Modelling the cost-effectiveness of diagnostic tests

Tristan Snowsill ORCiD: 0000-0001-7406-2819 t.m.snowsill@exeter.ac.uk

Health Economics Group, University of Exeter, Exeter UK

Abstract

Diagnostic tests are used to determine whether a disease or condition is present or absent in a patient, who will typically be suspected of having the disease or condition due to symptoms or clinical signs. Economic evaluations of diagnostic tests (e.g., cost-effectiveness analyses) can be used to determine whether a test produces sufficient benefit to justify its cost. Evidence on the benefits conferred by a test is often restricted to its accuracy, which means mathematical models are required to estimate the impact of a test on outcomes which matter to patients and health payers. It is important to realise the case for introducing a new test may not be restricted to its accuracy, but extend to factors such as time to diagnosis and acceptability for patients. These and other considerations may mean the common modelling approach, the decision tree, is inappropriate for underpinning an economic evaluation. There are no consensus guidelines on how economic evaluations of diagnostic tests should be conducted – this article attempts to explore the common challenges encountered in economic evaluations, suggests solutions to those challenges, and identifies some areas where further methodological work may be necessary.

Key Points for Decision Makers

- Economic evaluation of diagnostic tests typically requires economic modelling with significant structural assumptions
- The methodological approach adopted in economic models for diagnostics may vary according to the value proposition for the diagnostic
- Public and patient preferences for characteristics of diagnostics may not be captured by standard QALY calculations, and flexibility may be required to make optimal resource allocation decisions

1. Background

Medical tests are used extensively in healthcare to determine the presence or absence of a disease, the extent of a disease, the response to treatment, the presence of risk factors, the likelihood the disease will respond to a particular treatment (e.g., precision medicine), and other uses besides. Some medical tests require minimal equipment or consumables, such as the capillary refill test, auscultation and the Mini-Mental State Examination, while other medical tests require extremely expensive equipment, such as magnetic resonance imaging. Some tests require invasive procedures, such as biopsies and colonoscopies. Almost universally, the patient derives no benefit from the test, but the information obtained by the test is expected to lead to some change in how the patient is managed (e.g., initiating treatment). If a test does not change clinical management, it will have no clinical utility or economic value.

Diagnostic tests support clinicians to decide whether a particular disease or condition is present (or likely to be present) in a presenting patient, who will typically be symptomatic. This is distinguished from population and targeted screening, where individuals are invited to have a test for a disease without any indication they have the disease; they are simply *at risk* for the disease. A single test may be used for diagnosis, screening, surveillance, monitoring, etc., so clarity about the population receiving the test is important, since test performance can be affected by setting and population.

When it is proposed that a test should be introduced into healthcare, this may prompt an investigation of how effective the test will be and whether it will represent good value for money. In some cases a new test can be demonstrated to dominate (be better in every way and less costly than) an existing test used in a diagnostic pathway, and the new test can be introduced as a like-for-like replacement. But often new tests can be more expensive, can require changes to pathways, or can be better in some ways but worse in others. In this case, a clear *value proposition* for the test is necessary so that decision makers understand why it may be worthwhile to introduce the test, and an accurate assessment of the value can be undertaken.

Evidence for the effectiveness of diagnostic tests is frequently limited to the *clinical validity* of the test, i.e., how good is it at categorising patients as having or not having the disease (sensitivity, specificity, and related measures), or measuring some quantity relevant to the disease. The test (or set of tests) which define the “true” disease status in clinical validity tests is called the *reference standard*, while the test being evaluated is called the *index test*. Evidence rarely extends to a controlled assessment of how good the test is at producing meaningful benefits to the patient [1] (such studies are referred to as test-treatment or end-to-end studies), so it is impossible to estimate the full effect of a test on costs and health without the use of some modelling assumptions.

This approach of evidence linkage based on clinical validity studies and modelling (sometimes referred to as *indirect* evidence for the clinical effectiveness of a test) is embraced by some, but not all, health technology assessment organisations [2]. If a test-treatment study does exist, it may be possible to use it as the basis for an economic evaluation after due consideration of risks of bias and whether follow-up is sufficiently long to capture all differences in costs and health consequences.

In England, the National Institute for Health and Care Excellence (NICE) Diagnostics Assessment Programme invites clinicians and sponsors to submit diagnostic technologies for assessment, including economic evaluation by an independent technology assessment group (usually in the form of a model-based cost-utility analysis). Decisions by the committee are generally consistent with the application of a £20,000 per quality-adjusted life year (QALY) cost-effectiveness threshold, but with certain decision-modifying factors, such as uncertainty [3]. In Canada, the Canadian Agency for Drugs and Technologies in Health (CADTH) has undertaken health technology assessments of diagnostics, including cost-utility analyses.

Van der Pol et al. [4] have produced guidance on the design and reporting of economic evaluations of diagnostics, focusing rightly on the importance of having a very clear research question. This includes being clear about the population being tested – in which setting have they been identified? what symptoms do they have? Analysts should also ensure their economic evaluations comply with the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) [5].

The main purposes of this article are to describe the methodology commonly used in economic evaluations of diagnostics (particularly modelling methodologies), align these methodologies with common value propositions for diagnostics, highlight issues which may arise in the economic evaluation of diagnostics, and act as a tutorial paper for those interested in model-based economic evaluation of diagnostics. Occasionally the article touches on other uses of tests, e.g., surveillance and prognostic testing.

2. Common methodological approaches for economic evaluations of diagnostics

This section presents three methodological approaches for the economic evaluation of diagnostics. These are focussed on modelling the diagnostic pathway, i.e., determining which patients have a disease, whether there is a timely, correct diagnosis, how the patient will be managed in the future. They do not cover the long-term modelling of the disease (conditional on the outcome of the diagnostic pathway) because this will depend greatly on the nature of the disease, treatments, and the availability of data. Markov models are frequently employed and general best practice for modelling will apply.

2.1 Decision tree

The decision tree has long been used in economic evaluations. It calculates the expected costs and benefits of different competing options as a weighted average of the costs and benefits for different outcomes, where the weights correspond to the probabilities of those outcomes being realised. Decision trees include “chance nodes” which reflect things which are subject to chance, e.g., whether a patient responds to a treatment, receives an organ transplant, or in the case of diagnostics whether a test gives a positive or negative result.

When using a decision tree for an economic evaluation of a diagnostic, there will typically be at least four possible outcomes, corresponding to the combinations of the true disease status and the test result. The best modelling approach is to branch (split) first by true disease status and then by test result, as shown in Figure 1. Note that it does not need to be feasible in real life for patients to be split by true disease status – it is only important that the costs and outcomes can be appropriately modelled based on the true disease status and the results of testing. The key parameters to be modelled then are the prevalence (the probability the patient truly has the disease), sensitivity (the probability of a positive test result in a patient with the disease) and specificity (the probability of a negative test result in a patient without the disease). When modelling multiple competing tests the prevalence will be a common parameter across the different options, while each test will have its own estimates of sensitivity and specificity.

