INNOVATIONS in pharmacy

Volume 7 | Number 2 Article 12

7-5-2016

Sunlit Uplands: The Genius of the NICE Reference Case

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

Sunlit Uplands: The Genius of the NICE Reference Case

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

Abstract
The NICE reference case has received widespread acceptance in health technology assessment. The lifetime cost-per-QALY model and constructed claims for product impact have been widely emulated in country-specific guidelines for formulary submission as well as in publications in the leading health technology journals. Unfortunately, from the perspective of the standards of normal science, adherence to the reference case standard means that the claims made are typically non-evaluable. They have to be taken at face value. They may suggest potential evaluable hypotheses for clinical and cost-effectiveness claims, but there is no requirement in the reference case for claims to be put in an evaluable form and for manufacturers to suggest possible protocols for product impact assessment. This is not an acceptable situation. Absent the standards for falsification and replication, which are at the core of the scientific method, we have no idea whether the claims accepted by NICE are right or even if they are wrong. If we accept the reference case paradigm should we conclude that the sunlit uplands of formulary decisions based on non-evaluable simulated claims for cost-effectiveness has been reached? Have we rejected natural selection in favor of intelligent design?

Keywords: NICE, reference case, cost-effectiveness, cost-utility, modeling, credibility, imaginary worlds, scientific method

Introduction

The National Institute for Health and Care Excellence (NICE) has an enviable reputation. Since its formation in the late 1990s (and after a name change) it has assumed a leadership role in health technology assessment. This is due, not only to the resources allocated to its range of activities, but to its application of a reference case model to the process health technology assessment; a reference standard for health technology assessment that has been widely accepted [1 2 3]. The purpose of this review is to consider whether or not the reference case approach to health technology assessment has merit. Specifically, are the standards and processes of reference case evaluations for pharmaceutical products and devices compatible with what may be described as the scientific method? If they are not, are there possible explanations for the ongoing advocacy of the reference case methodology and its central role in NICE recommendations for formulary acceptance?

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 scientific method is not new; it can be traced back at least 350 years to the work of Bacon, Galileo, Boyle, Descartes and Newton [4]. The essence of the scientific method is captured in the motto of the Royal Society (founded 1660; Royal Charter 1662): Nullius in Verba ("take no man's word for it"). Unless there is the ability to test a hypothesis and if that hypothesis is tested, then we have some confidence in the worth of the hypothesis. Otherwise, the claim should not be taken at face value; irrespective of how reasonable and appealing it might be [5]. Popper has made abundantly clear the worth of the scientific method in his advocacy of the falsification hypothesis: a theory in the empirical sciences can never be proven [6 7]. It must be scrutinized by decisive experiments [8]. Theories that survive are not more 'true', they are more 'fit' for the problem at hand. For Popper it is from the interplay between the tentative theories (conjectures) and error elimination (refutation) that scientific knowledge advances. As Popper notes: 'non-reproducible single occurrences are of no significance to science'. Experimentation is central: a process described in a recent issue of *Science* as one where 'The deepest trust in scientific knowledge comes from the ability to replicate empirical findings directly and independently, whether through reanalyzing original data or by creating new data' [9].

Irrespective of whether or not we are developing hypotheses regarding the impact of minimum wage proposals on employment, the impact of rent control on the quality of a housing stock or the clinical and cost-effectiveness claims of competing pharmaceutical products, the scientific method dictates that they be presented in evaluable terms. In the

absence of experimental evidence, or the potential to demonstrate an experimental result, we have no idea, as Wolfgang Pauli has said, whether the claims made are right or even if they are wrong [10]. The key question for NICE is whether their standards for technology assessment are consistent with the standards of normal science or reflect acceptance of a relativist position in evidence-based decisions

The Reference Case

The NICE reference case requires submission for product review to conform to the following standards:

