of model outcomes. No one model, even the gold standard IMS CORE (IQVIA-CORE) model falls in this category [23]. It

is, after all, just one more of many assumption driven imaginary simulations; prospectively one of even more. A model may claim to be an accurate representation of the system it is designed to emulate, as defined typically by long-term clinical trials, but it will fail to provide, in any simulation, claims that have the potential to be evaluated empirically for specific therapy interventions. Claims that the IQVIA-CDM can be applied to estimate (under its choice of assumptions) the effect of interventions on clinical outcomes and a range of economic analyses is irrelevant; it is just one of a number of competing simulations with impossible outcome clams. At the same time, claims may also be mathematically impossible. An obvious example here are claims based on QALYs; as the QALY is an impossible mathematical construct claims for QALY adjusted life years, incremental cost per QALY and probabilistic sensitivity analyses should never be attempted. The assertion that IQVIA-CDM *can be used for multiple purposes to estimate the impact of interventions and cost outcomes as well as a range of economic analyses (cost-effectiveness, cost-utility, cost benefit or cost of disease* is just wishful thinking [23].

The question, for those who believe in lifetime simulations, is whether it is possible to make order from this multiplicity of model claims through ensuring transparency and providing check lists for model inputs; with, no doubt, dinner rolls lobbed across high table as favored assumptions are challenged, is whether there is a possible solution. If a model group wishes to argue that an assumption is validated as a claim on the future, then it has to demonstrate that it has knowledge of all possible instances; to claim all swans are white then it has to be able to demonstrate that all swans both deceased and currently alive globally are definitely white. Clearly an impossible standard.

The issue of induction, and its link to the now discredited logical positivism myth, leads to the issue of validation versus falsification; where the latter is taken by Karl Popper [1902-1994] as the demarcation line between science and non-science [24]. The notion of falsification or, to be more precise, Imre Lakatos's [1922-1974] 'sophisticated falsificationism' is absent from the Mt Hood Challenge debates in the focus on transparency of imaginary model structures and the jumble of competing imaginary claim [25]. The notion of the discovery of ne facts is absent.

MODELING IRRELEVANCE

Despite the efforts and the accolades received by the various diabetes models, to include the widely used IMS CORE model now rebadged to the IQVIA CDM or core diabetes model, they are irrelevant; they all fail the standards of normal science in failing to propose credible, evaluable and replicable value claims. The various My Hood Challenges, focusing on

transparency and assumptions, add nothing to value claims for therapy options and the contribution of innovative therapies in type 2 diabetes.

As a master class in irrelevance, consider the Eighth Mt Hood Challenge where modeling groups attempted to reproduce the results of two published studies, the Baxter study applying the IQVIA- CDM model to estimate the impact of improvements in glycemic control on the cumulative incidence of microvascular and macrovascular complications, and the UK Prospective Diabetes Study 72 (UKPDS-OM) which used the UKPDS Outcomes model (UKPDS-OM) version 1 to evaluate the cost-utility of intensive vs. conventional blood glucose control [26]. The results are not unexpected. In the Baxter reproduction for average cost reductions and complications avoided, none of the comparisons for four participating modeling groups consistently matched the IQVIA-CDM model results. If there was any intent to build an outcomes case for comparative therapies on the microvascular and macrovascular predictions (which stretched out to 25 years) then the sheer magnitude of the variability between the models (e.g., at 25 years avoidance of cases of eye disease ranged from 34,701 in the Cardiff model to 942,337 in the ECHO-T2RM model). The IVIA-CDM model estimate was 250,768 cases avoided.

The second assessment, evaluating cost-utility of intensive versus conventional blood glucose control should never have been attempted as the preferences (utilities) are, of course ordinal. In any event the imaginary results were just as jumbled with differences in total I-QALYs estimated at 0.27 for the UKPDS 72 modeling compared to a range of 0.15 to 1.17 I-QALY differences for 7 competing models. Differences in total costs (compared to 1,349 for the UKPDS 72) ranged from -1,537 for the IQVIA-CDM to 2,261 for the UKPDS-OM version 2.

The challenge report concluded, not surprisingly, that *the transparency challenge illustrated substantial difficulties in reproducing study results using published input data.* A more cynical observer might comment that why should anyone be interested in some 31 different models each producing their own imaginary claims even if they all meet the input checklist for transparency.

The Ninth Mt Hood challenge continued with these comparative assessments; in this case for the predictive accuracy of 10 diabetes models to predict 'hard outcomes' in two recent cardiovascular trials. The question addressed was whether recalibration could improve replication. Once again the results were, from the modeling perspective, disappointing. Commonly used risk equations were unable to capture the events, with an improvement with recalibration. The conclusion was that for other settings, time horizons and comparators new methods were required and/or new risk equations for capturing cardiovascular benefits.