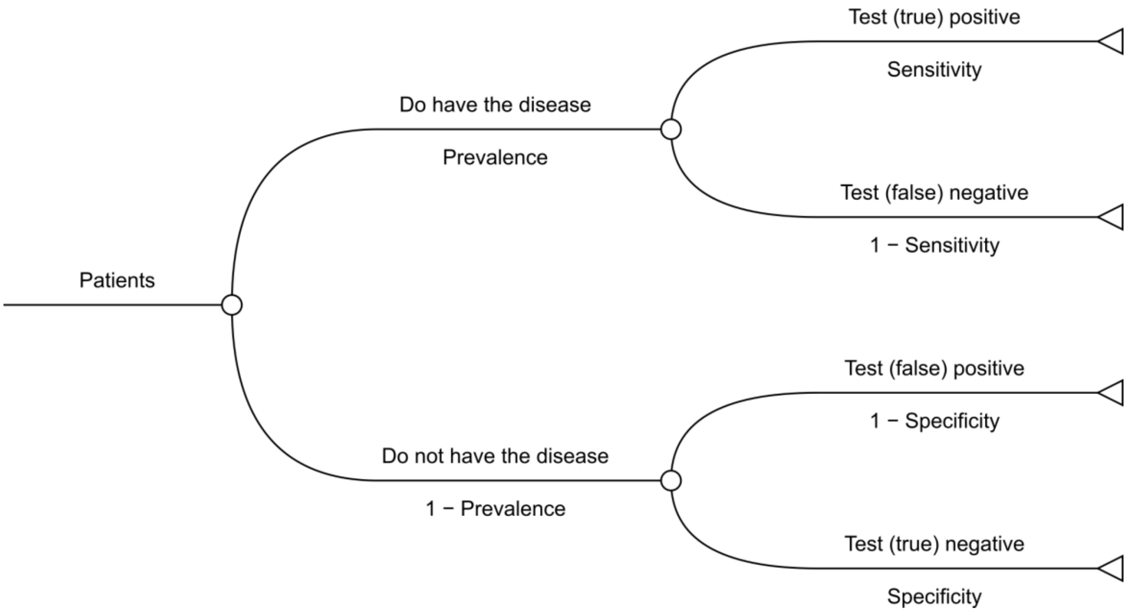

Figure 1: Decision tree for evaluating a diagnostic test

The decision tree component is fairly simple, but a model will also need to estimate the costs and benefits for the different outcomes so that these can be combined. For true positive and true negative results it may be possible to estimate these empirically, but generally assumptions will be required for incorrect test results and a modelling approach should be used.

If two or more tests are used *in combination* then errors can be introduced if sensitivity and specificity estimates are naively combined, i.e., assuming that the sensitivity and specificity measured in the whole population will also give the probability of a correct test result when another test has excluded some of that population [6]. Best practice in this case is to estimate the sensitivity and specificity of tests in populations which have been stratified according to prior test results. There are approaches for meta-analysis which allow the synthesis of studies which evaluate individual tests along with studies evaluating tests in combination [7].

Consider a simulated example where the true disease status can be defined by the true values of two characteristics, X_1 and X_2 (e.g., systolic and diastolic blood pressure). Tests 1 and 2 are imperfect measurements of X_1 (i.e., subject to some measurement error) which use different thresholds and Test 3 is an imperfect measurement of X_2, as shown in Figure 2.

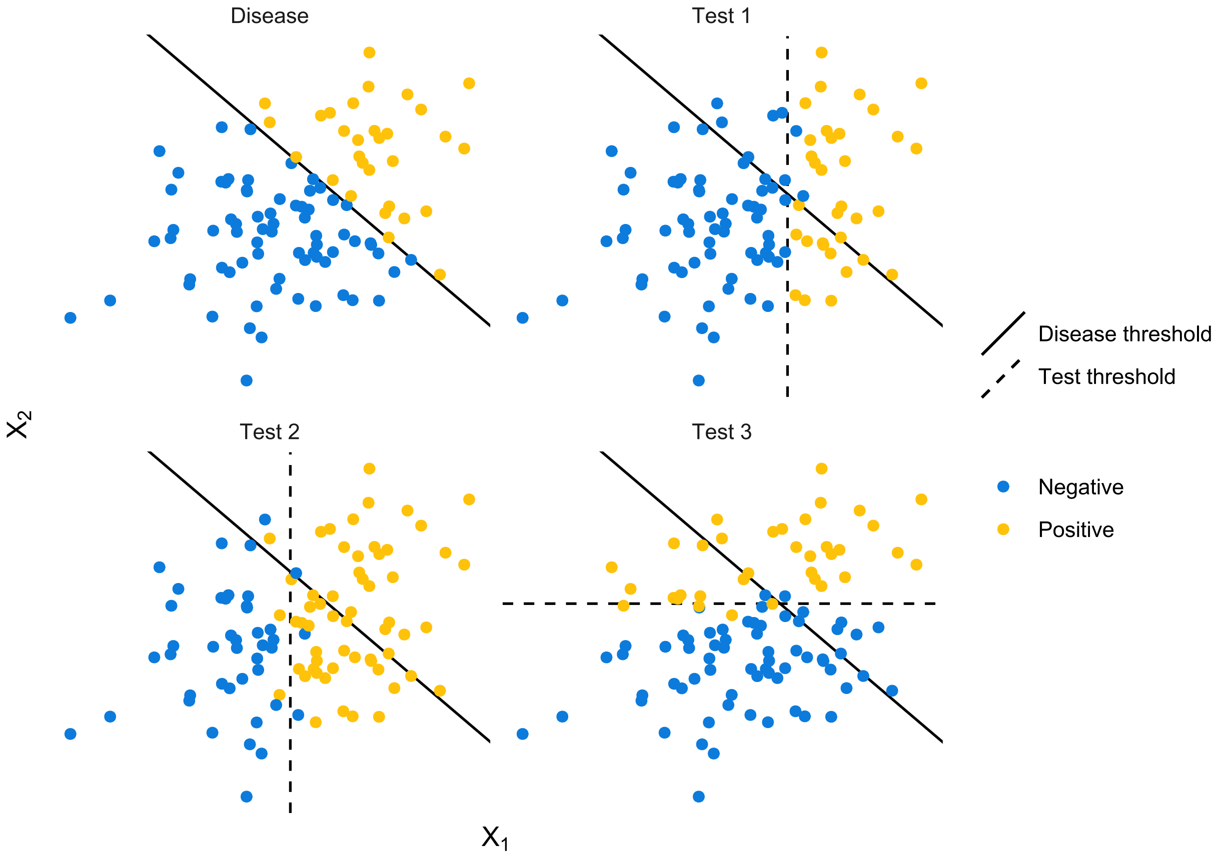

Figure 2: Results of tests in the simulated example

The sensitivity and specificity of each test will depend on whether any previous tests have already been used to stratify the population, as shown in Table 1. Tests 1 and 2 are highly correlated because they rely on measurement of the same characteristic, X_1. Note that even though Test 3 measures a different characteristic to Tests 1 and 2, its sensitivity and specificity are still affected by stratifying on the results of those tests. The sensitivity and specificity of particular combinations of tests are shown in Figure 3. Combining Tests 1 and 2 is pointless unless it results in reduced costs of testing. Combinations of Test 3 with one of Test 1 or Test 2 are particularly effective because Test 3 is measuring the characteristic X_2 and providing significant additional information.