- Decision problem as defined by NICE scope
- Comparator therapies are those listed in the NICE scope
- Required to include of all direct health effects for patients and caregivers
- Costs are those incurred by the NHS and PPS
- Required cost-utility analysis with fully incremental analysis
- Required time horizon long enough to reflect all important differences in costs and outcomes between technologies being compared
- Claims for health effects to be based on a systematic review
- Health effects to be expressed as quality adjusted life years (QALYs) with the EQ-5D the preferred instrument
- Data for measurement of health related quality of life should be that reported by patients and/or caregivers
- Preference data for valuation of changes in health related quality of life should be from a representative sample of UK population
- QALYs all have equal equity weight
- Resource use and costs should be valued using prices relevant to NHS and PPS
- Costs and benefits should be discounted at same annual rate (3.5%)

The reference case requires the decision model time horizon to be sufficiently long 'to reflect all important differences in costs and outcomes between the technologies being considered'. For NICE, as many technologies have an impact over the lifetime of a patient, a lifetime horizon is usually appropriate. Noting also that 'analyses that limit the time horizon to periods shorter than the expected impact of treatment do not usually provide the best estimates of benefits and costs'.

Application of NICE reference case standards, in the case of chronic disease interventions, results in a model or simulation that attempts to mimic the natural progression of the disease and the impact of competing interventions over the patient's lifetime or similar long-term time horizon. Stages of disease progression are captured by, for example, a Markov process which tracks the hypothetical cohort of patients through the disease stages. Each health state is defined in terms of associated utilities and costs. The results in scenario driven claims expressed in cost-per-QALY terms. By application of a willingness-to-pay threshold cost-per-QALY, products are either judged acceptable, rejected or accepted after agreement on a discounted price.

Responding to the Reference Case

Following the reference case criteria, a manufacturer typically prepares a modeled or simulated clinical and cost-utility case for their product. In a single technology appraisal (STA) the comparator is usually the standard of care. Understandably, the submission is inclined to put the manufacturer's product in a positive light in order to support the suggested product price. The manufacturer would endeavor also to claim consistency with standard proposed by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) And the CHEERS and similar input quality check lists [11] [12] [13].

Consider, as case studies in the application of the reference case, models developed to support the marketing of the four novel anti-coagulants (NOAC); dabigatran, rivaroxaban, apixaban and endoxaban. NICE has undertaken four single technology assessments of these products [14] [15] [16] [17]. The process of assessment is in three stages: (i) the manufacturer submits its model; (ii) the model is evaluated by an academic or similar evidence review group (ERG) that consider the model structure, the input assumptions, the basis for the assumptions and the possibility of sub-group analyses; and (iii) the reviewers comments and possible reformulation of the model are then assessed by the NICE Advisory Committee. A final version of the model or simulation is agreed together with the modeled claims for utility scores, costs (both discounted) and incremental cost-per-QALY scores. These imaginary scores are then applied to a NICE determined willingness to pay cost-per-QALY threshold (usually £20,000). At no stage is the issue of evaluating the claims raised.

The various NOAC models, extrapolating from their respective Phase 3 clinical trials, all adopted the recommended lifetime Markov framework. Discrete health states reflecting the range of possible adverse events (bleeds, stroke) were proposed together with cycle length (3 months or 1-year). Transition probabilities were estimated from the Phase 3 trials and the literature. Assumptions, often trial based, for therapy discontinuations were imposed. Utilities for the health states were constructed from the literature,

together with estimated direct medical costs for each health state. Utilities and costs, apart from an age-utility decrement factor, are invariant over the lifetime of the model; both are discounted. Then, voilà, base-case and supplementary ICERs emerge yielding QALY estimates, lifetime costs and cost-per-QALY estimates to set against the willingness to pay threshold.

In the case of edoxaban, the constructed discounted probabilistic cost-utility outcomes estimated by the ERG from the manufacturer's base case model yielded a lifetime cost per QALY for warfarin of £12,868 and a lifetime QALY gain of 6.56 years. The constructed results for edoxaban were £15,451 and QALY gain of 6.72; an increment of discounted 58 quality days over the lifetime of those initiated to edoxaban rather than warfarin. Corresponding constructed QALY estimates for the other NOACS were: rivaroxaban 6.65, dabigatran 110mg 6.66, dabigatran 150mg 6.75 and apixaban 6.77. Differences between the lowest QALY NOAC and the highest QALY NOAC (rivaroxaban vs. apixaban) is 76 quality lifetime days. Constructed discounted lifetime costs were rivaroxaban £16,313, dabigatran 110mg £15,732, dabigatran 150mg £15,293 and apixaban £15,531. The reviewers concluded that 'edoxaban, dabigatran etexilate 110 mg and rivaroxaban were strictly dominated by dabigatran etexilate 150 mg and apixaban extendedly dominated dabigatran etexilate 150 mg (more effective and less costly) with an ICER of £13,036 per QALY gained compared to warfarin'.

