INNOVATIONS in pharmacy

Volume 7 | Number 2 Article 9

7-1-2016

Imaginary Worlds: Modeled claims for cost-effectiveness published in PharmacoEconomics January 2015 to December 2015

Paul C. Langley
University of Minnesota, langley@maimonresearch.com

Imaginary Worlds: Modeled claims for cost-effectiveness published in PharmacoEconomics January 2015 to December 2015

Cover Page Footnote
I would like to thank T G Rhee for his assistance in collating the papers to be reviewed

Imaginary Worlds: Modeled claims for cost-effectiveness published in *PharmacoEconomics* January 2015 to December 2015

Paul C. Langley, PhD
College of Pharmacy University of Minnesota

Abstract
The purpose of this review is to assess whether or not economic evaluation studies published in PharmacoEconomics in 2015 meet the standards of normal science. Two questions are key to the assessment: (i) did the authors attempt to generate testable claims as to the impact of the pharmaceutical product in health care systems and (ii) did the authors suggest how the claims might be evaluated? A total of 31 studies were evaluated, including 14 research articles, 8 systematic reviews and 9 reviews. Although the majority of the studies met recommended standards for cost-effectiveness analysis, none met the standards of normal science. They were best categorized as imaginary worlds or thought experiments. The reader has no idea whether the claims for the products are right or even if they were wrong. Journal editors have two options: (i) require authors to submit cost-effectiveness claims that are evaluable with a protocol suggesting how the claim may be evaluated or (ii) continue to publish non-evaluable cost-effectiveness claims but insist authors indicate that the claims are non-evaluable.

Introductions
A number of recent publications have raised doubts as to the status of modeled cost-effectiveness claims in pharmacoeconomics [1 2 3 4]. The primary concern has been with the construction of modeled thought experiments (or imaginary worlds) to support claims for comparative cost-effectiveness. Rather than the development of models to generate testable claims or predictions for the anticipated impact of new products and devices in health care systems, the modeled claims are seen as an end in themselves. Unfortunately, readers are not advised, as a matter of course, that non-evaluable cost-effectiveness claims should not be taken at face value.

If a modeled claim is impossible to assess or if the model fails to generate testable claims, then the model fails to meet the standards of normal science. These standards are absolute and have been recognized since the 17th century. The core elements being: (i) the construction of empirically evaluable coherent theories and (ii) the testing of hypotheses through experimentation or systematic observation.

Empirical testability differentiates science from non-science or pseudoscience. Irrespective of whether or not the authors of a model argue that it is a reasonable reflection of reality, a correspondence that is sufficient and necessarily entails the claims made, the absence of testable claims means that the model should be put to one side. This may involve a reconsideration of the model to assess whether or not it is capable of being recast to generate testable claims. If the model is incapable of generating testable claims then it should be rejected.

In the absence of experimentation or observation, a formulary committee has no idea whether modeled claims are right or even if they are wrong. To an unknown and unknowable extent the claims may be misleading and even potentially harmful. Acceptance of the standards of normal science is in contrast, therefore, to a postmodern or relativist position which holds that if the model intends to reflect reality then, given the consensus view within the profession, we should accept the claims made and to factor them, even though they are impossible to evaluate, into formulary decisions.

While a debate over philosophical positions may seem something of a stretch when we address issues of cost-effectiveness modeling, the acceptability or otherwise of modeled claims is critical to an assessment of the worth of pharmacoeconomic modeling to support value claims. If we conclude that modeled yet untestable claims should be rejected as a basis for decision making, then we face the prospect of rejecting much of the modeling endeavors over the past 25 to 30 years as well as guidelines for good practice and recommendations for formulary submissions. If we accept the relativist position that modeled claims, even if untestable, are still credible as thought experiments or imaginary worlds then we run the risk of losing status as a 'science' in our rejection of the standards of normal science that underpin our belief in the discovery of new facts.

Corresponding author: Paul C Langley, PhD
Adjunct Professor
College of Pharmacy University of Minnesota
Phone: 520-577-0436
Email: langley@maimonresearch.com
Web: www.maimonresearch.com

The purpose of this review is to evaluate the cost-effectiveness or economic evaluation studies that have been published in *PharmacoEconomics* in the period January 2015 to December 2016. This evaluation is part of an ongoing program at the College of Pharmacy, University of Minnesota exploring the credibility of cost-effectiveness claims; to see whether published claims meet the standards for falsification and replication that are the core of the scientific method.

Methods

A systematic review, following the PRISMA-P checklist (MeSH terms 'cost', 'cost-effectiveness', 'QALY'), of all papers published in *PharmacoEconomics* in the period January 2015 to December 2016 identified 31 economic evaluation studies [5]. These comprised 14 original research articles, 8 systematic reviews and 9 reviews. These studies are detailed in Table 1.

In order to judge whether the modeled claims presented or reviewed met the standards of normal science, four questions were considered:

- Is the model capable of generating testable claims?
- Did the author(s) attempt to generate testable claims?
- Did the author(s) suggest how the claims might be evaluated?
- Did the author(s) caution readers as to the implications of generating non-testable claims?

A testable claim was defined as one that could be evaluated either experimentally or observationally in a timeframe relevant to the needs of a formulary committee (ideally a period of 2 to 3 years). This period was chosen because a testable claim was seen as provisional; a condition established in the WellPoint formulary guidelines issued almost a decade ago. [6 7] . A product or device could, in this context, be accepted by a formulary committee for provisional listing, but subject to an agreement with the manufacturer to report back to the committee with evidence to support the claims made. These claims could be for product comparative effectiveness, for the impact of the product on resource utilization or some combination of these to support a claim for incremental cost-effectiveness. The claim for comparative effectiveness could encompass clinical endpoints as well as those captured as patient reported outcomes.

The important point to note is that the modeled claims were not to be judged on the reasonableness or otherwise of the assumptions of the model; a point made recently by Ellis and Silk in reference to claims in string theory modeling [8]. The fact that the claims could not be tested led to those supporting string theory to argue that the inherent elegance of a model should be sufficient for its acceptance, without need to evaluate any testable claims. Ellis and Silk pointed out that this was unacceptable if the standards of normal science were to hold.

Certainly a cost-effectiveness model would be expected to cover comparator products, or least the key comparators, and to identify the target population for the claims. But this does not mean a model should necessarily conform to recommended standards from organizations such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) or to items such as the CHEERS checklist [9 10 11 12 13]. While the CHEERS checklist has been embraced by a number of journals, the point is that neither ISPOR recommended standards for good practice or the CHEERS checklist have endorsed the standards of normal science in accepting the central role of claims assessment. In this review there was no attempt to censor or categorize the studies reviewed by such criteria, although given that *PharmacoEconomics* subscribes to the CHEERS checklist, it was assumed that these criteria would have been addressed as part of the peer review process.

In judging whether or not a model might support testable claims, even if the possibility was not considered by the authors(s), three characteristics of the model are important. These are (i) the modeling framework, (ii) the choice of primary outcome measure; and (iii) the time frame for the model. A Markov or discreet event simulation model with a lifetime perspective and with discounted cost per quality adjusted life year (QALY) claims as the primary endpoints would be one that would be impossible to evaluate. Against this, a simple, trial-based decision model with a timeframe of 12 to 18 months with claims expressed in clinical terms (including PROs) and resource utilization endpoints would, given access to readily available data sources in the US, be open to hypothesis testing and feedback to a formulary committee.