Table 1: Sensitivity and specificity in the simulated example with stratification

		Test 1		Test 2		Test 3	
	Full population	+ve	-ve	+ve	-ve	+ve	-ve
Test 1							
Sens.	0.712	---	---	0.766	0.009	0.642	0.934
Spec.	0.886	---	---	0.665	0.999	0.985	0.870
Test 2							
Sens.	0.929	0.999	0.754	---	---	0.906	0.999
Spec.	0.663	0.006	0.747	---	---	0.788	0.642
Test 3							
Sens.	0.760	0.685	0.945	0.742	0.997	---	---
Spec.	0.858	0.981	0.843	0.911	0.832	---	---

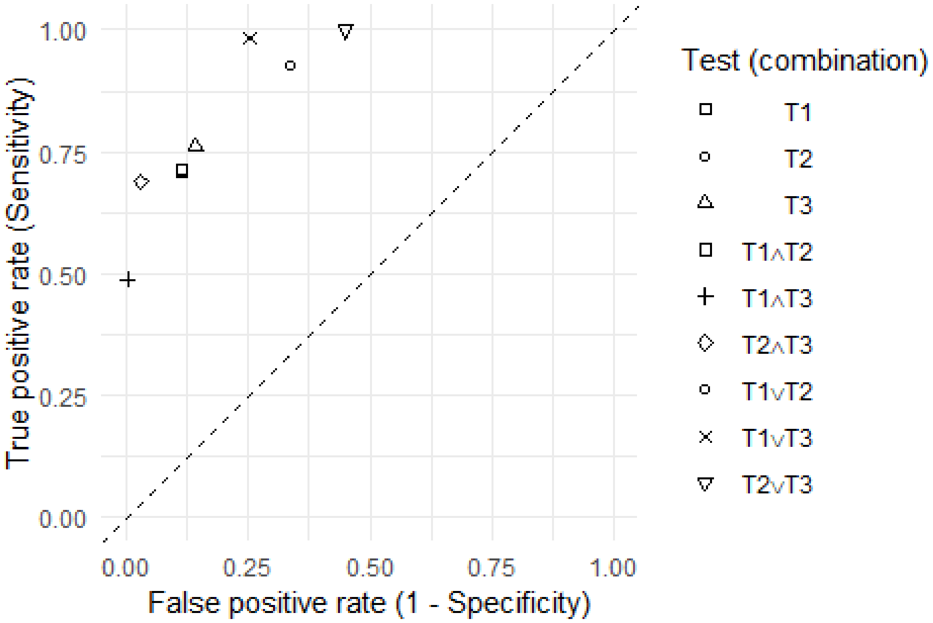

Figure 3: Summary receiver operating characteristic (ROC) plot.

Note: ∨ and ∧ mean OR and AND respectively in the Boolean logic sense, so that the combination T1 ∧ T2 gives a test positive result only if both T1 and T2 give a positive result. The same symbol is used for T1 as for T1 ∧ T2 because the sensitivity and specificity are indistinguishable in the plot (likewise for T2 and T1 ∨ T2).

Decision trees are ideal when the value proposition for the new test does not involve changes to timeframes and clinical pathways because future costs and health outcomes can be assumed to depend only on the result of testing (true positive, false positive, false negative, true negative), and not also contingent on the testing strategy.

2.2 Discrete event simulation

Here we refer specifically to models where the *diagnosis* is modelled using discrete event simulation, not when discrete event simulation is used to forecast future costs and health outcomes which are then combined using a decision tree. A single discrete event simulation can incorporate the diagnostic pathway and long-term outcomes, or another modelling methodology can be used to estimate long-term outcomes.

Discrete event simulations are particularly valuable if diagnosis is time critical or if new technologies may disrupt service pathways.

An instructive example is stroke patients. The management of stroke depends on what has caused it, so determining whether it is an ischaemic stroke (caused by a clot preventing bloodflow to the brain) or a haemorrhagic stroke (caused by bleeding in the brain) is important and must be done in a timely fashion. In England, suspected stroke patients are taken by ambulance to the nearest comprehensive stroke unit or acute stroke unit. If brain imaging suggests that thrombectomy (surgical removal of a clot) is needed, a patient will be transferred to a comprehensive stroke unit if they are not already at one. An alternative strategy, which may reduce the length of time between stroke and thrombectomy (improving outcomes) is to use mobile stroke units (ambulances with onboard CT scanners) to determine whether thrombectomy is indicated on-scene [8] – an economic evaluation of this strategy would likely involve discrete event simulation, since outcomes are so dependent on the time between onset of symptoms and initiation of treatment.

Discrete event simulation is also likely to be important if the use of the test is not diagnostic, but instead testing over time in an at-risk population, e.g., colorectal cancer patients at risk of a second colorectal cancer [9], or screening for cancer in the asymptomatic general population. In this case the discrete event simulation simultaneously allows for the disease state to evolve over time and for the health service to attempt to intercept disease as early as possible.

2.3 Patient-level analysis

If patient-level data are available from a diagnostic accuracy study, and there are no other accuracy studies which could contribute to a meta-analysis, the best approach to economic evaluation may be to base it on patient-level data [10].

Consider the study design shown in Figure 4; for each participant we have the results of index test 1 and index test 2 (two tests which we are considering introducing to clinical practice), and we have the results of the reference standard for any participants with at least one index test positive and a random subsample of participants with both index tests negative. While it would be preferable to have the reference test for all participants, this may not be feasible if the prevalence of the disease is low and the reference standard is costly and/or invasive.

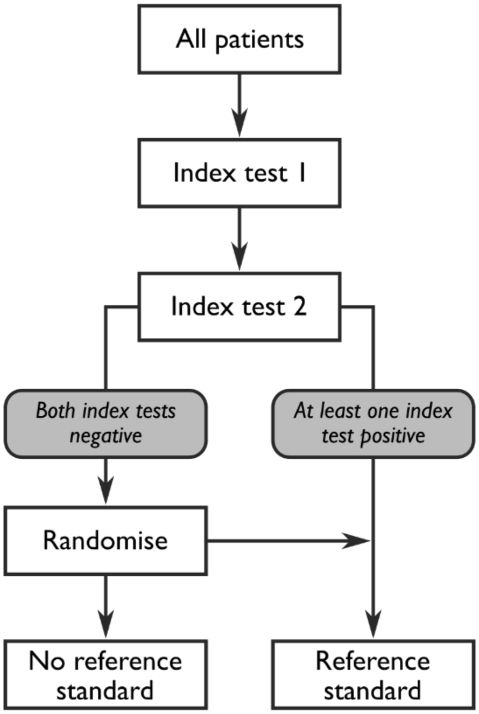

Figure 4: A diagnostic accuracy study design which can underpin an economic evaluation; participants who test negative by both index tests are randomly assigned to receive the reference standard or not

For each participant, we can forecast their future costs and health outcomes based on the results of the reference standard and whether or not they are appropriately diagnosed. If the participant truly has the disease (according to the reference standard) then we forecast the future costs and health outcomes for a true positive and a false negative outcome of testing (and the relevant costs and health outcomes are selected for each strategy based on the test outcomes); if the participant truly does not have the disease we forecast for a true negative or false positive outcome; if the participant did not undergo the reference standard then we can impute the probability they have the disease and proceed accordingly. We can also link the forecasts of future costs and health outcomes to participant characteristics, e.g., age, sex, comorbidities.

