
Article:
Soares, Marta O orcid.org/0000-0003-1579-8513, Walker, Simon orcid.org/0000-0002-5750-3691, Palmer, Stephen J orcid.org/0000-0002-7268-2560 et al. (1 more author) (2018) Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. Medical Decision Making. pp. 495-508. ISSN 1552-681X

https://doi.org/10.1177/0272989X17749829

Establishing the value of diagnostic and prognostic tests in Health Technology Assessment

Running title: Value of diagnostic and prognostic tests

For submission to MDM

Soares MO*, MSc: Centre for Health Economics, the University of York, UK; marta.soares@york.ac.uk

Walker S, MSc Centre for Health Economics, the University of York, UK; simon.walker@york.ac.uk

Palmer SJ, MSc: Centre for Health Economics, the University of York, UK; Stephen.palmer@york.ac.uk

Sculpher MJ, PhD: Centre for Health Economics, the University of York, UK; mark.sculpher@york.ac.uk

* corresponding author

This study was unfunded. The authors were independent in designing the study, interpreting the data, writing the manuscript and publishing it.

ABSTRACT

In the recent years, Health Technology Assessment (HTA) processes specific to diagnostics and prognostic tests have been created as a response to an increasing pressure on health systems to decide not only which tests should be used in practice, but also the best way to proceed, clinically, from the information they provide. These technologies differ in the way value is accrued to the population of users, by depending critically on the value of downstream health care choices. This paper defines an analytical framework for establishing the value of diagnostic and prognostic tests for HTA in a way that is consistent with methods used for the evaluation of other health care technologies. It assumes a linked-evidence approach where modelling is required, and incorporates considerations regarding a number of different areas of policy such as personalised medicine. We initially focus on diagnostic technologies with dichotomous results, and then extend the framework by considering diagnostic tests that provide more complex information, such as continuous measures (for example, blood glucose measurements) or multiple categories (such as tumour classification systems). We also consider how the methods of assessment differ for prognostic information or for diagnostics without a reference standard. Throughout, we propose innovative graphical ways of summarising the results of such complex assessments of value.

INTRODUCTION

Funding decisions regarding health technologies are increasingly supported by an explicit examination of the available evidence, with the aim of determining which technology is expected to confer most value for use in clinical practice – a process called Health Technology Assessment (HTA). This process is well established for medicinal products such as drugs, and is explicitly used to support policy decisions in many jurisdictions(1). However, some jurisdictions have extended their scope and defined separate HTA processes for diagnostic (including screening) and prognostic technologies, examples being the Diagnostics Assessment Programme in the UK(2) (created in 2010) and the HTA of co-dependent technologies in Australia(3) (created in 2011). This responds to the increasing pressure on health systems to decide not only which tests should be used in practice, but also the best way to proceed, clinically, from the information they provide; HTA provides the ideal framework to inform these two interrelated questions.

Diagnostics and prognostics (for simplicity we will refer to these as *tests*) are clinical investigations that provide information on the patient, the patient's health, or on the (observed or expected) effects of treatment. This information may contribute to diagnosis by helping to detect or to exclude disease, or to prognosis by predicting the chance of relevant future health outcomes in a particular patient.(4) These technologies differ from medicinal products in a crucial way, relating to the indirect mechanism of accruing value which has been recognised by HTA agencies: "most outcomes of interest [from diagnostic tests] follow from treatments that are either initiated or not initiated based on the results of the tests",(5) and there is thus "the need to consider the benefits of their joint use [of the diagnostic technology and the treatment], as distinct to the benefit of each technology in isolation"(6). Basing decisions on these technologies over value to patient outcomes such as Quality Adjusted Life Years (QALY) has been increasingly recognised, not just in the policy context(7, 8) but also in the context of the design of evaluative research(9-11). However, while the principles for a distinct HTA process may be well justified, there is insufficient guidance on how specifically to adapt the methods of HTA to tests, in a way that reflects the features of these technologies and to allow decision makers to clearly understand the drivers of value.(8, 12) A better, and more integrated, evaluation can also align interests across stakeholders.(8, 13, 14)

Early work by Phelps and Mushlin(15) set out such an approach to the evaluation of diagnostic tests using cost-effectiveness as a basis. The authors focussed on a single dichotomous diagnostic test, i.e. one that distinguishes two subgroups such as presence or absence of disease. Also, the authors' viewpoint was from a Research and Development (R&D) context and considered, for example, pricing strategies and development priorities. Despite some additional work specific to a particular application(16), there has been limited effort to extend such a framework in a comprehensive and general fashion.