In the case of the systematic reviews and reviews, the focus was on, not the individual papers but on the whether or not the reviewer took into account the issue of whether the models presented in the papers under review were capable of generating testable claims, whether any attempts were made to evaluate claims and whether the authors made any suggestions as to how the claims made might be evaluated by, for example, a formulary committee as part of ongoing disease area and therapeutic reviews. A number of these reviews utilized a checklist to evaluate the 'quality' of the study, such as the CHEERS or the Quality of Health Economics Studies (QHES) checklist [14]. The QHES checklist does not consider either the ability to generate testable claims or proposals for how claims might be evaluated as part of a quality assessment.

A study may be judged 'high quality' by application of the QHES or CHEERS checklist but still may fail to meet the standards of normal science. It may meet the National Institute for Health and Care Excellence (NICE) reference case standards, but as a result fail to provide operational content in meeting the standards of normal science [15]. Irrespective of the internal 'mathematical elegance' of the cost-effectiveness model, it would be considered by the standards of normal science to be what is described here as an imaginary world or thought experiment.

Results

Original Research Articles

The primary outcome measure in 13 of the 14 original research articles was incremental cost-per-QALY. Subsidiary endpoints included costs, years of life gained and events avoided (e.g., falls). In a number of the articles discount rates were applied to both costs and outcomes to yield discounted cost-per-QALY claims. The majority of the papers, although not necessarily stating it explicitly, attempted to follow NICE reference case standards or those proposed by ISPOR for good modeling standards in generating estimates of direct medical costs and utilizing a generic QALY measure. None of the models utilized measures of generic QALYS generated directly from the comparative products' RCTs. The only paper that reported cost-per-QALYs from a trial generated these indirectly from the SF-12v2 instrument utilized in the trial [24]. Other studies mapped utility values from other instruments or captured utilities from studies in 'similar' product scenarios.

With the exception of one paper, all utilized a Markov state-transition model framework. The timeframes for the various state transition models ranged from 5 years to the lifetime of the modeled cohort. Seven of the articles employed a time frame of 30 years or longer (4 the lifetime of the cohort) and 4 articles reported on a 10 year timeframe. In the case of the lifetime models, three of these focused on either older patients (with limited lifespans) or those whose health states were in the end stage of the disease. The youngest lifetime cohort at age 40 years was modeled for dermatomysotis, with 56 years for hyperphosphataemia, 70 years for atrial fibrillation and 84 years for injury prevention.

In respect of the questions raised for this review, none of the studies presented considered how the claims made might be evaluated. None of the studies presented testable claims. There were no discussions of how these modeled claims might be factored into formulary decisions. Discussions on the limitations of the studies focused on the deficiencies in the evidence base used to populate the model parameters, the presence of significant evidence gaps, the lack of direct generic QALY measures for the target populations, the difficulty of translating costs estimates across health systems and the degree of uncertainty attached to the modeled claims for comparative effectiveness.

Systematic Reviews

Five of the 8 systematic reviews utilized a quality assessment checklist to grade the individual studies: 3 papers used the CHEERS checklist while 3 utilized the QHES checklist. Zhang et al concluded that of the 53 studies evaluated in psoriasis the majority were of low quality. In their view high quality studies should apply a reasonably long time horizon (with 30% adopting a time horizon of < 1 year), adopt a valid and comparable effectiveness measure (QALY), consider all relevant cost items and conduct a sensitivity analysis [31]. The question of whether or not the model should be capable of generating testable hypotheses was not considered. Hiligsman et al in their review of postmenopausal osteoporosis drugs also reported on an insufficient quality of reporting for several articles [30]. These included methods for identifying and synthesizing clinical effectiveness data, the description of the population and methods used to value preference based outcomes (all but one of the studies used QALYs as the outcome measure) and all analytic methods supporting the evaluation. Again, the question of whether or not the model should be capable of generating testable hypotheses was not considered. None of the reviews raised any questions as to the appropriateness of Markov models in generating testable predictions and the implications for evaluating model findings where, for example, the timelines of the models (9 out of 20 Markov simulations modeled claims for 40 years or more).

The most frequently cited outcomes were QALYs and life years gained (LYG). In postmenopausal osteoporosis, 38 out of 39 studies used QALYs; in psoriasis PASI (Psoriasis Area and Severity Index) were reported in 26 studies and QALYs in 15 studies (out of a total of 53 studies) [30]; in diabetes all 11 studies reported QALYs [34]; in tuberculosis QALYs were the outcome measure in 15 studies and TB cases prevented in 7 studies (out of 24 studies) [35]; in gout the humanistic burden was reported as either generic or disease specific HRQoL [36]; and in ovarian cancer, out of 28 studies, 15 studies reported cost per LYG and cost per QALY in 13 studies [38]. Once again, the question of testable claims was not raised or even how differences in comparative cost-per-QALY claims within disease or therapeutic areas or for individual products across the various studies might be resolved. The issue of whether QALYs could actually be evaluated or their absence in administrative claims or other health data sets was not raised. Any comparison is made the more difficult, presumably, given the disparate instruments and matching (or crosswalking) techniques employed in generating utility scores. There was no discussion of how these disparities might be resolved.

Review Articles

Eight of the 9 review articles reported the findings of NICE single technology appraisals (STA) [39 40 41 43 44 45 46 47]. Under the STA process manufacturers are invited to submit a clinical and cost-effectiveness case for a product in seeking a recommendation within its marketing authorization. The role of the Evidence Review Group (ERG) is to evaluate the submission, possibly seek additional input from the manufacturer, and if necessary develop its own model and report to the NICE Advisory Committee. The ERG report is considered by the NICE Appraisal Committee, with (after possibly further review) a final recommendation, and a NICE guidance. Outside of NICE, there is no independent review of findings.

The focus of the ERG's evaluation of single product submissions is on the validity of the underlying clinical claims that are made, typically the techniques applied to indirect product comparisons, and the structure and assumptions of the submitted cost-effectiveness model. The critical questions are: (i) whether the evidence submitted conforms to the methodological guidelines issued by NICE; (ii) whether the company's interpretation of the evidence is appropriate; and (iii) whether there are other evidence sources or alternative interpretations that could be useful. In respect of the reference case cost-effectiveness model the ERG is required (i) to comment on the robustness and accuracy of the model; (ii) the data used to populate the data used to populate the mode; and where possible (iii) to carry out a sensitivity analysis.

None of the eight STAs reviewed gave any recognition to the possibility of modeled claims generating testable predictions, as to how possible claims might be evaluated or whether recommendations might be revisited if additional evidence emerges that challenged the ERG summaries of clinical and cost-effectiveness claims. Instead, the focus of the STA is on checking and challenging modeled assumptions. These challenges range from the choice of clinical data, the techniques applied for indirect comparisons on clinical effect, the relevance of trial protocols to the target UK population, the choice of model, the choice of health states, the state transition probabilities, cycle lengths, application of survivorship techniques, the measurement of utilities and the scope of resource units and costs. The model is checked against the gold standard of the reference case.

While standards for developing evaluable predictions, falsification and replication are outside the reference case remit, three off the products reviewed had the potential for generating comparative evaluable claims. These are: alteplase for ischaemic stroke; aflibercept for metastatic colorectal cancer and ipilimumab for unresectable malignant melanoma [41 43 47]. The models proposed could be modified to generate assessable claims, given the expected survival profiles in these late-stage interventions at (say) six or 12 months.

Discussion

There is no doubting the popularity of modeled claims in the health technology assessment literature; models which in the majority cases conform to the standards proposed in checklists such as CHEERS and QHES. Unfortunately, the conformity in these models to what are perceived as 'quality standards', exemplified in the NICE reference case, means (i) they fail to meet the standards of normal science and (ii) are unlikely, outside of groups such as NICE in the UK, the PBAC in Australia, the AMCP in the US and CADTH in Canada to be of interest to formulary committees and other health system decision makers [48 49 50].