It is worth noting that, as an outcome of the Ninth Mt Hood Challenge a study was undertaken to assess the extent to which bias may be present in the cohort modeling approach compared to micro-simulation using the IHE-DCM cohort model and the ECHO-T2DM micro-simulation model [5]. The metric outcome considered included absolute and incremental costs, I-QALYs, event rates and cost-effectiveness. To the extent that decision makers, recognizing the impossible or I-QALY, would take these model results seriously, these imaginary comparisons the cohort model produced larger estimates of absolute life years, I-ALYs incremental cost-effectiveness ratios, net monetary benefits and costs. Bias, for these imaginary metrics as reflected in cost-effectiveness claims was not considered consequential. This entire analysis is, of course, as emphasized, a waste of time as the health outcomes, discounted or otherwise, are meaningless.

Whether the hoped for effect of improved transparency and the role of an input check list is supposed to minimize these differences is unclear. Even if future Mt Hood Challenges manage to resolve these issues, there is still the issue of the presence of competing models and the basis (if any) for choosing one model over another. If there is a 'gold standard' reference point for modelling, it must exist in some other domain free from the constraints of the standards of normal science. Information may be claimed to be approximate, but is not clear to what it is approximate. Even if transparency was achieved for the various models, we can go no further in modeling for cost-outcomes claims. Indeed, the concept of transparency is unclear; when can we say we have achieved the sought after level of transparency for model structure and assumptions? None of the output claims from the various Mt Hood models are remotely empirically evaluable; indeed, they have not been designed to be evaluable. This approach is consistent with the dominant meme in health technology assessment where hypothesis testing is put to one side in favor of approximate information, supported by the leading textbook in health technology assessment as a primer for creating imaginary claims and by organizations such as the Institute for Clinical and Economic Review (ICER) in the US [27].

CONCLUSIONS

ISPOR in its advocacy of good research practices claims advocating a central role for invented evidence has performed a major disservice [28]. Health technology assessment, as detailed in ISPOR's good research practices has ensured that this is a discipline which has turned its back on the scientific method, advocating instead a commitment to invented evidence [29]. Unfortunately, this has been uncritically accepted by many who lack the training to recognize the importance of the standards of normal science. Proposing standards to replicate model claims seems absurd when the model claims in the first instance are impossible; a point that was made by the present author and a colleague some years ago [30].

As the Mt Hood Challenge models are consistent with the now discredited health technology assessment meme dedicated to the construction of lifetime modeled imaginary claims, it is not clear what the last 20 years has achieved. From the perspective of the standards of normal science and the application of standards required consistent with the axioms of fundamental measurement, the Mt Hood Challenge claims are irrelevant. If value claims for competing therapies, in type 2 diabetes or other therapy areas are the focus, then we must insist of single attribute claims that meet the standards for normal science as opposed to metaphysical and pseudoscientific claims that characterize the Mt Hood Challenges. In other words, to fall back on an evidence based and a more coherent way of presenting and evaluating value claims as part of an agreed and ongoing research program in therapy evaluation for type 2 diabetes.

Supporting the Mt Hood Challenge program, including the IV QIA-CDM model requires us to reject the standards of normal science, insisting that invented claims for competing products in what is admittedly a complex disease area resources are to be allocated, including the acceptance or rejection of therapies for type 2 diabetes, then modeled claims supporting this allocation in therapy choice, must be clinically credible and valid for the target patient populations. These models fail on both counts. These various models in their invented evidence must be rejected by health care decision makers in terms of the health economic outcomes which, with the focus on the I-QALY, are meaningless. Unfortunately, even if we accept this contribution of non-science there is no obstacle to number of imaginary modeled claims we can create. At present there are some 30 plus models that have appeared in the Mt Hood Challenge events. We have no idea if these various claims are right or wrong, we will never know and, to be cynical, we were never intended to know. ISPOR in its advocacy of good research practices in claiming a role for invented evidence has performed a major disservice. Formulary committees, physicians and patients certainly deserve better. It is foolhardy to believe that formulary committees would accept invented claims from assumption driven simulations or cohort designs that stretch decades into the future. A problem made all the more vexed with the competing number of models jostling for an imaginary pole position; with the opportunity for new models and equally impossible claims on the horizon for decades ahead. It is doubtful if a centenary Mt Hood Challenge meeting could provide any relief.

Given its apparent popularity in the diabetes modeling community, special mention should be made of the IQVIA-CDM (formerly IMS-CORE) model. As detailed in this commentary this model exhibits the same mistakes as all other models. The model builders and those contributing to supporting it in its on-line version appear to have no concept of the standards of normal science and, more specifically, the axioms of fundamental evidence. The most

glaring error is to assume that the I-QALY is a feasible mathematical construct when preference measures only have ordinal properties. Claims that it can provide estimates of lifetime I-QALYs, incremental cost-per-I-QALY assessments and probabilistic sensitivity analysis are just pie in the sky. It might present them but the results are meaningless. Again it is just one of many assumption driven simulations that lack credibility in attempting to apply economic analysis to invented clinical outcomes.

Comparing claims from imaginary constructs that involves an arbitrary choice of assumptions, denying simple logic, is no way forward to the discovery of new yet provisional facts in diabetes therapy, supported by a commitment to 'sophisticated falsificationism'. If the effort that has been put into the creation of imaginary worlds through the Mt Hood Challenge meetings had been applied to a research program that recognized evidence gaps we will probably be further ahead than we are now in creating such a strong evidence base for type 2 diabetes. Given the commitment to long-term clinical trials, it is difficult to see what the Mt Hood Challenge contributes by focusing on models that try to replicate these trial outcomes.