Finally, we estimate the costs and health outcomes for each participant under each testing strategy and add the costs of the technologies themselves.

A key advantage of doing the study like this is that associations between participant characteristics, disease characteristics and test outcomes will carry through to associations with costs and health consequences (e.g., QALYs) without us needing to even be aware of them. For example, Lynch syndrome (a hereditary cancer syndrome) is more likely to be present in a younger cancer patient, younger patients have greater potential to benefit from life-long preventive measures, and the specificity of tumour tests for Lynch syndrome decrease with increasing age [10].

3. Value propositions

It is essential that the value proposition for a new technology is understood. Diagnostics can have a wide variety of different value propositions [11], and in some cases (e.g., when conducting an early economic evaluation) this value proposition may not yet be clearly articulated by its sponsors. Healthcare payers may also identify the characteristics of tests they wish to see developed using target product profiles [12].

It is also important to bear in mind that a holistic view of the benefits and *risks or harms* of a technology is necessary. The sales pitch for a technology may omit the downsides – it is essential that the economic evaluation does not.

The following sections outline a variety of value propositions and how these influence the methods of economic evaluation.

3.1 Replacing a test with a cheaper, better test

If the test is intended to replace an existing test, but it is cheaper than the existing test and better (i.e., has better sensitivity and specificity), then it is generally not necessary to do a full cost-effectiveness analysis where future costs and health outcomes are forecast. If these conditions can be demonstrated to hold (at least on the balance of probabilities), then the new technology is dominant. Attention should be focused on whether the new test is superior in all patients, since if the new test is inferior in certain populations (e.g., tests for gynaecological malignancy can be less accurate in premenopausal women), dominance will not hold.

3.2 Replacing a test with a more expensive test

If the new test is intended to replace an existing test, but it is more expensive than the existing test, then a full cost-effectiveness analysis where future costs and health outcomes are forecast is necessary. The new test does not need to have superior sensitivity *and* specificity, but it should be superior on at least one of those measures. A decision tree approach will generally be appropriate.

3.3 Replacing a slow test with a faster test

If the new test is intended to replace an existing test, but it generates results quicker, then it is important to consider what value this actually adds. Some cases where a quicker test may add significant value (to justify a potentially greater cost) are:

- Acute conditions where a diagnosis is needed urgently as the patient’s condition may deteriorate
- Tests used during operations where surgery is paused while test results are produced, since quicker results will reduce operative time

- Point-of-care testing where quicker results can avoid the need for additional consultations or operations (e.g., rapid diagnostics to detect lymph node involvement during breast cancer surgery [13])
- Conditions where current testing times mean patients are left highly anxious for several weeks and this time can be brought down substantially

It is important to remember that the new test may, for example, sacrifice accuracy in favour of speed. An economic evaluation of a test or tests should generally include accuracy and the costs and consequences of diagnostic errors unless there is absolute certainty that tests have totally equivalent accuracy.

3.4 Replacing a test with a more acceptable test

There is no doubt that some medical tests are painful, uncomfortable, or inconvenient. A new diagnostic test may be more acceptable to patients because it is less painful or uncomfortable, or it takes less time out of their day (which would not appear as a cost using the common third-party payer perspective).

If a new test is less painful or uncomfortable, how should this difference in acceptability be incorporated into an economic evaluation? Traditional cost-utility analyses will attempt to estimate a health state utility value for undergoing the test [14–16] and will apply this utility value for the length of the test. However, even if a test leads to a health state utility value worse than death (e.g., -0.865, the lowest utility value in the Indonesian EQ-5D-5L value set [17]), if it only lasts for one hour then it will at most lead to a loss of 0.00032 QALYs, with a monetary value of $15.97 implied by a cost-effectiveness threshold of $50,000 per QALY. Is this a remotely sensible value to assign in an economic evaluation? Craig et al. [18] found significant violations of the constant proportionalilty assumption underpinning QALYs, lending credence to the idea that short term impacts on health-related quality of life may have an outsized influence on preferences. It has been specifically argued that QALYs are not appropriate when the health condition is acute, and argued instead that an alternative methodology such as willingness-to-pay should be employed [19]. There is evidence that for acute conditions, QALYs are not predictive of willingness-to-pay [20], but QALYs are accepted by many policymakers while willingness-to-pay is generally not.

It is likely that when the value proposition for a new diagnostic is that it reduces pain or discomfort, health economists will need to present economic evaluations using QALYs, but should supplement these with analyses incorporating willingness-to-pay.

3.5 Inserting a triage test

Sometimes a new test is not intended to replace an existing test, but to be used before the existing test in a sequence in order to rule out the disease in some patients. Typically this is because the existing test is expensive, invasive or time-consuming. A triage test should be highly sensitive so that the rate of false negatives is controlled. Provided the triage test is rapid a decision tree is likely to be appropriate. The analyst should check that there is no risk of patients “falling out” of the diagnostic pathway because of the addition of an extra test.

3.6 Companion diagnostics

If the sole purpose of a test is to identify patients who can receive a single targeted treatment, the test should be viewed as a companion diagnostic and the economic evaluation should include both the test and the treatment. The population for the economic evaluation should be everybody who would receive the test, not just those who get selected for treatment.

Things get more complicated if one test can determine eligibility for multiple targeted treatments (e.g., DNA mismatch repair deficiency testing to determine eligibility for immunotherapy). If the different targeted treatments are appraised individually (as is often the case with reimbursement agencies), then each of them will have to bear the cost of the companion diagnostic until one of the treatments is reimbursed, at which point the companion diagnostic becomes standard care. From that point on, further treatments being appraised (including treatments once rejected being reappraised) will arguably not have to bear the cost of the companion diagnostic. In order to ensure fairness, the population for these subsequent economic evaluations should be only those selected for treatment (provided the selection criteria are identical for all treatments), all treatments should be included in a single fully incremental analysis, and the cost of the companion diagnostic can be ignored.

3.7 Expanding the population that can be tested

Perhaps the existing test for a disease is so expensive, invasive, or otherwise deleterious that some patients never receive the test. If the patients are symptomatic, they may have their symptoms managed, rather than the definitive cause of those symptoms identified and treated. If the test is instead for a risk factor (e.g., a hereditary predisposition to cancer) then perhaps only individuals with a high chance of having the risk factor undergo testing.

A novel test may mean that the population being tested expands; it is critically important that when conducting an economic evaluation the population should not differ between the study arms at baseline. Figure 5 shows how such economic evaluations should be approached – any characteristics which were previously used to determine who gets testing at present should also be present when estimating costs and outcomes with the novel test. For example, if patients are not currently getting tested because they are at high risk of dying from a comorbidity before they could benefit from treatment of the disease of interest, those patients should still be modelled as at high risk of dying from that comorbidity after having the novel test.