The accepted models can extend, in the case of chronic disease, for the lifetime of the patient cohort. Commonly applying Markov models or the more mathematically complex discreet event modeling, the analyst presents results in terms of the recommended gold standard outcome measure of lifetime quality adjusted life years saved, claiming benefits from one product over another. The models are justified by their ability to 'reflect reality', a present and future reality of 'what is', in choice of target population characteristics, treatment arms, assumed resource utilization and costs and outcomes defined by constructed quality of life indices. The models rely for their appearance of 'reasonableness' on their foundation in disparate (yet peer reviewed) literature sources and results reported for RCTs. With due account taken in the modeling technique of parameter and outcome uncertainty, the models are presented as evidence for the comparative effectiveness of competing therapies.

Can this practice be justified? Can we make claims for the comparative impact of competing products that might extend decades into the future and expect them to be taken seriously? Could we argue, for example, that the modeled claims 'reflect reality' and that they should be considered as equivalent to modeled claims that generate testable hypotheses which can be evaluated from existing evidence; an appeal to the facts?

The Standards of Normal Science

The fact that the modeled or simulated claim is defended on the grounds that it 'reflects reality' or that it is 'reasonable' in is representation or correspondence to the target treating environment and the anticipated impact of competing products and devices in not, unfortunately, a justification for accepting the model and claims generated as supporting coverage decisions. If a formulary committee is to consider modeled claims as a basis for a coverage decision then the model claims need to meet the standards of normal science:

the claims should be capable of experimental or observational evaluation. If not, the model and the claims should be put to one side as 'not fit for purpose'.

If a model is said to 'reflect reality' the obvious questions are 'what is reality' and what is a 'reflection'? From the perspective of the model builder (or collaborative group of model builders), the reality they perceive is presumably the state of things that they think actually exists and are 'expected' to exist over the time frame of the model (an imaginary future); their belief in the simulation's correspondence with the real world. If the correspondence is sufficient, then the outcomes claimed are necessarily entailed.

Unfortunately, no two groups may share the same vision of correspondence with the real world. The NICE Evidence Review Group (ERG) may not agree with the manufacturer's submission. Their different realities may generate different models which, while subscribing to the same set of standards, may result in quite different non-testable claims for the superiority of the same competing products. This presents a quandary to a formulary committee where competing models jostle for attention. Whose claims should be accepted? Should the formulary committee attempt to set 'acceptable' modeling parameters, as in the NICE reference case, in order to ring-fence modeling options? Even so, there is still the possibility that different groups will propose different models. Indeed, manufacturers may support competing models that meet common standards. Both are justified on the grounds that they 'reflect reality'. Both are defended on the grounds that they meet the required standards of the commonly held belief system.

Feedback, Information, Evidence

As noted above, neither the CHEERS nor the QHES checklists address the question of whether or not the studies evaluated are capable of generating testable predictions. There is no concept of how, through evaluating claims, new facts might be uncovered. There is also no concept of how claims evaluation may generate feedback to formulary committees and other health decision makers. The view seems to be that this approach is too difficult, time consuming and of little interest to decision makers; building imaginary worlds which have little if any chance of creating testable predictions is the easy option.

Another argument for 'accepting' modeled claims even if they fail to generate testable hypotheses is that we have limited information. Even though the US is well served in access to health data, ranging from administrative claims, possibly linked to laboratory data, together with electronic medical records from in-patient and ambulatory environments, the rest of the world is less well served. Our ability to assess claims for competing products and devices, particularly if we wish to capture patient reported outcomes as a primary endpoint, means that we may either to invest significant resources in data capture in targeted treatment settings or we have to rely on a less resource intensive approach to supporting claims. As a result, it could be argued, we fall back to a 'needs must' justification with our belief in comparative modeled claims driving our research agenda and formulary decisions.

The NICE reference case is clearly not designed to generate empirically evaluable claims. It rejects the standards of normal science. Rather, subject to the ministrations of contracted academic assessment centers, manufacturer's submissions are scrutinized; models are tinkered with, adjustments are made to cost-per-QALY claims and thresholds re-calibrated. This sets the stage for pricing negotiations and agreement on the terms for formulary listing – all driven by an imaginary construct which, in most cases, is a reformulation of an earlier imaginary construct.

NICE arrives at a determination on the acceptability of a product, couched in terms of threshold cost-per-QALY performance. As these claims for threshold performance are impossible to validate, there is little chance that the NICE decision can be effectively challenged (other than though public opinion and the redoubtable 'Daily Mail'). Paradoxically, while there is presumably evidence to justify building the imaginary reference case, there is no appeal to evidence to validate claims made. Indeed, the reference case itself ensures, whether intentional or not, that the evidence is most unlikely ever to eventuate to question NICE decisions. This is seen in what Popper refers to as the problem of demarcation given the possibility, which distinguishes empirical science from pseudoscience, of immunizing any theory against criticism [51].

What this process overlooks is the fact, as noted by Popper, that never in science are inferences drawn from mere observational experience to the prediction of future events [51]. There is no sense, in these single technology appraisals of any interest in (or any commitment to) a program to discover new facts.

Equivalent Belief Systems

There is, however, a school of thought that could, its adherents would argue, defend the role of imaginary worlds in decision making: relativism. A relativist subscribes to what is known as the equivalence postulate. This postulate holds that it is illegitimate to maintain that one belief system is superior to another. We cannot argue for 'superiority' because we have 'good evidence' for it; in other words, because we have validated a belief or claim by an appeal to the evidence. Application of the standards of normal science is not, for a

relativist, a means of coming to grips with reality. From a sociological perspective, for the relativist (or postmodernist) to accept the standards of normal science in accepting or rejecting claims is to say that one belief system is superior to another; a position which is unacceptable if we accept the equivalence of all belief system [52]. Discovery is put to one side in favor of rhetoric, persuasion and authority [53].

For a relativist the focus is on constructing truth [54]. If there is a group that can convince others of their standards, then they have the power to create belief and decide what they would label 'knowledge'. The result is that modeled claims are not subject to the scrutiny that comes from generating and testing claims. Hypothesis testing is put to one side. The simulated or modeled claim is the end product.

If the objective is to 'construct truth' in the choice of model and its assumptions, then many of the models reviewed here rest on somewhat shaky empirical foundations. Study limitations detailed by authors include: limited clinical data, lack of clinical follow-up data, inappropriate clinical comparator data, lack of data on-follow up or secondary therapy choices, lack of data to support modeling state transition probabilities, limited direct medical cost data (particularly for second-line therapy), lack of indirect cost data, lack of non-clinical primary endpoint data (e.g., utility scores), reliance on underpowered secondary endpoints from clinical studies and an absence of therapy adherence data. Many of these limitations are self-inflicted, due to the long-term perspective of the model itself and the acceptance that the gold standard endpoint is the QALY.

PharmacoEconomics reports that over the past 30 years it has published over 3,000 papers. If the proportion of non-evaluable modeled claims published in the last 12 months are indicative of the weight given to those publications, then the journal has probably published over one thousand thought experiments to support claims for comparative effectiveness (not to mention publication of systematic reviews which bring in dozens more papers). If this is the case (and given the time span involved) a reasonable question might be to ask if any author has attempted to revisit modeled claims to evaluate whether those claims have been substantiated? After all, if the journal has been prepared to publish, say, a modeled cost-per-QALY claim for competing products over a 10 year timeframe, then it should have been possible in this timeframe to revisit this claim.