This paper thus aims to define an analytical framework for establishing the value of diagnostic and prognostic tests for HTA in a way that is consistent with methods used for the evaluation of other health care technologies. It extends the work by Phelps and Mushlin to consider i) the HTA perspective, typically focussing on informing funding decisions, ii) the specificities of prognostic, and not just of diagnostic tests, and iii) the fact that tests are becoming increasingly complex, with results going beyond dichotomous. We begin by clearly laying-out the specific characteristics of these technologies that are of relevance to HTA, notably, the mechanism of accrual of value. Then we present a methodological framework for HTA of tests, first by focussing on technologies with dichotomous results. We then extend the framework by considering tests that provide more complex information, such as continuous measures or multiple categories (such as tumour classification systems). Throughout, we suggest graphical ways of summarising the results of such complex assessments of value.

MECHANISM OF VALUE FOR DIAGNOSTICS AND PROGNOSTICS

Within HTA, technologies can be considered of value if they present health benefits to the patient population that will receive them (benefits discounted of any harms, or the 'health' impact of the technology). Some jurisdictions consider, alongside health impact, the expected health losses to other patients from displaced treatments as result of any additional funding needs (a net health benefit approach).(17) This paper will consider either case, as long as the metric of value can be represented using a single unit, encompassing health or net health. The framework is equally applicable where a wider perspective on benefits is taken or where patient preferences are incorporated into metrics of benefit.

In terms of the health impact of diagnostic and prognostic technologies, the mechanism of value accrual is more complex to that of other health care technologies. Diagnostic and prognostic technologies identify the level or magnitude of attributes that determine (diagnostic) or predict (prognostic) health outcomes with and without treatment (where treatment refers to a course of action which potentially impacts on health, for example, treatment with a drug, undergoing a biopsy, adopting a life-style intervention or even watchful waiting). While these technologies may affect health directly (e.g. as a result of adverse events from undergoing a test procedure), their main value typically lies in identifying patients expected to benefit from distinct treatments (or other regimens of health care such as strategies for the prevention of disease, illness or injury).(18) The mechanism by which value is generated is not direct, but instead arises from tailoring treatment decisions to patient characteristics. It can be structured using three interlinked components(15):

i) *Classification*: When a test is applied to individual patients it may return one of a set of possible results. In some cases, the test results can be directly used to define the patients being treated, as is the case with a test returning a positive or negative result. However, where treatment options are

fewer than the number of possible results from the test, there is the need to apply a classification rule. The classification rule pools some of the test results to identify the groups of patients who will be treated differently – for convenience we refer to these as treatment groups. For example, for a test reporting results on a continuous scale, a cut-off may be needed to classify patients into two treatment groups (i.e. those above and below the cut-off), one of which may benefit more from treatment than the other. It is, however, important to evaluate the use of alternative classification rules. For example, we could alternatively choose not to treat patients with extremely high or extremely low test results, if that proves more valuable, and use two cut-off values instead of one. Patients with low and high test results, not treated, are, according to our definition part of the same treatment group. However, given these two subgroups may have different prognosis and resource use profiles, they may need to be considered separately in a decision model.

ii) *Choice*: When testing, and after knowing the results of the test and classifying patients into treatment groups, there are choices to be made about which treatments those treatment groups will receive. Such choices are critical as the value of a test often does not depend on the test itself but on the capacity to benefit from better therapeutic choices. This component thus relates to therapeutic choices made for each treatment group after knowing the classification.

iii) *Outcomes*: This component involves the quantification of the consequences of treating the distinct treatment groups (or subgroups within these if the model is non-linear) in terms of (net)health. This is an important component as the value of testing is bound by the outcomes associated with each treatment option.