After the manufacturer's submission, the ERG assessment by the Liverpool Reviews and Implementation Group (LRiG), involving eleven reviewers, the 'evidence' was considered by the Appraisal Committee on whether or not the claimed ICER met the willingness-to-pay threshold [19]. The Appraisal Committee noted that the warfarin monitoring costs were the main cost driver in the cost-effectiveness analysis. Unfortunately, there was no agreement on what that cost might be. The Appraisal Committee considered that the ICER relative to warfarin was likely to be closer to the ERG estimate of £26,000 per QALY compared to the company estimate of £2,500 per QALY. However, due to perceived flaws in the company model and the lack of definitive warfarin monitoring costs the committee could not accept the lower estimate. At the same time the upper ICER was considered uncertain. In a Solomon-like decision the Appraisal Committee concluded that, as the clinical evidence was similar to the other NOACs previously reviewed and the price of edoxaban was similar to rivaroxaban, it was likely that the plausible (but unknown) ICER for edoxaban was likely to be in line with the other NOACs. Edoxaban could therefore be recommended as a cost-effective use of NHS resources.

On the face of it nothing could be simpler or more compelling. The precision of these constructed results for these projected lifetime discounted costs and QALYs is impressive; certainly more impressive than a 10-day weather forecast or economists' projections for GDP growth in the next 12 months (which are invariably incorrect). The evidence base for the input assumptions is well documented and the disease stages reflect the clinical process of the disease. If this correspondence with the real world is agreed by the academic review group and the NICE advisory committee to be *sufficient*, then the *necessary* cost-utility outcome is entailed [18].

Obviously these cost and QALY claims are not only constructed but are impossible to put in testable terms for empirical assessment and replication. The current NHS costs identified for the reference case are expected to remain unchanged (apart from discounting) over the modeled time horizon. Estimated lifetime costs can only be challenged by challenging cost assumptions. The QALY estimates are typically based on the literature. In the case of edoxaban while the initial event quality of life estimate was trial based all subsequent utilities for the treatment and event stages in the Markov 2-week cycle model were literature based.

If we accept the reference case methodology as appropriate for formulary decisions it seems reasonable to conclude that there was never any intention that the reference case should generate evaluable claims. After all, QALYs are not collected in the UK by regional health authorities and, as far as can be judged, these health authorities have no interest in collecting QALYs. Decisions were to be based on constructed evidence. Any dispute over the claims made could only be resolved by challenging the model or simulation. Given this, the question is then one of whether we have crossed the boundary between science and pseudoscience; rejecting natural selection in favor of intelligent design? [20]

Truth is Consensus

But perhaps we are being unreasonable. Perhaps NICE in its advocacy of the reference case is not rejecting the scientific method but taking a relativist position. Rather than subscribing to the position that the standards of normal science are the only standards to apply in health care decisions, the relativist believes that all perspectives are equally valid. In their advocacy of the equivalence or symmetry, health care decisions are to be understood sociologically. No one body of evidence is superior to another. Results of a simulation are on an equal basis with those of a randomized clinical trial (RCT). For the relativist the success of a scientific research program, in this case one built on models and simulations, rests not on its ability to generate new knowledge but on its ability to mobilize the support of a

community. Basing decisions on models and simulations underpins the consensus view that evidence is constructed, never discovered. Instead of coming to grips with reality, science is about rhetoric, persuasion and authority. Truth is consensus.