Acceptance of a belief system, of standards for establishing the superiority of comparative product claims that lies outside of the standards accepted for normal science, assumes that those making formulary decisions share that belief system. While this may be true in the UK where NICE and academic research groups have embraced the reference case model (and indeed these groups were party to its development), in a country such as the US there is little evidence for formulary committees or other health decision makers agreeing on the decision criteria for formulary listing. Given the literally thousands of formulary committees that have emerged following the passing of the Patient Protection and Affordable Care Act in 2010, it is most unlikely that any more than a minority are aware (or, at least adhere to) standards proposed by ISPOR and the format for formulary submissions recommended by the AMCP. It is even more unlikely that there are many that would subscribe to the reference case model. They would probably consider it rather odd to base decisions for product value on lifetime cost-per-QALY models, let alone base decisions on conformity of a claim that is patently non-evaluable to a notional threshold value.

The risk, therefore, is that a wider audience, the audience that comprises decision makers for pharmaceutical products, fails to share the decision standards for modeled claims. The effort put into modeling, systematic reviews and organizational frameworks may be seen, at best, to be odd, but irrelevant to ongoing formulary decisions which seek evidence-based value claims and feedback from prior claims for product performance.

Rejecting Equivalence

It is not possible to subscribe to competing belief systems. This is an impossible and indefensible position: either subscribe to the recognized standards of normal science or admit to participating in a pseudoscience. In other words, a metaphysical exercise which is intended to persuade rather than establish new facts; a belief system, from the relativist perspective, which has no claim to superiority over other belief systems. The reason for this is quite obvious: if a claim is not amenable to empirical testing then we don't know whether it is right or whether it is wrong. It may, to an unknown and unknowable extent, be misleading. As such, untestable claims should be relegated to the status of imaginary worlds or thought experiments; a relegation that is effectively summarized in the motto of the Royal Society (1660 first meeting; 1662 Royal Charter): *nullius in verba* ('take no man's word for it').

The task of relegating economic evaluations to the category of imaginary worlds can be resolved once we abandon attempts to apply standards for modeled claims that are inconsistent with the standards of normal science. Since the 17th century, standards to be applied to modeled claims have been clear-cut: a model is judged on the basis of the hypotheses or claims it creates; its ability to generate new facts. The first hurdle is to agree that the claims are testable, either through observation (e.g., an appeal to existing evidence) or through

experiment (e.g. an RCT). The second hurdle is to test the claim. If the claim is not falsified then it receives provisional acceptance. All claims are provisional; subject to further evaluations which may overturn them. This process of 'conjecture and refutation', as described by Popper, is at the core of the scientific method [55]. It supports the notion of progress in science. It is a process that explores our understanding of the real world and sets the stage for generating, testing and the discovery (and exploration) of new facts.

Conclusions

Rejecting or recasting modeled claims that are 'not fit for purpose' is a necessary step, not only for the formulary committee but for other health care decision makers to recognize the importance of a firm and defensible evidence base for decision making. We can still, of course, subscribe to a hierarchy of evidence and work with decision makers to identify and, hopefully, close evidence gaps. In recognizing that modeled claims may not be fit for purpose does, however, raise the issue of whether or not the publication of modeled claims which are essentially imaginary worlds or thought experiments should be encouraged? Could we argue that they are suggestive of more tractable hypotheses, of projections, and should be published even if there is no assessable hypothesis presented? In the last resort it is presumably up to journal editors and staff to agree on whether or not they wish to subscribe to standards which are relativistic or culturally determined. Should they support postmodernist standards which are clearly at variance with those of normal science in publishing claims for comparative product performance? Or should they come out and declare support for clinical and cost-effectiveness models that drive a research agenda that meets the standards of normal science?

The position taken here is unambiguous: if claims based on imaginary worlds or thought experiments are published, the readership should be advised of this by the authors of the paper. This is not to deny the right to publication but merely to ensure that the evidence presented is 'fit for purpose'. If the editorial policy of a journal such as *PharmacoEconomics* is to subscribe to and support the beliefs accepted in a discipline such as pharmacoeconomics, then this should be made explicit, pointing out that these are not the standards of normal science.

References

1. Langley P. The status of modeled claims. *J Med Econ.* 2015;18(12):991-992
2. Langley P. Validation of modeled pharmacoeconomic claims in formulary submissions. *J Med Econ*. 2015;18(12):993-999
3. Schommer J, Carlson A, Rhee G. Validating pharmaceutical product claims: questions a formulary committee should ask. *J Med Econ.* 2015;18(12):1000-1006
4. Belsey J. Predictive validation of modeled health technology assessment claims: lessons from NICE. *J Med Econ*. 2015;18(12):1007-1012
5. Moher D, Shamseer L, Clarke E et al. Preferred reporting items for systematic reviews and met-analyses protocol (PRISMA-P) 2015 statement. *Systematic Reviews*. 2015;4:1 DOI: 10.1186/2046-4053-4-1
6. Langley PC. *Recent developments in the health technology assessment process* in Fulda TR, Wertheimer AI (eds). *Handbook of Pharmaceutical Public Policy*. New York: Pharmaceutical Products Press. 2007.
7. Sweet B, Tadlock CG, Waugh W et al. The WellPoint Outcomes Based Formulary: Enhancing the Health Technology Assessment Process*. J Med Econ*. 2005;8:13-25
8. Ellis G, Silk J. Defend the integrity of physics. *Nature*. 2014 (516): 321323; see also A Frank and M Gleiser. A crisis at the edge of physics. *New York Times*, 7 June, 2015.
9. Marshall DA, Burgos-Liz L, Ijzerman MJ et al. Applying dynamic simulation modeling methods in health care delivery research – the SIMULATE checklist: Report of the ISPOR Simulation Modeling Emerging Good Practices Task Force. Value Health. 2015;18:5-16
10. Marshall DA, Burgos-Liz L, Ijzerman MJ et al. Selecting a dynamic simulation modeling method for health care delivery research – Part 2: Report of the ISPOR Dynamic Simulation Modeling Emerging Good Practices Task Force. Value Health. 2015;18:47-60
11. Caro JJ, Briggs A, Siebert U et al. Modeling good research practices – Overview: A report of the ISPOR-SMDM Modeling good practices task force – 1. *Value Health*. 2012 (15:796-803
12. Eddy DM, Hollingworth W, Caro JJ et al. Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices Task Force – 7. *Value Health*. 2012;15:843-850
13. Husereau D, Drummond M, Petrou S et al. Consolidated health economics evaluation reporting standards (CHEERS) statement. *J Med Econ*. 2013;16(6):713-19
14. Chiou C-F, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care. 2003;41:32-44.
15. National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal*. London: NICE, April 2013