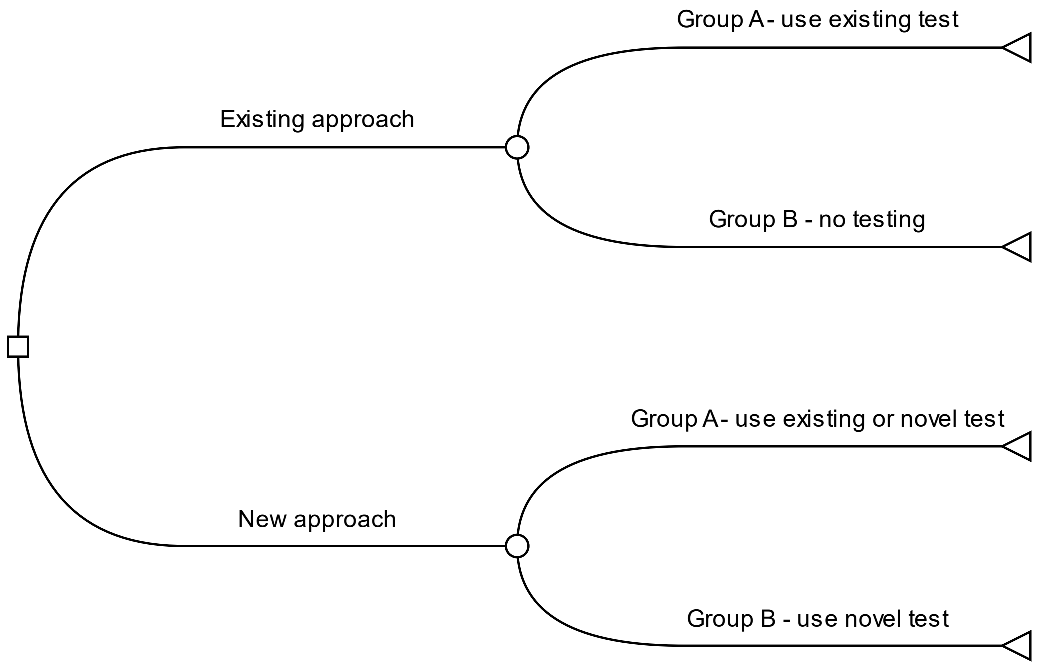

Figure 5: Decision tree when we expand the population being tested

3.8 Lowering the rate of test failures

In this value proposition, the novel test promises fewer test failures. By test failures, we do not mean when a test gives a false positive or false negative result, but when the test fails to give a result at all. The consequences of a test failure can vary substantially. The analyst should consider: can the test even be repeated (or did the test destroy the only available sample)? does repeating the test require the patient to be recalled? is the test very likely to fail again if it already failed once? will a different test be used in the event of a test failure?

3.9 Replacing a test with a cheaper, worse test

If the existing test is very expensive, it may be worth replacing it with a much cheaper test, even if the cheaper test results in worse health outcomes on average. In this case, the new technology is in the south-west quadrant of the cost-effectiveness plane (less costly and less effective). Economic evaluation proceeds as normal, but some thought should be given to whether healthcare professionals will adopt the new test if the existing test continues to be reimbursed.

3.10 Prognostication or prediction

The results of a test may be highly informative for the prognosis for a patient already affected by disease or may indicate their risk of developing future disease. There is evidence that patients are willing to pay for a test which predicts their future risk of disease, even if there is no intervention which can modify that risk, i.e., they are willing to pay for the information alone [21].

As we discussed earlier (Replacing a test with a more acceptable test), just because patients are willing to pay for something does not mean that healthcare payers will, particularly if the effect on QALYs is negligible or even negative (e.g., being told one is at risk of developing a disease in the future may lead to long-term mental health consequences).

4. Populating diagnostic models

The component of a diagnostic model which estimates long-term costs and health outcomes will need to be populated (parameterised) just like any other health economic model, so here we focus on the diagnostic component of the model. We consider the diagnostic accuracy parameters, the pre-test probability, and costs associated with the diagnostic test.

4.1 Diagnostic accuracy

The key diagnostic accuracy parameters will be the sensitivity and specificity of the test, the accuracy of the test among diseased and non-diseased populations respectively. These parameters are usually sufficiently important to justify conducting a systematic review, and a meta-analysis if this is appropriate after studies have been identified. The Cochrane Screening and Diagnostic Tests Methods Group has published a handbook for conducting systematic reviews and meta-analyses of diagnostic tests [22]. Particular attention should be paid to the risk of bias in diagnostic studies.

In an economic evaluation we may be particularly interested in the comparative accuracy of two or more tests, and if there are studies which have evaluated those tests simultaneously then our estimates of the accuracy of each test will not be statistically independent. Our economic evaluation should include the statistical dependence of accuracy estimates because it may have a significant effect on the results, particularly the amount of uncertainty in results. The second edition of the handbook (in draft at the time of writing) contains sections on meta-analysis of comparative accuracy studies but this is an area of active methodological development.

4.2 Pre-test probability

The pre-test probability that a patient has a disease is an essential factor in cost-effectiveness. The pre-test probability of disease across a population is often referred to as the prevalence, although this can be confusing since prevalence has a related but different interpretation in epidemiology.

The pre-test probability can be heterogeneous and depend on the presence or absence of particular symptoms and risk factors. If these factors resulting in heterogeneity in pre-test probability have no influence on the performance of the test or the future costs and health consequences conditional on the test outcomes, then they can be ignored, but this is a strong assumption unlikely to hold. We show, in the Electronic Supplementary Material (a simple decision tree model built in Microsoft Excel®), that it is possible a test appears to be cost-effective when heterogeneity is ignored, when in fact it is not cost-effective in either of the subgroups considered.

If the economic evaluation is for use of a test in the same setting as its diagnostic accuracy has been evaluated, then the prevalence of the disease (according to the reference standard) in that study (or those studies) is a suitable estimate for the pre-test probability. This should be stratified where possible according to known risk factors. If no such studies exist (e.g., if the diagnostic accuracy was estimated using a case-control or two-gate design), then expert elicitation may be an appropriate alternative [23].

4.3 Costs

When assigning costs to resources, it is of course important that these reflect opportunity costs, where possible, and that they follow the economic perspective on costs (e.g., third-party payer, societal). If a diagnostic technology requires very high fixed costs, an approach to allocating those fixed costs to each use of the technology should be adopted [24].

5. Challenges in economic evaluation of tests

5.1 What if there are more than two disease states?

Often diagnostic tests are intended to determine whether a disease is present or absent (two options). But in some cases the disease state may not be binary. For example, pathologists will aim to determine the *histotype*, *stage* and *grade* of lung cancers because these inform prognosis and the most appropriate treatment. Histotype is a categorical classification while stage and grade are ordinal (lower stages and grades are less advanced and less aggressive respectively).