From this perspective, the construction of evidence in the modeled or simulated chronic disease environment is readily understandable. There is no need to discover new facts. Supported by a community of academic advisors and professional organizations, NICE embraces the relativist position. Simulations or models are accepted because in the consensus view, the view of the authorities in the discipline, including academics from leading technology assessment groups, the ability to capture the assumed critical or similar features of the reality of a decision is all that is required. The mandated reference case is to demonstrate who is in the driving seat. Reality is achieved through rhetoric, persuasion and authority.

The simulations are, for NICE, the reality of the decision environment. Unlike the Royal Society we have to take their word for it. The simulations are seen as sufficient to support formulary decisions with the claims for competing products necessarily entailed by the model assumptions and core mechanism. Any thoughts of interplay between tentative theories (conjectures) and error elimination (refutation), the process of the discovery of new facts, are not only put to one side but they are irrelevant. The broad, sunlit upland of non-evaluable cost-effectiveness decisions based on constructed evidence has been reached.

Stimulations Can Fail

But simulations can be challenged; simulations can fail. Practically, there is unlikely to be agreement on correspondence, sufficiency and necessary entailment. Practitioners can agree that a Markov process is appropriate to capture the natural course of a disease, yet disagree on the cycle length, the number of health states and transition probabilities that the model accommodates. The flexibility allowed in constructing simulations means that simulations, by their nature, can always be challenged. Indeed, any number of competing simulations coming to different conclusions, are entirely possible. Rather than capturing the essence of a decision problem, the simulation captures the perception of the essence of the problem held by the authors of the simulation guided by an existential reference case or similar standards. In the last resort, the validation of one simulation can only be achieved by matching to the performance of another.

Consider the novel oral anticoagulant (NOAC) case. A recent study by Yao et al compares adherence patterns for warfarin with those for rivaroxaban, dabigatran and apixaban and their impact on risk of stroke and major bleeding [21]. During a median follow up of 1.1 years, only 47.5% of NOAC patients were adherent, defined as a medication possession ratio (MPR) of ≥ 80% of therapy days. Adherence to warfarin was 40.2%. Apixaban had the highest unadjusted adherence (61.9%) and dabigatran the lowest (38.5%). The rivaroxaban rate was 58.4%. Applying a multivariate logistic regression, adjusted adherence rates were 38.7% for warfarin and 47.5% for all NOACs. Higher rates of adherence were found across all treatments for those at higher risk. For those with a CHAa2DS2-VASc ≥ 4 the warfarin adherence was 53.4% and the average for the NOACs 59.8%.

Persistence with NOACs has been reported in three recent observational studies. Forslund et al utilizing data from the administrative health register of the Stockholm region evaluated crude and adjusted persistence from the index OAC prescription in the period April 2011 to December 2014 [22]. At the end of the first year crude overall persistence was 88.2% and 82.9% at the end of the second year. Persistence with warfarin at the end of the first year was 85.0%, apixaban 85.9%, dabigatran 74.4% and rivaroxaban 77.4%. In the UK, Martinez et al reported on persistence with longitudinal data from the Primary Care Clinical Practice Research Datalink between January 2011 and May 2014 . Persistence with warfarin at the end of the first year was 63.6$ and 79.2% for all NOACs. In Germany, Beyer-Westendorf et al reported persistence from primary care patients at 180 days of 66.0% for rivaroxaban, 60.3% for dabigatran and 58.1% for VKA. At 1 year corresponding persistence estimates were 53.1%, 47.3% and 25.45% respectively. An MPR ≥ was found for 61.4% of rivaroxaban and 49.5% of dabigatran patients . Experience in Australia for the period 2006 to 2009 for a sample of 1,108 patients, mean age 74 years, found that 15% (95% CI: 13-17%) failed to collect the first repeat warfarin prescription. Median persistence time on medication was only 12 months (95% CI: 10-13%) with long-term persistence at 33 months at 26% (95% CI: 23-27%).