16. Sawyer LM, Wonderling D, Jackson K et al. Biological therapies for the treatment of severe psoriasis in patients with previous exposure to biological therapy: A cost-effectiveness analysis. *PharmacoEconomics*. 2015;33(2):163-177
17. Blank PR, Filipits M, Dubsky P et al. Cost-effectiveness of prognostic gene expression signature- based stratification of early breast cancer patients. *PharmacoEconomics*. 2015;33:179-90
18. Mensch A, Stock S, Stollenwerk B et al. Cost effectiveness of rivaroxaban for stroke prevention in German patients with atrial fibrillation. *Pharmacoeconomics*, 2015;33(3):271-83
19. Hamid R, Loveman C, Millen J et al. Cost-effectiveness analysis of onabvotulinumtoxin A (BOTOX) for the management of urinary incontinence in adults with neurogenic detrusor overactivity: A UK perspective. *Pharmacoeconomics*. 2015;33:381-93
20. Schawo S, van der Kolk A, Annemans L et al. Probabilistic Markov model estimating cost-effectiveness of methylphenidate osmotic-release oral system versus immediate-release methylphenidate in children and adolescents: which information is needed? *PharmacoEconomics*. 2015;33:489-509
21. Delea TE, Amdahl J, Wang A et al. Cost-effectiveness of dabrafenib as a first-line treatment fin patients with BRAF V600 mutation-positive unresectable or metastatic melanoma in Canada. *PharmacoEconomics*. 2015;33(4):367-80
22. Janzic A, Kos M. Cost effectiveness of novel oral anticoagulants for stroke prevention in atrial fibrillation depending on the quality of warfarin anticoagulation control. *PharmacoEconomics*. 2015;33(4):395-408
23. Bamrungsawad N, Chaiyakunapruk N, Upakdee N et al. Cost-utility analysis of intravenous immunoglobulin for the treatment of steroid-refractory dermatomyositis in Thailand. *PharmacoEconomics*. 2015;33(5):521-31
24. Finkelstein EA, Kruger E, Karnawat S. Cost-effectiveness of Qsymia for weight loss. *PharmacoEconomics*. 2015;33:699-706
25. Vestergaard AS, Ehlers LH. A health economic evaluation of stroke prevention in atrial fibrillation: Guideline adherence versus the observed treatment strategy prior to 2012 in Denmark. *PharmacoEconomics.* 2015;33(9):967-79
26. Takabayashi N, Murata K, Tanaka S et al. Cost-effectiveness of proton pump inhibitor co-therapy in patients taking aspirin for secondaty prevention of ischemic stroke. *PharmacoEconomics*. 2015;33:1091-1100
27. Schremser K, Rogowski WH, Adler-Reichel A et al. Cost-effectiveness of an individualized first-line treatment strategy offering erlotinib based on EGFR mutation testing in advanced lung adenocarcinoma patients in Germany. *PharmacoEconomics*. 2015;33:1215-28
28. Church JL, Haas MR, Goodall S. Cost effectiveness of falls and injury prevention strategies for older adults living in residential aged care facilities. *PharmacoEconomics*. 2015;33:1301-10
29. Gutzwiller FS, Pfeil AM, Ademi Z et al. Cost Effectiveness of Sucroferric Oxyhydroxide Compared with Sevelamer Carbonate in the Treatment of Hyperphosphataemia in Patients Receiving Dialysis, from the Perspective of the National Health Service in Scotland. *PharmacoEconomics*. 2015;33(12):1311-24
30. Hiligsmann M, Evers SM, Sedrine WB et al. A systematic review of cost-effectiveness analyses of drugs for postmenopausal osteoporosis. *PharmacoEconomics*. 2015;33(3):205-24
31. Zhang W, Islam N, Ma C. Systematic review of cost-effectiveness analyses of treatment for psoriasis. *PharmacoEconomics*, 2015;33(4):327-40
32. Woolacott N, Hawkins N, Mason A et al. Etanercept and efalizumab for the treatment of psoriasis: a systematic review. Health Technol Assess. 2006;10:1-233
33. Srivastava K, Thakur D, Sharma S et al. Systematic review of humanistic and economic burden of symptomatic chronic obstructive pulmonary disease. *Pharmacoeconomics*, 2015;33(5);467-488
34. Geng JG, Yu H, Mao Y et al. Cost-effectiveness of dideptidyl peptidase-4 inhibitors for type-2 diabetes. *PharmacoEconomics*. 2015;33(6):581-97
35. Diel R, Lampoenius N, Nienhaus A. Cost effectiveness of preventive treatment for tuberculosis in special high-risk populations. *Pharmacoeconomics*. 2015;33(8):783-809
36. Shields GE, Beard SM. A systematic revciew of the economic and humanistic burden of gout. *PharmacoEconomics*. 2015;33:1029-47
37. Zakiyah N, Postma MJ, Baker PN et al. Pre-eclampsia diagnosis and treatment options: A review of published economic assessments. *PharmacoEconomics*. 2015;33:1069:82
38. Poonawalla IB, Parikh RC, Du XL et al. Cost effectiveness of chemotherapeutic agents in targeted biologics in ovarian cancer: A systematic review. *PharmacoEconomics*. 2015;33:1155-85
39. Fleeman N, Bagust A, Beale S et al. Pertuzuma in combination wIh trastuzumab and docetaxel for the treatment of HER2-positive metastatic or locally recurrent unresectable breast cancer. *PharmacoEconomics*. 2015;33:13-23
40. Greenhalgh J, Bagust A, Boiland A et al. Eribulin for the treatment of advanced or metastatic breast cancer: A NICE single technology appraisal. *PharmacoEconomics*. 2015;33(2):137-148
41. Holmes M, Davis S, Simpson E. Alteplase for the treatment of acute ischaemic stroke: A NICE single technology appraisal; an evidence review group perspective. *PharmacoEconomics*. 2015;33:225-33

42. Haas M, Lourenco R. Pharmacological management of chronic lower back pain: A review of cost-effectiveness. *Pharmacoeconomics*. 2015;33(6):561-69
43. Wade R, Duarte A, Simmonds M et al. The clinical and cost-effectiveness of aflibercept in combination with irinotecan and fluorouracil-based therapy (FOLFIRI) for the treatment of metastatic colorectal cancer which has progressed following prior oxaliplatin-based chemotherapy: a critique of evidence. *Pharmacoeconomics* . 2015;33(5):457-466
44. Stevenson M, Pandor A, Stevens JW et al. Nalmefene foir reducing alcohol consumptions in people with alcohol dependence: An evidence review group perspective of a NICE single technology appraisal. *PharmacoEconomics*. 2015;33(8):833-47
45. Fleeman N, Bagust A, Beale S et al. Dabrafenib for treating unresectable, advanced or metastatic BRAF V600 mutation positive melanoma: An evidence review group perspective. *PharmacoEconomics*. 2015;33(9):893-904
46. Simpson EL, Davis S, Thokala P et al. Sipuleucel-T for the treatment of metastatic hormone-relapsed prostate cancer: A NICE single technology appraisal; an evidence review group perspective. *PharmacoEconomics*. 2015;33:1187-94
47. Giannopoulou C, Sideris E, Wade R et al. Ipilimumab for previously untreated malignant melanoma: A critique of the evidence. *PharmacoEconomics*. 2015;33:1269-79
48. Australian Government. Department of Health. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (version 4.4).* Canberra: July 2013
49. Canadian Agency for Drugs and Technologies in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada (3rd Ed). CADTH: Ottawa, 2006
50. Academy of Managed Care Pharmacy. AMCP Format for Formulary Submissions, Version 4.0 April 2016
51. Popper KR. Objective Knowledge (Rev ed.) (Oxford: Clarendon Press, 1979)
52. Shapin S, Schaffer S. Leviathan and the Air-pump:Hobbes, Boyle and the Experimental Life. Princeton: Princeton University Press, 1985.
53. Wootton D. The Invention of Science. New York: Harper Collins, 2015
54. Foucault, M. The Order of Things, London: Vintage Books, 1970 (1966)
55. Popper KR., The logic of scientific discovery .New York: Harper, 1959