Manufacturers, who may have supported the program of Mt Hood Challenges over the years, have an embarrassing choice. Which model do they choose to support claims for their product? Or do they put the models aside and opt for value claims which meet the standards of normal science, respecting the axioms of fundamental measurement. The Mt Hood Challenges are, unfortunately, a classic example of the commitment in health technology assessment to inventing impossible claims to support formulary decisions. This may seem an unreasonably harsh judgement but there is little doubt that the model builders for these various type 2 diabetes cohort and simulation models have had, and continue to have, no concept of the standards of normal science; of the demarcation between science and non-science. Their embrace of non-science, of metaphysics and pseudoscience, has devalued the Mt Hood Challenges to nothing more than a debate over competing imaginary models, risk equations and randomized clinical trials; a debate which looks certain to continue over the next 20 or even 100 years.

REFERENCES

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

[2]Langley P. Value Assessment, Real World Evidence and Fundamental Measurement: Version 3.0 of the Minnesota Formulary Submission Guidelines. *InovPharm*. 2020;12(4): No. 12
https://pubs.lib.umn.edu/index.php/innovations/article/view/3542/2613

[3]Langley PC and McKenna SP. Measurement, modeling and QALYs. *F1000Research 2020*, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[4] 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

[5] Willis M, Fridhammar A, Gundgaard J et al. Comparing the cohort and mict-simulation modeling approaches in cost-effectiveness modeling of type 2 diabetes mellitus: A case study of the IHE diabetes cohort model and the economic and health outcomes model of T2DM. *Pharmacoeconomics*. 2020; 38:953-969

[6] American Diabetes Association Consensus Panel. Guidelines for Computer Modeling of Diabetes and its Complications. *Diabetes Care*. 2004;27: No. 9

[7] Si l, Willis M, Asseburg C et al. Evaluating the ability of economic models of diabetes to simulate new cardiovascular outcomes trials: A report on the Ninth Mount Hood Diabetes Challenge. *ValueHealth*. 2020;23(9):1163-1170

.

[8] Caro JJ, Briggs AH, Siebert U, et al. Modeling good research practices - overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Value Health*. 2012;15(5):796-803.

[9] Kent S, Becker F, Feenstra T et al. The challenge of transparency and validation in health economic decision modelling: A view from Mt Hood. *Pharmacoeconomics*. 2019;37:1305-1312

[10] 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

[11] Wootton D. The invention of science: A new history of the scientific revolution. New York: Harper Collins, 2015

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

[13] McKenna S, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality. *J Med Econ*. 2020; 23(10):1196-1204

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

[15] Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*. 1989;70:308-12

[16] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960

[17] Luce R, Tukey J. Simultaneous conjoint measurement. A new type of fundamental measurement. *J Math Psychol*. 1964; 1(1);1-27

[18] Nair R, Kachan P. Outcome tools for diabetes-specific quality of life. *Canadian Fam Physician*. 2017;63:e310-e315

[19] Palamenghi L, Carlucci M, Graffigna G. Measuring the quality of life in diabetes patients: A scoping review. *J Diabetes Res*. 2020;ID5419298

[20] Langley P, McKenna S. Fundamental Measurement: The Need Fulfillment Quality of Life (N-QOL). *InovPharm.* 2021;11(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/3798/2697

[21]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

[22] Hume D. An Enquiry Concerning Human Understanding. 1748

[23] IQVIA-CORE Diabetes Model (IQVIA-CDM). https://www.core-diabetes.com/Index.aspx?Page=About

[24] Popper K. Logik der Forschung. 1934 translated as The Logic of Scientific Discovery. London: Hutchinson 1959 with second edition 1968.

[25]Lakatos I, Worrall J, Curry G (eds). The Methodology of Scientific Research Programmes Vol 1: Philosophical Papers. Cambridge: Cambridge University Press, 1980

[26] Palmer A, Si L, Tew M et al. Computer modeling of diabetes and its transparency: A report of the eighth Mount Hood Challenge. *ValueHealth*. 2018;21:724-33

[27] Langley P. Supping with the Devil: Belief and the Imaginary World of Multiple Myeloma Therapies Invented by the Institute for Clinical and Economic Review. *InovPharm*. 2021;12(3): No. 6
https://pubs.lib.umn.edu/index.php/innovations/article/view/4215/2937

[28] Caro J, Briggs A, Siebert U et al. Modeling good research practices – overview: a report of the ISPOR-SMDM Modeling Good Practices Task Force – 1. *Med Decis Making*. 2012;32:66777

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

[30] Langley PC, Rhee TG. Imaginary Worlds: The status of simulation modeling in claims for cost-effectiveness in diabetes mellitus. *InovPharm* . 2016;7(2): No. 17.
https://pubs.lib.umn.edu/index.php/innovations/article/view/440/435