These multiple, interlinked, components make these technologies distinctive from treatments(19). In practice, these principles should be used when specifying the decision problem, the clinical pathways for modelling and the value proposition of the new test (9, 12). Where evidence is available separately for each of the components of value, (net)health can be determined by mathematical models using the relations established above – the generally endorsed linked-evidence approach (5, 19-21) akin to a decision modelling approach often used in cost effectiveness.(22)

Diagnostics and stratified medicine

The literature on the value of heterogeneity and stratified decision making directly relates to this mechanism of value accrual with diagnostic and prognostic tests.(23-25) Heterogeneity is defined as the variation in outcome of a population (variability) that can at least partly be explained by some attribute of interest.(26) Heterogeneity is valuable insofar as it allows treatment decisions to be stratified across different subgroups so as to generate gains in (net)health; but, for heterogeneity to be identified, tests need to be applied that identify the subgroup an individual patient belongs to. However, the body of literature on evaluation of heterogeneity(27, 28) is seldom explicit about the need to test, and whilst there may be some consideration for the direct health effects of testing such as adverse events and the costs of the tests, it rarely makes explicit the potential for classification

errors and the fact that there may be alternative tests. For the purpose of this paper, we will interpret the value of heterogeneity as the value of optimal treatment once patients have been correctly identified for the heterogeneous attribute of interest (value of the perfect testing). Such value is unlikely to ever be realised in practice, but it is important to have an idea of the potential gains from a perfect test.

With the recent developments in genetic profiling, another policy area receiving much attention is personalised care(8 , 14, 29), which considers treatment stratification to its' limit of the individual. Despite full personalisation still being some way off achievable, it is believed that the identification of relevant sources of heterogeneity through pharmacogenetics will allow further targeting of R&D to allow for full personalised care.(30) A few such developments have been through HTA processes, most comprising of targeted treatment to patients more likely to respond as identified through a genomic test: recent examples are trastuzumab (e.g. NICE TA257), imatinib (e.g. NICE TA209) or gefitinib (e.g. NICE TA258). The HTA of co-dependent technologies in Australia was created in response to these test/treat strategies.(3)

Identifying relevant options for comparison when establishing value

Value is in essence relative and can only be determined in comparison to something else. The well-established HTA process for treatments specifies that all possible alternatives should be included in an appraisal so that the value of each can be compared(31). This same principle should also apply to tests. Options that do not involve testing should be considered alongside those involving relevant tests (including tests focussing on different attributes that can also explain heterogeneity) and the possibility of using multiple tests or sequences of tests. For each test, alternative classification options from test results should be considered, and alternative treatment decisions should also be evaluated. Therefore, all possible combinations of tests, classifications and treatments need to be considered.

VALUATION FRAMEWORK

Dichotomous test results

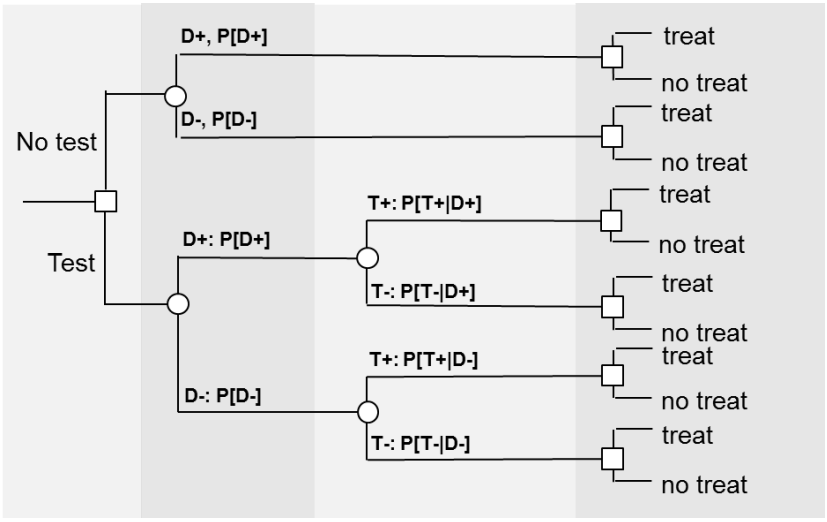
We start by revisiting the Phelps and Mushlin(15) framework (based on a single dichotomous diagnostic test and a single treatment option), taking an evaluative perspective instead of an R&D perspective. True 'disease' status is assumed known in evaluating the tests through the concomitant application of a reference test, a test that is definitive for a particular disease.