TABLE 1: IMAGINARY WORLDS: ECONOMIC EVALUATION STUDIES PHARMACOECONOMICS JANUARY 2015 TO DECEMBER 2015

Paper (author)	Target Population and Intervention	Sponsor (if any)	Modeling Technique and Claims Status	Claims Assessment and Credibility
Original Research Article				
Sawyer et al [16]	Cost effectiveness of sequential biologic therapies in patients with psoriasis exposed to previous biologic therapy	National Guideline Centre (UK) with funding from NICE	A two part model following the NICE reference case recommendations with a 10-year time horizon. Model divided into an initial short 'trial' period built as a simple decision tree and a longer term Markov transition model with annual cycles and half cycle correctives where patients either continue care or drop out and move to best supportive care. Outcomes: Cost per QALY. Results: Further biologic therapy for patients with psoriasis who have previously been treated with a biologic may be cost effective although considerable uncertainty in results.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a 10-year time horizon precludes claims assessment
Blank et al [17]	Cost effectiveness of prognostic gene expression signature based stratification of early breast cancer patients (EPclin)	Part sponsor Sividon Diagnostics GmbH	A lifetime Markov state transition model with 3 health states (disease free, metastasis and death) and a 1-year cycle. Comparing standard guidelines with guidelines plus molecular tests. Outcomes: cost per patient treated and QALYs. Results: EPclin strategy was dominant with 13.173 QALYs and lower costs. [Note: the study results support the sponsors product].	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a lifetime horizon precludes claims assessment
Mensch [18]	Cost-effectiveness of rivaroxaban for stroke prevention in German patients with atrial fibrillation		A Markov state-transition with 6 health states model comparing fixed dose rivaroxaban with variable dose warfarin. Time line 35 yearas or death for patients aged 65 years with monthly cycles. Outcomes: cost per QALY. Results: QALY adjusted life expectancy for rivaroxaban arm 11.06 years vs. 10.35 years for warfarin. Corresponding total costs €20,238 and €9,464.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a 35-year time horizon precludes claims assessment
Hamid et al [19]	Cost-effectiveness of BOTOX for management of	Allergan Ltd (UK)	Markov state-transition cost per QALY model with 5-year time frame and 6 health states. Treatment arms: supportive care +	No testable hypotheses or consideration given to

	urinary incontinence with neurogenic detrusor overactivity		BOTOX versus best supportive care alone (including anticholinergic drugs ACHDs). Cycle length 6-weeks after week 12. Outcomes: cost per QALY (including reduced frequency of UI). Results: BSC + BOTOX resulted in an increase in 0.4388 discounted QALYs gained with an increase in costs of £1,689 over 5 years. [Note: the study results support the sponsors product].	how the claims might be evaluated in a shorter term model as a 5-year time horizon precludes claims assessment
Schawo et al [20]	Cost-effectiveness of methylphenidate osmotic-release oral systems (OROS) versus immediate-release methylphenidate (IR) in children and adolescents with ADHD	Jansse-Cilag BV	Markov state-transition model with four states with a duration of 12-years. Patients entered the model at 6 years of age following ADHD treatment guidelines. Cycle length of 1-day. Four states were: optimal response, suboptimal response, treatment stopped and remission (no medication). Outcomes: cost per QALY (parent/caregiver evaluation). Results: dominance of OROS compared to IR. Note: the study results support the sponsors product.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a 12-year time horizon precludes claims assessment
Delea et al [21]	Cost-effectiveness of dabrafenib as first-line treatment in BRAF V600 metastatic melanoma in Canada	GlaxoSmithKline	A 5-year partitioned survival analysis model with 3 health states and a cycle length of 1-week . Base treatment naïve patients comparing dabrafenib, dacarbazine and vemurafenib. Time horizon 5-years. Outcomes: incremental cost per QALY adjusted life year. Results: dabrafenib unlikely to be cost-effective compared to dacarbazine; no reliable conclusions regarding dabrafenib versus vemurafenib	No testable hypotheses or consideration given to how the claims might be evaluated although the time line for the model is only 5 years and data accessible for 2-3 year observational study
Janzic et al [22]	Cost-effectiveness of novel oral anticoagulants (NOACs) for stroke prevention in atrial fibrillation		Lifetime state transition Markov model comparing dabigatran, rivaroxaban, apixaban, high-dose edoxaban with standard warfarin treatment. Starting cohort 70-years of age with increased risk for stroke. Outcomes: cost-per-QALY saved. Results: NOACs more likely to be cost-effective in settings with poor warfarin management	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a lifetime time horizon precludes claims assessment
Bamrungsawad et al [23]	Cost-utility of Intravenous Immunoglobulin (IVIG)		Markov four state-transition model with a 12-week cycle for the lifetime of patients initiating therapy at age of 40 years.	No testable hypotheses or consideration given to

	in steroid-refractory dermatomyositis in Thailand		Outcomes: cost-per-QALY saved. Results: IVIG dominant	how the claims might be evaluated in a shorter term model as a lifetime time horizon precludes claims assessment
Finkelstein et al [24]	Cost-effectiveness of Qsymia for weight loss	VIVUS Inc	A 56-week trial based estimate of cost and QALY outcomes (imputed from SF-12v2). Outcomes: Cost per QALY over 1 and 2 years with residual benefits in years 3 and 4. Linear regression modeling of change in QoL as dependent variable and treatment arm. Analogous regressions for PCS and MCS scores. Imputed direct costs. Results: Qsymia cost-effective at $50.000 threshold. May be cost-effective vs surgery if benefits associated with therapy extend beyond medication cessation.	Although no testable hypotheses or consideration given to how the claims might be evaluated in an observational study, the short time-horizon makes testing claims achievable from existing data
Vestergaard et al [25]	Cost-effectiveness of competing strategies for stroke prevention in atrial fibrillation		Markov state-transition model to simulate costs and outcomes of two treatment strategies of guideline adherence versus observed treatment strategy over 10-year time horizon with a 3-month cycle length and a starting age of 70 years. Outcomes: cost-per QALY saved over 10 years. Results: Guideline adherence was cost-effective.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a 10-year time horizon precludes claims assessment
Takabayashi et al [26]	Cost effectiveness of proton pump co-therapy in patients taking aspirin for stroke secondary prevention	None but authorship potential conflict of interest	Markov state-transition model-based simulation of one-year cycle with half-cycle correction and a time horizon of 30 years. Outcomes: life years gained with willingness to pay threshold of US$48,077 and lifetime risk calculated from probability of one health state divided by total of probabilities of all health states. Results: ASA plus PPI co-therapy cost-effective vs ASA in Japan.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a 30-year time horizon precludes claims assessment
Schremser et al [27]	Cost-effectiveness of epidermal growth factor recipient receptor (EGFR)-tyrosine kinase inhibitors as first line	German Research Center for Environmental Health	Markov state-transition model with 3 mutually exclusive health states with a cycle length of 3 weeks and a time horizon of 10 years (assumed patient's lifetime). Outcomes: cost-per-QALY. Results: Individualized therapy with EGFR likelihood	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as

	therapy with erlotinib in advanced lung adenocarcinoma		of 50% cost-effective over standard therapy.	a 10-year time horizon precludes claims assessment
Church et al [28]	Cost effectiveness of falls and injury prevention strategies		A three-state Markov lifetime transition model with a cycle length of 1-year based on starting mean age of residents in care facilities (84 years). Outcomes: falls avoided and QALYs gained. Results: vitamin D supplementation, medication review multifactorial interventions yield QALYs gained and fall averted compared to no-intervention. In cost-effectiveness terms Vitmain D and medication are dominant.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model as a lifetime horizon precludes claims assessment
Gutzwiller et al [29]	Cost-effectiveness of sucroferric exyhydroxide(PA21) versus sevelamer carbonate (SC) in patients receiving dialysis for hyperphosphataemia in Scotland	Vifor Pharma Ltd	Lifetime Markov cohort model to assess cost-effectiveness of therapy with PA21 versus SC with six health states and cycle length of 1-month and half-cycle correction. Mean age at model entry 56 years with adopted model lifetime represented by a maximum of 44 years (528 cycles), corresponding to age 100 years. Outcomes: undiscounted survival, QALYs and cost per QALY gained[Note: Us data indicated that for patients with dialysis at age 40-44 years expected survival 8-years; for those aged 60-64 years survival approximately 4.5 years]. Results: average base case survival 7.61 years. Incremental QALY gain 0.009 QALYS in favor of SC strategy and total cost difference (in favor of PA21) was £1,609.	No testable hypotheses or consideration given to how the claims might be evaluated in a shorter term model. As the effective observational period given average base case survival is only 7.61 years an observational study in warranted given the clinical benefits of PA21 (possibly combined with SC).
Systematic Reviews				
Hiligsman et al [30]	Cost-effectiveness of drugs for postmenopausal osteoporosis		The review identified 39 articles that met inclusion criteria from 42 articles were that assessed using full text. Quality assessment followed the CHEERS checklist. All but one of the studies used QALYs as the outcome with model based cost-effectiveness analyses. A Markov cohort model was used in 28 studies with 8 studies using a microsimulation model and one using a discreet-event simulation model. Seven studies used a fixed time horizon (3, 5 or 10	The review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee. Application of the CHEERS instrument