Overall, these estimates suggest that by the end of one year after the index prescription persistence with warfarin is in the range 60 to 70% with a corresponding NOAC rate of 70 to 80%. By the end of year 2, persistence is likely to be 15 to 20% lower. Beyond two years is sheer speculation, although it would not be unreasonable, given evidence for persistence in other chronic disease states, to argue that the overwhelming majority of patients have discontinued within 5 years. Given the age at which treatment is usually initiated for atrial fibrillation, deaths to patients need to be factored in to persistence estimates. In the edoxaban pivotal trial, for example, 10-8% of patients died before the end of the trial.

Under reasons for discontinuation death was given in 3.1% of warfarin patients and 2.8% of edoxaban patients.

Adherence patterns add a further dimension. Although not captured in any of the pivotal trials, other than reporting time in therapeutic range for warfarin, adherence as reported by Yao et al., is a potentially important offset to claimed therapy gains. While it is often difficult in observational studies to distinguish adherence from discontinuation, their additive effect on projected model end-points in therapy could be significant.

If the majority of patients initiated to an oral anticoagulant (OAC) have discontinued therapy, for event related reasons, non-event related reasons and death within 4 to 5 years of their initial prescription, then it seems rather odd to focus on creating projections for discounted direct medical costs and utilities over a lifetime horizon. Instead of modeling switching and discontinuation behavior over the lifetime of a treatment cohort, a more practical and useful approach would be to recognize the likelihood of early discontinuation (or switching to successor compounds) and generate comparative predictions for events and costs for a timeframe that can be captured from existing data sources as feedback to formulary committees.

There are, in fact, example in the literature over the past 5 years of cost-effectiveness studies explicitly taking a short-term modeling approach that focus on clinical outcomes and eschew a QALY endpoint [25 26]. These models adopt a two-year time frame with a Monte Carlo simulation. The claims presented are readily evaluable in treatment settings. QALYs are put to one side because of doubts as to their validity with observed and calculated utility values varying significantly, together with concerns over preference consistency [27]. In short, if it is possible to commit resources to developing and reviewing lifetime cost-per-QALY models is should be relatively straightforward to apply those resources to short term models that generate evaluable claims. These, at least, would have the merit of meeting the standards of normal science.

The Genius of the Reference Case

The genius of the reference case is that it puts to one side hypothesis testing, falsification and replication. There is no need to demonstrate empirically the worth of a claim. Replication is out of the question. Once the NICE appraisal is complete, guidance can be drawn up safe in the knowledge that they rest on a sound and incontrovertible evidence base. Clinical guidance can be implemented safe in the knowledge that any challenge or feedback will be impossible.

The virtually unassailable position that adoption of a reference case and its putative popularity with the NHS, other single payer health systems and model builders in technology assessment, means that there would be a considerable push back to recommendations for a defensible assessment process.

Whether these systems embraced a reference case model (with some rejecting willingness-to-pay thresholds such as PHARMAC in New Zealand) in the full recognition of the advantages of mandating or recommending non-testable modeled claims is an open question. A more reasonable interpretation is that they simply followed the crowd: lifetime cost-per-utility models were the flavor of the month, so let's hop on the bandwagon. One day, perhaps, the penny will drop.

Conclusions

There are two reasons for being optimistic; for believing that at least some practitioners in this discipline will see the penny drop in favor of the scientific method. First, the growing concern over the failure to replicate clinical claims, particularly with outcome switching, and second, increasing pressure from competing manufacturers, physicians and patient groups on pharmacy and formulary decision makers to justify formulary decisions and to present, at least in the US evidence-based value propositions to support product adoption [28 29]. The ready access in the US to 'big healthcare data' offers a 'relatively' low cost and timely access to administrative claims and electronic records to address the question: can we substantiate the claims for that product or device? [30 31 32] Together with the 'supplementary' question: why were the claims made not put in testable form with protocols to detail how the claims should be assessed?

References

1. National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal*. London: NICE, April 2013
2. New Zealand. Pharmaceutical Management Agency. Prescription for Pharmacoeconomic Analysis: Methods for cost-utility analysis (v 2.1). PHARMAC, 2012.
3. Ireland. Health Information and Quality Authority. Guidelines for the Economic Evaluation of Health Technologies in Ireland. Dublin: HIQA; 2014
4. Wootton D. *The Invention of Science*. New York: Harper, 2015.
5. 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.