Figure 1 (top panel) represents the decision problem using the conventional decision tree diagram. Such diagrams are useful to identify the different options for comparison, and to list and structure the inputs needed to evaluate the (net)health impact of the alternative options. The first node from the left (a decision node) represents the decision of whether or not to use a certain test. Note that, for simplicity, the diagram does not include the option of using the reference standard for diagnosis,

although this should be considered where relevant. The second node identifies the true status of patients with diseased patients identified as D+ and healthy patients as D-; this is a stochastic node reflecting the expected likelihood of any single patient being diseased (probability of disease (P[D+], or prevalence). The third node available describes the distribution of test results within disease status groups. In the case of a dichotomous test result no classification rule is required. Thus, the test directly classifies patients into two treatment groups, one where the test is positive, T+, and another where the test is negative, T-. The decision tree in Figure 1 includes the test results conditional on the true disease status – i.e. the test performance (its accuracy). To describe performance, two quantities are often used: sensitivity, referring to the probability that the test of interest is positive when the disease is present (true positives, P[T+|D+]); and specificity, which refers to the probability that the test is negative when the patient is not diseased (true negatives, P[T-|D-]).(32) With an 'imperfect' test some of the patients may be incorrectly identified, in which case sensitivity and/or specificity will assume values lower than 1.

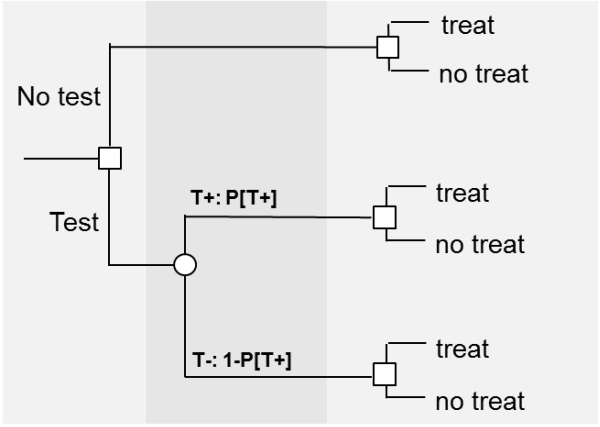
Figure 1: Decision tree schematic: dichotomous results and two treatments available

Diagnostic with reference standard



Inputs	Values assumed
P[D+]	0.6
P[T+|D+]	0.6
P[T-|D-]	0.95
$NH_{D+,treat}$	1.8 QALY
$NH_{D+,notreat}$	1 QALY
$NH_{D-,treat}$	1.4 QALY
$NH_{D-,notreat}$	2 QALY
NH_{test}	-0.01 QALY

Prognostic or diagnostic without reference standard



Inputs
P[T+]
$NH_{T+,treat}$
$NH_{T+,notreat}$
$NH_{T-,treat}$
$NH_{T-,notreat}$
NH_{test}

Treatment choices follow classification. If the test is not made available, should all or no patients be treated? When testing, should only patients who tested positive be treated, or also those who tested negative? Setting up such an explicit decision tree allows the exhaustive identification of all options for treatment -- if only one treatment is available (as is considered in Figure 1) six alternative decision options can be defined for comparison. These are listed in Table 1 with numbers 1 to 6.