			years) with 32 studies using a lifetime perspective. Active osteoporotic drugs were generally cost-effective when compared to no treatment at commonly accepted thresholds of around €45,000 per QALY gained. It was not possible to make recommendations, given heterogeneity of studies, on relative cost-effectiveness of drugs.	effectively precluded these questions being addressed.
Zhang et al [31]	Cost-effectiveness of existing treatment options for psoriasis		The review identified 53 articles that met inclusion criteria from 500 articles than merited a full text review. Quality assessment was evaluated by application of the QHES instrument, with the focus on the drivers of cost-effectiveness instead of cost-effectiveness outcomes. Overall, 11 studies used a decision tree framework 10 a Markov model and 7 the York (Woolacott) model [i]. The authors concluded that most cost-effectiveness analyses were of low quality – determined by short time horizons, not using quality adjusted life years as the effectiveness measure , failure to include all relevant resource units or failing to perform a sensitivity analysis.	The review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee. Application of the QHES instrument effectively precluded these questions being addressed.
Srivastava et al [33]	Humanistic and economic burden of Chronic Obstructive Pulmonary Disease (COPD).	GlaxoSmithKline	A total of 32 studies reporting humanistic burden and 74 economic burden studies were identified with 6 studies reporting both. Outcomes in humanistic studies were PROs (including HRQoL); the economic burden was restricted to cost and resource utilization.	Although focused on evidence describing and quantifying the burden of symptomatic COPD the review did not address the issues of testable claims (e.g., in evaluating the burden of disease), the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee in the management of gout.

Geng et al [34]	Cost-effectiveness of dideptidyl peptidase-4 inhibitors for type 2 diabetes		A total of 11 studies met the inclusion criteria following full text review of 36 articles. Quality assessment followed the CHEERS checklist. Seven studies used the Cardiff Diabetes Model, 3 the UK Prospective Diabetes Study (UKPDS) and one the Januvia Diabetes Model. Four studies were of good quality, six of moderate and one of low quality. Four studies were based on a lifetime model and five based on 40 years. One of the latter studies also included 5 years and the remaining study 3 years. Seven of the studies were funded by manufacturers. All studies reported outcomes as QALYs per person.	The review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee.
Diel et al [35]	Cost-effectiveness of preventive treatment for tuberculosis in special high risk populations		A total of 24 cost-effectiveness studies were identified covering six high-risk groups. Of these, 20 studies used a Markov simulation model and one a decision analytic model. QALYs were the outcome measure in 15 studies, with TB case prevented in 7 studies. Time frames ranged from one years to a lifetime. Apart from 6 studies which did not report a timeframe, 5 reported a lifetime framework, and 8 modeled from 20 to 40 years.	Apart from a brief mention of the unlike possibility of a long term prospective study to assess the costs and effects of preventive treatment, The review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee.
Shields et al [36]	Economic and humanistic burden of gout		A total of 39 studies met inclusion criteria, 17 and 26 respectively were relevant to economic and humanistic burden of gout. No mention of application of a quality checklist. The economic burden was restricted to cost studies; the humanistic burden on health related quality of life. Generic HRQoL measures were most frequently reported applied: HAQ-DI (n=6), HAQ-II (n=3) and SF-36 (n = 12). Gout specific measures; GAQ-GI (n= 2 and GIS (n=3).	Although focused on evidence describing and quantifying the burden of gout, the review did not address the issues of testable claims (e.g., in evaluating the burden of disease), the possibility of evaluating any claims or how the claims might be evaluated by

				a formulary committee in the management of gout.
Zakiyah et al [37]	Cost-effectiveness of screening diagnosis and treatment options in pre-eclampsia		The review identified 6 studies that met inclusion criteria from 8 full text reviews. Five of these were economic revaluations and one a budget impact analysis. Quality assessment was evaluated by application of the CHEERS checklist. Four studies used a decision model and two a trial based cost effectiveness analysis (with 2 studies using a 30 year time horizon). No unequivocal conclusions could be drawn as to cost-effective care in pre-eclampsia or the cost-effectiveness of biomarkers for pre-eclampsia	The review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee. Even so, the budget impact analysis and trial based evaluations could have been evaluated in these terms.
Poonawalla et al [38]	Cost-effectiveness of chemotherapeutic agents and targeted biologics in ovarian cancer		A total of 73 full text reviews yielded 28 publications for inclusion. Quality assessment used the QHES checklist. Covering a period of 18 years (1996-2014) made comparisons between studies difficult and any consensus. Cost per life year gained (LYG) was the outcome measure in 15 of the studies; cost per QALY gained the outcome in 13 of the studies.	The range of studies included make any general conclusion difficult. Even so, the review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee.
Review Article				
Fleeman et al [39]	Pertuzamab in combination with trastuzumab and docetaxel for HER-2 metaststic or locally recurrent unresectable breast cancer	NICE Single Technology Appraisal	Roche was asked to submit evidence for the clinical and effectiveness of pertuzumab + trastuzumab + docetaxel vs. placebo + trastuzumab + docetaxel for the treatmentof human epidermal growth factor in HER-2 metastatic or locally recurrent or resectable breast cancer based on one ongoing RCT (CLEOPATRA). While the ERG judged the trial to be of good methodological quality the overall survival data were problematic as the trial was ongoing. Also, the protocol for the trial did not reflect current clinical practice with	The NICE appraisal made no reference to the possibility of testable claims or the possibility of evaluating claims for the product in an NHS environment.

			trastuzumab. The cost-effectiveness case rested on a lifetime partitioned survival model with 3 health states. Given the lack of maturity in the overall survival data application of the model was limited. A provisional extrapolation by the WERG for overall survival yielded a lower survival rate and difference between the two arms. The estimated ICERs were larger than for the manufacturer's base case. It was deemed impossible to set a price for pertuzamab. While the Appraisal Committee rejected the product a further decision was pending at the time of publication	
Greenhalgh et al [40]	Eribulin for advanced or metastatic breast cancer	NICE Single Technology Appraisal		The NICE appraisal made no reference to the possibility of testable claims or the possibility of evaluating claims for the product in an NHS environment
Holmes et al [41]	alteplase for acute ischaemic stroke	NICE Single Technology Appraisal	Boehringer Ingelheim GmbH was asked to submit evidence for the clinical and cost-effectiveness of alteplase for the prevention of strokes within a 0 – 4.5 hour window from the onset of stroke symptoms with the comparator of standard medical and supportive management that did not include alteplase. A lifetime model was utilized generating cost-per-QALY estimates as the primary endpoint. The ERG generally accepted the clinical claims and the modeled cost-effectiveness case. Treatment effect was captured by modeling the distribution of patients between the health states dependent, independent, dead at 6 months following treatment. Probabilities of transition between the health states beyond 6 months were modelled from a stroke registry, with patients remaining in that health state until they experienced a recurrent stroke or died. Age specific	Although there were opportunities given the structure of the modelling to evaluate claims at three or six monthly intervals, the NICE appraisal made no reference to the possibility of testable claims or the possibility of evaluating claims for the product in an NHS environment