This does not make economic evaluation impossible, but it is important to be able to estimate future costs and health outcomes according to the different disease states (even if with significant uncertainty) and what happens if the healthcare system misclassifies the patient. Once this has been achieved, a decision tree approach is still viable – the population is split initially into the different disease states and then the probabilities of different test results (according to the technology used) determine how patients are ultimately classified and treated.

5.2 What if there are more than two possible test results?

Even if it is agreed that the objective is to determine whether a disease is present or absent, it is possible that the test technology can produce something other than positive or negative. For example, a urine dipstick test can have different strips which indicate the concentration of an analyte is within a particular range, or a test may be fully quantitative, or it may be an imaging test.

A key role for the health economist is to understand how test results are, or could be, used in clinical practice. What test result will lead to a disease being ruled out with no further testing? What test result will lead to commencement of treatment? What test result will lead to further testing? If the pre-test probability of having a disease is heterogeneous (e.g., some patients have more specific symptoms than others, or the disease is associated with age) or the consequences of mistakes are heterogeneous then there may not be a simple answer to these questions, and we may instead need to simulate how test results are interpreted and acted upon.

If we are interested in simulating the interpretation of a test result, Bayes' theorem tells us how to correctly update our belief that a patient has the disease according to the results of a test (though it may not correlate well with how physicians in fact interpret those results [25]). Bayesian interpretation of test results using likelihood ratios is mathematically convenient. The likelihood ratio, $LR^\star$, for a particular test result, $T^\star$, is given by

$$LR^\star = \frac{\Pr(T^\star \mid D^+)}{\Pr(T^\star \mid D^-)}$$

where $\Pr(T^\star \mid D^+)$ and $\Pr(T^\star \mid D^-)$ are respectively the probability of getting the result $T^\star$ in a diseased and non-diseased patient. Then if we have estimated the pre-test probability of the disease as p, the post-test probability of the disease, p', is given by

$$\frac{p'}{1-p'} = LR^\star \cdot \frac{p}{1-p}$$

Note that this works regardless of whether $T^\star$ is positive, negative, takes a semi-quantitative or fully quantitative value.

Even if physicians would not naturally interpret test results using Bayes' rule, it may be feasible to construct and populate a simple model of physician behaviour, or to produce a simple decision aid to accompany the test which correctly applies Bayes' rule.

5.3 Optimising test thresholds

If a test is quantitative (producing a single continuous value, e.g., the concentration of an analyte) then the threshold can be optimised from an economic perspective, by identifying the threshold which leads to the maximum net benefit [26]. Net benefit is the value of health benefits (B) offset by costs (C), and is calculated based on the willingness-to-pay for a unit of health benefit (λ). The net health benefit (NHB) is expressed in units of health benefit, while the net monetary benefit (NMB) is expressed in monetary units:

$$\begin{aligned} \mathrm{NHB} &= \mathrm{B} - \frac{1}{\lambda} \cdot \mathrm{C} \\ \mathrm{NMB} &= \lambda \cdot \mathrm{B} - \mathrm{C} \end{aligned}$$

A test cannot change the true disease status of a patient, it can only change whether the diagnosis is correct or not, i.e., changing the threshold of a test can only turn true positives into false negatives (or vice versa) and true negatives into false positives (or vice versa). What matter, therefore, are the net benefit gained by converting a false positive into a true negative ($\mathit{\Delta}\mathrm{NB}_{D-}$) and the net benefit gained by converting a false negative into a true positive ($\mathit{\Delta}\mathrm{NB}_{D+}$). Then if sensitivity and specificity are functions of the threshold, θ, denoted $\alpha(\theta)$ and $\beta(\theta)$, and the disease prevalence is π, then the optimal threshold ($\theta^\star$) will be determined by

$$\theta^\star = \mathrm{argmax}_\theta\{\mathit{\Delta}\mathrm{NB}_{D+} \cdot \pi \cdot \alpha(\theta) + \mathit{\Delta}\mathrm{NB}_{D-} \cdot (1-\pi) \cdot \beta(\theta)\}$$

In a health technology assessment of different technologies to detect preterm labour, one technology was fully quantitative (quantitative fetal fibronectin, qfFN) [27]. The economic evaluation considered the use of qfFN at thresholds of 10, 50, 200 and 500 ng/ml, but we can use the approach above to estimate the economically optimal threshold. Using linear regression on costs (minus the cost of each test) and QALYs we can estimate that $\mathit{\Delta}\mathrm{NB}_{D+} \cdot \pi \approx 1340$ and $\mathit{\Delta}\mathrm{NB}_{D-} \cdot (1-\pi) \approx 1040$. We use maximum likelihood estimation to estimate the distribution of the analyte conditional on the true disease status and therefore the sensitivity and specificity depending on the threshold:

$$\begin{aligned} \alpha(\theta) &= 1 - \Phi\left(\frac{\log\theta - 5.99}{1.71}\right) \\ \beta(\theta) &= \Phi\left(\frac{\log\theta - 2.83}{2.29}\right) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Based on these assumptions, the optimal threshold is 98.5 ng/ml, as shown in Figure 6. Of course, we have (for the sake of simplicity) ignored uncertainty in the various estimates – the aim should be to select the threshold which gives the greatest expected net benefit, taking the expectation across the distributions representing uncertainty in all parameters.

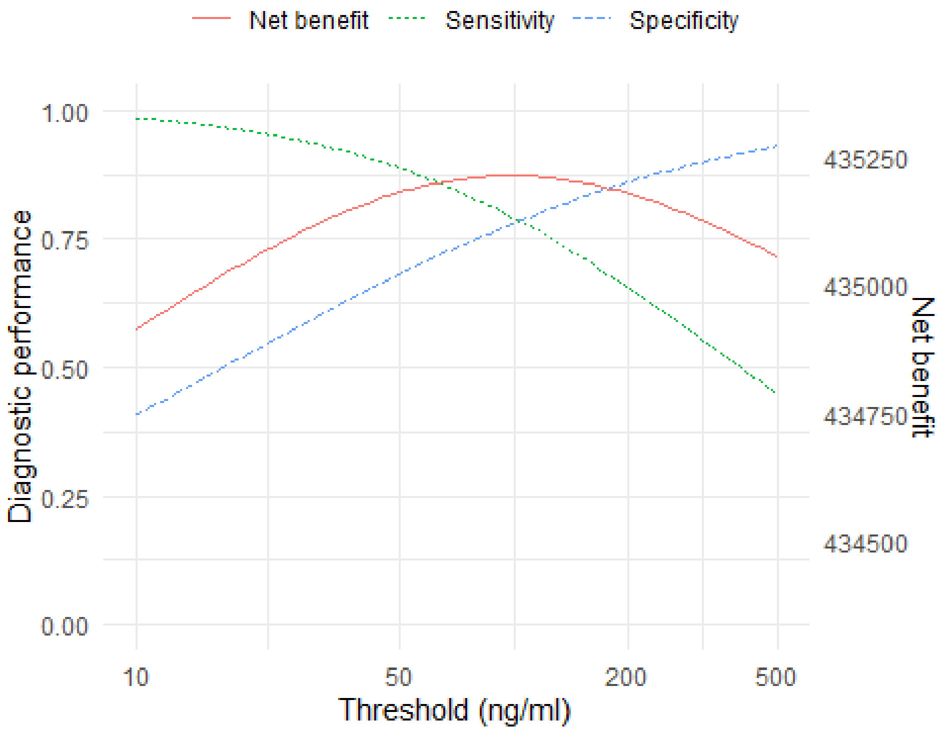

Figure 6: Optimising the threshold for a test based on net-benefit

5.4 What if there is no reference standard?

Soares et al. [28] considered the case of evaluating a test for which there is no reference standard, i.e., no way of knowing whether a patient truly had the disease or not at the time of receiving the test under consideration. They considered prognostic tests (where the test is intended to predict a future outcome rather than the current disease state) to be a special case of this problem.