Fully structuring the problem using a decision tree also allows the identification of parameter inputs, estimates of which are needed for evaluation. Probability parameters inform stochastic nodes, and each is formulated as conditional on reaching the previous node on the pathway. For example, the probability of having a positive test is conditioned on disease status being positive, and alternatively, negative. Also, all possible pathways defined by the tree will need to be associated with expected health and cost outcomes (outcomes component). By knowing true disease status, the outcomes of treatment can be comfortably assumed independent of test results and only conditioned on true disease status. For the example in Figure 1, evidence on the outcomes of the following patients would be needed: D+ and treated, D+ and not treated, D- and treated and D- and not treated. These highlight the population-level trade-offs imposed by the imperfect tests: the gains of treating diseased patients compared to not treating them are expected to be valued at 0.8 QALY (1.8 QALY on average for diseased patients treated minus 1.0 QALY for diseased patients left untreated), as opposed to the 0.6 QALYs expected to be lost by treating healthy patients in comparison to not treating them (2 QALY for healthy patients untreated minus 1.4 QALYs for those inappropriately treated). Any direct (adverse events) and indirect (opportunity costs) effects of the tests should also be considered.

Evaluating total (net)health for each possible decision option consists of rolling-back the decision tree using the estimated parameter inputs described above; a detailed explanation of the calculations required and the results for the hypothetical example are shown in Table 1.

Table 1: Total (net) health associated with alternative strategies: dichotomous test results and single treatment

strategies	Test	treat if t+	treat if t-	tNH, QALY	Incremental tNH
strategies not involving testing					
1	No test	Yes	Yes	1.64	
2	No test	No	No	1.40	
strategies with test A alone					
3	Test A	Yes	Yes	1.63	
4	Test A	Yes	No	1.666	
5	Test A	No	No	1.39	
6	Test A	No	Yes	1.354	
			value of test A (4 vs. 1)		0.026
Strategies with Reference Standard (RS)					
7	RS	Yes	No	1.68	
8	Test A followed by RS (on - to A)	Yes	No	1.734	
9	Test A followed by RS (on + to A)	No	No	1.602	
			value of testing (8 vs. 1)		0.094
P	Test without misclassification, without costs or direct health effects)			1.88	
			value of heterogeneity (P vs. 1)		0.24

Notes on calculations: The (net)health of option 1 (not testing but treating all) can be calculated by considering the outcomes of diseased and healthy patients after treatment (i.e. 1.8*0.6+1.4*0.4), totalling 1.64 QALY. For options that involve testing, evaluating the decision tree implies determining the proportion of patients correctly and incorrectly classified according to disease status, and attributing the appropriate health outcomes. For option 4 (treating only if test is positive), the outcomes of each of four groups of patients are summed: (i) diseased patients that have been treated, $P[D+] \times P[T+|D+] \times NH_{D+,treat}$, (ii) diseased patients that have not been treated, $P[D+] \times (1-P[T+|D+]) \times NH_{D+,no\ treat}$, (iii) healthy patients that have been treated, $(1-P[D+]) \times (1-P[T-|D-]) \times NH_{D-,treat}$ and, finally, (iv) those healthy patients that have not been treated, $(1-P[D+]) \times P[T-|D-] \times NH_{D-,no\ treat}$. Additionally, the (net)health associated with the test itself is important, NH_{test} (here is assumed to be associated with -0.01 QALY). For option 4, the total (net)health totals 1.666 QALY.

Out of the 6 options (1 to 6), option 4 is the one that confers maximum population value and, as such, would be expected to be funded by the health care system. The value of the diagnostic test can be determined by comparing option 4 (i.e. the testing strategy associated with the highest value) with option 1 (i.e. the option that does not involve testing with the highest value) – and for this example the value of the diagnostic test is 0.026 QALY per patient.

It may also be of interest in an evaluation to describe the value of heterogeneity. This entails establishing the value of perfectly distinguishing subgroups and treating them appropriately. This can be simply done by assuming perfect sensitivity and specificity (both with probability of 1), and no costs or adverse events of testing (i.e. $NH_{test} = 0$) -- strategy P in Table 1. The value of heterogeneity (value of the perfect test) compares the value of P to the value of the best non-testing strategy (strategy 1 in our example). For the example in Figure 1, the value of heterogeneity is 0.24 QALY. The value of the perfect test provides a necessary condition for establishing the value of any new test (i.e. if the test costs more than this it can never be of value). The value of heterogeneity is much larger than the 0.026 QALY conferred by the imperfect test A, indicating how much is lost from current suboptimal strategies

Table 1 includes, for completeness, strategies 3 and 5, where all patients are tested but regardless of its results all are treated, or not treated, respectively. Such strategies are, in this example, always