			mortality risk for those without or experiencing a recurrent stroke were applied. The product was recommended for marketing authorization.	
Haas et al [42]	Pharmacologic management of chronic lower back pain		A total of 7 studies were identified. Two of these were modelled studies. Both scored highly on the QHES; other studies were considered poor due to lack of good quality clinical evidence. Most common outcome measures were pain and disability. In the two studies that used a reference case semi-Markov model the QALYs were measured indirectly using a transfer to utilities regression equation. Neither model specified the ICER timeframe. The authors concluded that in the absence of RCT data, economic models should be used to estimate future costs and outcomes using robust methods.	The review did not address the issues of testable claims, the possibility of evaluating any claims or how the claims might be evaluated by a formulary committee.
Wade et al [43]	Aflibercept in combination with irinotecan and FOLFIRI for metastatic colorectal cancer	NICE Single Technology Appraisal	Sanofi was asked to submit clinical and cost-effectiveness evidence for aflibercept in combination with irinotecan and fluorouracil for treatment of metastatic colorectal cancer. The clinical data were from one RCT that compared aflibercept + FOLFIRI with placebo + FOLFIRI. The RCT found a small but significant difference in median overall survival and progression free survival. The ERG considered the manufacturer's extrapolation of survival curves to 15 years to be excessive, instead truncating the analysis to 5 years to reflect a more realistic survivorship profile. Cost-effectiveness was based on a 3-state Markov model with a 15 year time horizon; outcomes were cost-per-QALY. A major concern was with the impact of adverse events on continuation and HRQoL r The ERG defined an alternative base case which increased the ICER to £54,368 per QALY. The product was not recommended; a decision upheld on appeal.	The NICE appraisal made no reference to the possibility of testable claims or the possibility of evaluating claims for the product in an NHS environment
Stevenson et al	Nalmefene for	NICE Single	Lundbeck was asked to submit evidence for	The NICE appraisal

[44]	reducing alcohol consumption	Technology Appraisal	the clinical and cost-effectiveness of nalmefene for reducing alcohol consumption in people with alcohol dependence. Clinical evidence was from 3 phase III trials that compared nalmefene plus psychosocial support with placebo plus psychosocial support. Cost-effectiveness claims were based on Markov cohort model with a timeframe of 5 years. The model comprised a one year short term phase (cycle length1 month) and a long term phase (cycle length 1 year). Utility in the first year was from trial-based pooled EQ-5D results; in the long term utility was assumed to be a function of drinking status. The ERG reworked the model, in part to draw comparisons with the NICE CG15 recommendations for psychosocial interventions. Although based on limited data (with one comparison dropped) the reworked ICERS were little different from those first submitted. The methodological issue that had the largest impact was the fact that the pivotal RCTs did not use the psychosocial support listed in the scope. Nalmefene, was recommended within its marketing authorization, prescribed in conjunction with continuous psychosocial support..	made no reference to the possibility of testable claims or the possibility of evaluating claims for the product in an NHS environment.
Fleeman et al [45]	Dabrafenib to treating unresectable, advanced or metastatic BRAF V600 for mutation-positive melanoma	NICE Single Technology Appraisal	GlazoSmithkline was asked to submit evidence for the clinicakl and cost-effectiveness of dabrafenib. Although the ongoing BREAK-3 trial compared dabrafenib with dacarbazine, the Evidence Review Group (ERG) considered vemurafenib to be the appropriate comparator. Evidence for vemurafenib was taken from the BRIM-3 trial where the comparator was dacarbazine. The company presented an indirect treatment comparison model that demonstrated no clinical difference between the two products. The ERG rejected the comparison, mainly for the validity of the assumptions. This undermined modeled	The NICE appraisal made no reference to the possibility of testable claims or the possibility of evaluating claims for the comparator products in an NHS environment for either clinical or cost-effectiveness claims.

			claims. The model used to establish cost-effectiveness had additional issues that undermined its credibility. A 30-year Markov model was presented with 3 states and a cycle length of 1 week. The ERG rejected the company's claims for long-term survival which were significant in making the case for the company's product vs. the comparator. There were limited data on HRQoL. The company submitted evidence did not conform to the NICE methodological guidelines with an inappropriate analysis of the evidence. The two products were considered to be identical in clinical practice	
Simpson et al [46]	Sipulencel-T for metastatic hormone-relapsed prostate cancer	NICE Single Technology Appraisal	Dendreon, the manufacturer of sipuleucel-T was asked to submit evidence for clinical and cost-effectiveness the product where the comparator was abiraterone acetate (AA) or best supportive care (BSC). Based on the balance of evidence presented, the ERG concluded that sipuleucel-T in two of three trials improved overall survival but none showed prolonged time to disease progression. The Advisory Committee concluded that the product improved overall survival compared to APC-PBO. However, in the low PSA-subgroup (which was considered the relevant group for marketing authorization) there was no evidence for superiority in overall survival. The company cost-effectiveness model considered both the whole population as well as subgroups. A lifetime horizon was assumed with monthly time cycles with parametric survival curves for overall survival. The primary outcome was QALYs gained. The ERG had a number of concerns with the model and undertook a separate modelling exercise on a set of nine preferred assumptions. The ERGs probabilistic sensitivity analysis suggested either a zero probability or a very low probability of the product meeting a	The NICE guidance recommended against the product. The NICE appraisal made no reference to the possibility of testable claims or the possibility of further evaluating claims for the product in an NHS environment for either clinical or cost-effectiveness claims.

			£50,000 threshold	
Giannopoulou et al [47]	Ipilimumab for previously untreated unresectable malignant melanoma	NICE Single Technology Appraisal	Bristol-Myer-Squibb were asked to submit clinical and cost-effectiveness evidence for ipilimumab in previously untreated advanced melanoma compared to dacarbazine (DTIC) and vemurafenib at a recommended dose of 3mg/kg. The primary source of clinical data was the CA184-024 trial which was based on a dose of 10mg/kg. Results over a 5-year period supported the 10mg/kg dosing in demonstrating a significant increase in overall survival. The cost-effectiveness analysis assumed the equivalence of the 3mg/kg and 10mg/kg regimen. This was questioned as lacking robust evidence support together with evidence to inform the sequential use of treatment. The manufacturer responded with additional evidence to support ipilimumab as first line therapy. The analyses yielded an ICER of £47,900 gained compared to DTIC and £28,600 compared with vemurafenib. The ERG were concerned with the levels of uncertainty associated with both the clinical and cost-effectiveness claims. From the clinical perspective there was no evidence that the dosing regimens were equally clinically effective or that ipilimumab plus DTIC may be more effective than ipilimumab alone. The cost-effectiveness analysis applied a semi-Markov model with ipilimumab as first line therapy, second-line active therapy and third line best supporting care. The ERG was critical of the lack of data to support the sequencing outcomes modeled and suggested a first-line model only to be more appropriate. Even then, there was considerable uncertainty. This uncertainty was not resolved for the ERG in the manufacturer's further response to the Appraisal Committee's initial decision and NICE recommendation for recommending the product only in the context of research	Although the ERG admitted considerable uncertainty with respect both to clinical claims for equivalence in dosing and in combination therapy together with the lack of clinical evidence to support treatment sequencing, Even with the short survival time frames with unresectable melanoma the NICE appraisal made no reference to the possibility of testable claims or the possibility of evaluating claims for the product in an NHS environment for either the clinical or cost-effectiveness claims.

			as part of a clinical study. Marketing approval was finally given after a discount was agreed with the manufacturer under the market access scheme. IT was admitted that there were concerns that the appraisal process was initiated before there was sufficient evidence to inform the cost-effectiveness assessment.	