6. Popper KR., *The logic of scientific discovery* .New York: Harper, 1959.
7. Popper KR. *Objective knowledge* (Rev. Ed.) Oxford: Clarendon Press, 1979.
8. Editorial. *Nature*. 2016;529:256.
9. Camerer CF, Dreber A, Forswell E et al. Evaluating replicability of laboratory experiments in economics. *Science*. 10.1126/science.aaf0918 (2016)
10. Prochnow HV. The Successful Toastmaster: A treasure chest of introductions, epigrams, humor and quotations. Harper: New York, 1966
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. National Institute for Health and Care Excellence. Dabigatran etexilate for the prevention of stroke and systemic embolism in atrial fibrillation. (NICE technology appraisal guidance 249). March 2012.
15. National Institute for Health and Care Excellence. Rivaroxaban for the prevention of stroke and systemic embolism in people with atrial fibrillation. (NICE technology appraisal guidance 256). May 2012.
16. National Institute for Health and Care Excellence. Apixaban for preventing stroke and systemic embolism in people with non-valvular atrial fibrillation. (NICE technology appraisal guidance 275). February 2013.
17. National Institute for Health and Care Excellence. Edoxaban for preventing stroke and systemic embolism in people with non-valvular atrial fibrillation. (NICE technology appraisal guidance 355). September 2015
18. Grim P, Rosenberger R, Eason RE et al. How simulations fail. *Synthese*. 2011 DOI 10 1007/s11229-011-9976-7
19. Greenhalgh J, Longworth L, Crossan Cet al. Edoxaban tosylate for preventing stroke and systemic embolism in people with non-valvular atrial fibrillation: A Single Technology Appraisal. LRiG, 2015.
20. Pigliucci M. Nonsense on stilts: How to tell science from bunk. Chicago: University of Chicago Press, 2010
21. Yao X, Abraham NS, Alexander GC et al. Effect of adherence to oral anticoagulants on risk of stroke and major bleeding among patients with atrial fibrillation. *J Am Heart Assoc*. 2016;5:e003074 doi: 10.1161/JAHA.115.003074
22. Forslund T, Wettermark B, Hjemdahl PO. Comparison of treatment persistence with different oral anticoagulants in patienbts with atrial fibrillation. *Eur J Clin Pharmacol*. 2016;72(3):329-38
23. Martinez C, Katholing A, Wallenhorst C et al. Therapy persistence in newly diagnosed non-valvular atrial fibrillation treated with warfarin or NOAC. A cohort study. *Thromb Haemost*. 2015;115(1):31-9
24. Beyer-Westendorf J, Enliken B, Evers T. Real-world persistence and adherence to oral coagulation for stroke risk reduction in patients with atrial fibrillation. Europace. 2016 Jan 31. Pii:euv421
25. Beresnisk A, Ariza-Ariza R, Garcia-Llorente JF et al. Modelling cost-effectiveness of biologic treatments based on disease activity scores for the management of rheumatoid arthritis. *Int J Inflammation*. 2011. Article ID 727634
26. Beresniak A, Caruba T, Sabatier B et al. Cost-effectiveness modelling of percutaneous coronary interventions in stable coronary artery disease. *World J Cardiology*. 2015;7(10):594 – 591
27. Beresniak A, Medina-Lara A, Auray JP et al. Validation of the underlying assumptions of the quality-adjusted life-years outcome: results from the ECHOUTCOME European project. *PharmacoEconomics*. 2015;33:61-69
28. Centre for Evidence Based Medicine. *COMPare Project*. Oxford, 2016.
29. The Economist. *For my next trick* …. London, 25 March 2016
30. Langley P. The status of modeled claims. *J Med Econ.* 2015;18(12):991-992
31. Langley P. Validation of modeled pharmacoeconomic claims in formulary submissions. *J Med Econ*. 2015;18(12):993-999
32. Wasser T, Haynes K, Barron J, Cziraky M. Using 'big data' to validate claims made in the pharmaceutical approval process. *J Med Econ.* 2015;18(12):1013-1019