Their solution to such problems was to abandon modelling the true disease state and for the outcomes to be based only on the test result and whether the patient undergoes treatment. If a test-treatment study happens to exist where patients are randomly assigned to receive a test or not, and lifetime health outcomes and resource use are observed, this approach is entirely appropriate. But in the likely situation that no such study exists, we are left to attempt to model outcomes according to the results of a test rather than the true disease status, which is in conflict with general best practice guidance on modelling [29,30].

An alternative approach is to have some form of latent disease state. For a test intending to diagnose disease now but where there is no reference standard, approaches exist relying on latent class modelling [31]. The results of a prognostic test can also be linked to a modelled latent characteristic of the patient being simulated.

Consider, for example, a prognostic test which gives the estimated probability, $\hat{p}$, that a patient develops a disease within the next five years, unless an intervention is put in place (which will happen if $\hat{p}$ exceeds some threshold). A natural modelling approach may be to use an exponential distribution for the time to developing the disease, where the rate parameter, λ, is taken to be the solution to $p = F(5) = 1 - e^{-5\lambda}$, and p is the "true" risk for a patient with those characteristics. It is likely important to include some error in the transmission between p and $\hat{p}$ since the prognostic model is not perfect, and overestimating or underestimating the true risk for a patient could both lead to losses. Prognostic models can also be imperfectly calibrated [32], meaning that the $\hat{p}$ are systematically biased (at least in some ranges) – this should be represented in a model also.

6. Recommendations and areas for further research

While we have found no consensus recommendations for modelling diagnostics, there is a strong argument that, when a linkage approach is adopted, the economic model structure should include the true disease status, and this should simultaneously drive test outcomes and the future course of the disease (according to whether the patient receives a correct diagnosis at baseline). The possible exception to this is if test-treatment studies exist which are judged to be a superior basis for economic evaluation than the linkage approach. This is especially relevant if there is no “gold standard” for clinical validity studies and/or the ways in which test information can lead to clinical benefit are highly complex or not well understood (e.g., in psychiatry).

Analysts should consult with clinical experts and/or producers of a test to determine its value proposition, since different value propositions may require different modelling approaches. Certain value propositions (e.g., when a new test is more acceptable than an existing test) may not sit well within the standard economic evaluation methodology of cost-utility analysis, where QALYs are the measure of health benefit, and analysts should be prepared to investigate willingness-to-pay as an alternative.

All general purpose good practice guidelines on the conduct and reporting of health economic modelling apply to modelling diagnostics, and the specific recommendations by van der Pol et al. [4] are well-made.

Economic modelling of diagnostics touches on many areas of active research, including approaches to synthesising evidence from diagnostic accuracy studies. Three factors may yet be under-researched in the economic modelling of diagnostics: differential diagnoses, clinical factors, and inconsistent evidence.

By differential diagnoses, we refer to the alternative diseases which could explain the symptoms (assuming that we are evaluating the use of a test in symptomatic patients). Analysts tend to give little thought to these, and how misclassified results are modelled in general. A fairly typical assumption is that a false positive result will lead to some attempt at managing a disease which is not there, but that the mistake will be rapidly corrected, with overall fairly limited effects on costs and usually no effect on QALYs. Analysts should generally question whether this is an appropriate assumption, especially when:

- the disease is rare and false positive findings may be quite numerous;
- treatment for the primary disease in question carries a significant risk of harm;
- an alternative disease which could explain the symptoms is progressive (so there is a risk the real disease will be made worse by the false positive finding leading to a diagnostic delay); or,
- the symptoms are severe and cannot be managed without addressing their true cause.

One example where a number of different diseases were potentially relevant was the use of reflectance confocal microscopy in suspected skin cancers, where some of the clinical validity studies found that technologies sometimes incorrectly diagnosed one type of skin lesion as another, including benign lesions being misclassified as melanoma [33].

By clinical factors, we refer to how a clinical system may not incorporate a test result in the way it is modelled. Frequently a model will assume that once a certain diagnostic pathway is completed, the patient will either be discharged or treated but this may not be an accurate assumption. A clinician may conduct more tests than is assumed, either before discharging a patient (because they perceive the patient is at higher than average risk of the disease or risks worse consequences than average if a diagnosis is missed) or initiating treatment. A clinician may conduct fewer tests than are assumed, e.g., incorrectly treating a triage test as a definitive test. A patient may become “lost in the system” if the diagnostic pathway is complex and they are not highly visible (e.g., an emergency department attendee or an inpatient). Economic models can incorporate such possibilities, but they rarely do.

In the case of inconsistent evidence, we note the possibility that there will be studies which measure only the results of tests (accuracy studies) and others which measure the outcomes for patients (test-treatment or end-to-end studies), and that these studies may give inconsistent findings. For example, diagnostic accuracy studies may show that a new test has improved sensitivity and specificity compared to an existing test, while end-to-end studies show no benefit from the new test. Health technology assessment agencies have not issued any guidance on how to handle such inconsistencies [2]. Test-treatment studies are far from immune from bias [34] and there are a variety of study designs with different advantages and disadvantages, as well as the possibility of adaptive studies [35]. In principle, a Bayesian economic model can incorporate evidence on the accuracy of tests as well as on the longer-term outcomes observed in a test-treatment trial, but this will require strong assumptions about the correctness of the linkage approach and it is unclear how studies at high risk of bias should be handled.

7. Conclusion

Although there have been substantial developments in how evidence for the effectiveness of diagnostics is appraised and synthesised, methods for the economic evaluation of diagnostics remain unstandardised and have not become markedly more sophisticated. There are a number of pitfalls to avoid when modelling the cost-effectiveness or cost-utility of diagnostics, and it is important to understand that not all diagnostics “add value” in the same way.

References

1. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. Journal of Clinical Epidemiology. 2012;65:282–7.

2. Ferrante di Ruffano L, Harris IM, Zhelev Z, Davenport C, Mallett S, Peters J, et al. Health technology assessment of diagnostic tests: A state of the art review of methods guidance from international organisations. medRxiv. 2022;

3. Chen G, Peirce V, Marsh W. Evaluation of the National Institute for Health and Care Excellence Diagnostics Assessment Program decisions: Incremental cost-effectiveness ratio thresholds and decision-modifying factors. Value in Health. 2020;23:1300–6.

4. van der Pol S, Rojas Garcia P, Antoñanzas Villar F, Postma MJ, van Asselt ADI. Health-economic analyses of diagnostics: Guidance on design and reporting. PharmacoEconomics. 2021;39:1355–63.

5. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated health economic evaluation reporting standards (CHEERS) 2022 explanation and elaboration: A report of the ISPOR CHEERS II good practices task force. Value in Health. 2022;25:10–31.

6. Novielli N, Cooper NJ, Sutton AJ. Evaluating the cost-effectiveness of diagnostic tests in combination: Is it important to allow for performance dependency? Value in Health. 2013;16:536–41.

7. Novielli N, Sutton AJ, Cooper NJ. Meta-analysis of the accuracy of two diagnostic tests used in combination: Application to the Ddimer test and the Wells score for the diagnosis of deep vein thrombosis. Value in Health. 2013;16:619–28.

8. Fassbender K, Walter S, Grunwald IQ, Merzou F, Mathur S, Lesmeister M, et al. Prehospital stroke management in the thrombectomy era. The Lancet Neurology. 2020;19:601–10.

9. Erenay FS, Alagoz O, Banerjee R, Said A, Cima RR. Cost-effectiveness of alternative colonoscopy surveillance strategies to mitigate metachronous colorectal cancer incidence. Cancer. 2016;122:2560–70.

10. Snowsill TM, Ryan NA, Crosbie EJ. Cost-effectiveness of the Manchester approach to identifying Lynch syndrome in women with endometrial cancer. Journal of Clinical Medicine. 2020;9:1664.

11. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: A framework for designing and evaluating trials. BMJ. 2012;344.

12. Cocco P, Ayaz-Shah A, Messenger MP, West RM, Shinkins B. Target Product Profiles for medical tests: A systematic review of current methods. BMC Medicine. 2020;18:119.

13. Huxley N, Jones-Hughes T, Coelho H, Snowsill T, Cooper C, Meng Y, et al. A systematic review and economic evaluation of intraoperative tests [RD-100i one-

step nucleic acid amplification (OSNA) system and Metasin test] for detecting sentinel lymph node metastases in breast cancer. Health Technology Assessment. 2015;19.

14. Wright DR, Wittenberg E, Swan JS, Miksad RA, Prosser LA. Methods for measuring temporary health states for cost-utility analyses. PharmacoEconomics. 2009;27:713–23.

15. Ogwulu CB, Jackson LJ, Kinghorn P, Roberts TE. A systematic review of the techniques used to value temporary health states. Value in Health. 2017;20:1180–97.

16. Stoniute J, Mott DJ, Shen J. Challenges in valuing temporary health states for economic evaluation: A review of empirical applications of the chained time trade-off method. Value in Health. 2018;21:605–11.

17. Purba FD, Hunfeld JAM, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Ramos-Goñi JM, et al. The Indonesian EQ-5D-5L value set. PharmacoEconomics. 2017;35:1153–65.

18. Craig BM, Rand K, Bailey H, Stalmeier PFM. Quality-adjusted life-years without constant proportionality. Value in Health. 2018;21:1124–31.

19. Bala MV, Zarkin GA. Are QALYs an appropriate measure for valuing morbidity in acute diseases? Health Economics. 2000;9:177–80.

20. Franic DM, Pathak DS, Gafni A. Quality-adjusted life years was a poor predictor of women’s willingness to pay in acute and chronic conditions: Results of a survey. Journal of Clinical Epidemiology. 2005;58:291–303.

21. Neumann PJ, Cohen JT, Hammitt JK, Concannon TW, Auerbach HR, Fang C, et al. Willingness-to-pay for predictive tests with no immediate treatment implications: A survey of US residents. Health Economics. 2012;21:238–51.

22. Deeks J, Bossuyt P, Gatsonis C, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy [Internet]. The Cochrane Collaboration; 2013. Available from: http://srdta.cochrane.org/

23. Bojke L, Soares MO, Claxton K, Colson A, Fox A, Jackson C, et al. Reference case methods for expert elicitation in health care decision making. Medical Decision Making. 2022;42:182–93.

24. Sassi F, McKee M, Roberts JA. Economic evaluation of diagnostic technology: Methodological challenges and viable solutions. International Journal of Technology Assessment in Health Care. 1997;13:613–30.

25. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychological Science in the Public Interest. 2007;8:53–96.

26. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Medical Decision Making. 2008;28:650–67.

27. Varley-Campbell J, Mújica-Mota R, Coelho H, Ocean N, Barnish M, Packman D, et al. Three biomarker tests to help diagnose preterm labour: A systematic review and economic evaluation. Health Technology Assessment. 2019;23.

28. Soares MO, Walker S, Palmer SJ, Sculpher MJ. Establishing the value of diagnostic and prognostic tests in health technology assessment. Medical Decision Making. 2018;38:495–508.

29. Haji Ali Afzali H, Bojke L, Karnon J. Model structuring for economic evaluations of new health technologies. PharmacoEconomics. 2018;36:1309–19.

30. Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M. Conceptualizing a model: A report of the ISPOR-SMDM modeling good research practices task force–2. Medical Decision Making. 2012;32:678–89.

31. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, de Groot JAH. Latent class models in diagnostic studies when there is no reference standard—a systematic review. American Journal of Epidemiology. 2013;179:423–31.

32. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. JAMA. 2017;318:1377–84.

33. Edwards SJ, Mavranezouli I, Osei-Assibey G, Marceniuk G, Wakefield V, Karner C. VivaScope 1500 and 3000 systems for detecting and monitoring skin lesions: A systematic review and economic evaluation. Health Technology Assessment. 2016;20.

34. Ferrante di Ruffano L, Dinnes J, Sitch AJ, Hyde C, Deeks JJ. Test-treatment RCTs are susceptible to bias: A review of the methodological quality of randomized trials that evaluate diagnostic tests. BMC medical research methodology. 2017;17:1–12.

35. Hot A, Bossuyt PM, Gerke O, Wahl S, Vach W, Zapf A. Randomized test-treatment studies with an outlook on adaptive designs. BMC medical research methodology. 2021;21:1–12.

Declarations

Funding None

Conflicts of interest The author declares no conflict of interest

Ethical approval Not applicable

Consent to publish Not applicable

Consent to participate Not applicable

Author contributions TS is the sole author

Acknowledgements

I thank Prof Chris Hyde (University of Exeter) for reviewing a draft of this manuscript, which prompted valuable discussion.

Data availability

There are no data included in this manuscript that are not in the public domain. The simulated example shown in Figure 2 can be replicated by drawing (X_1, X_2) from a bivariate normal distribution with mean $(0,0)$ and covariance matrix $\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$, then simulating $Z_1 \sim \mathcal{N}(X_1, 0.2), Z_2 \sim \mathcal{N}(X_1, 0.2)$ and $Z_3 \sim \mathcal{N}(X_2, 0.2)$, and finally simulating the test results as $Y_1 = I(Z_1 > 0.6)$, $Y_2 = I(Z_2 > 0)$ and $Y_3 = I(Z_3 > 0.5)$. The true disease status is given by $I(X_1 + X_2 > 1)$.

Electronic Supplementary Material

Example decision tree implemented in Excel showing the importance of accounting for heterogeneity in pre-test probability