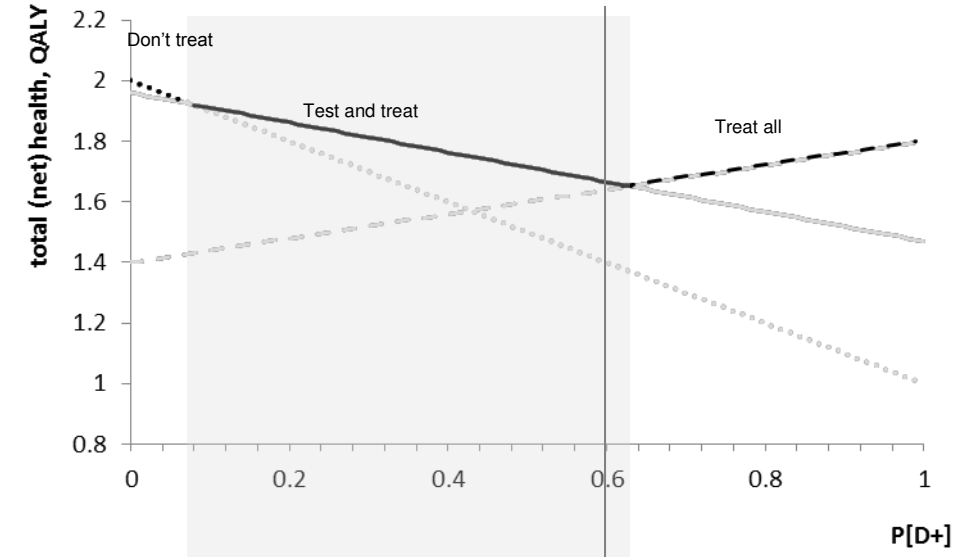
associated with worse (net)health outcomes than strategies 1 and 2, respectively, due to the test's costs or adverse events. These strategies could have been reasonably excluded from consideration a priori.

Drivers of value

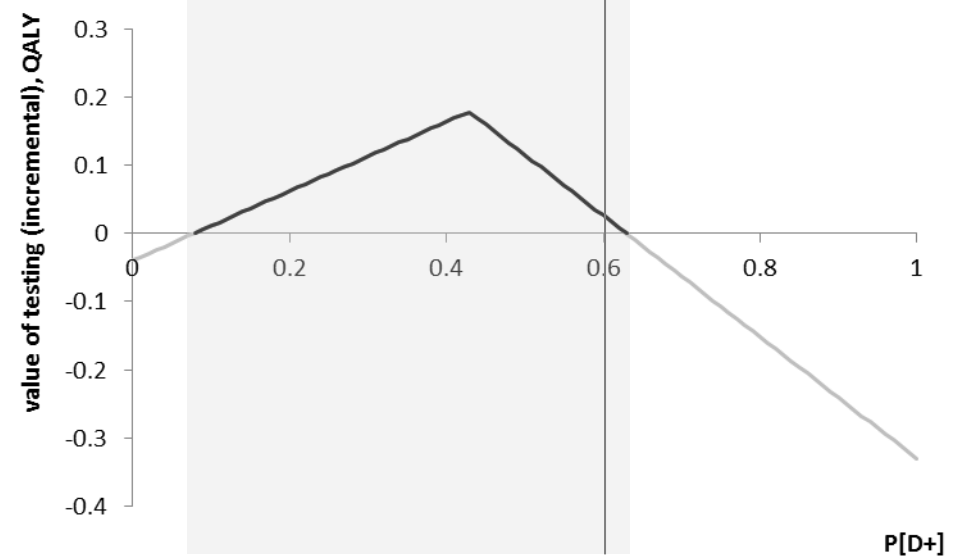
Given the mechanism of accrual of value specific to tests, there are three main aspects driving value. The first is the prevalence of disease. The higher is the prevalence the more likely it is that 'treating all' is preferred to testing – ultimately, as prevalence approximates 1, the costs that testing imposes may not compensate the ability to distinguish the few existing healthy patients. Conversely, the lower the probability of disease the more likely it is that 'treating none' is preferred to testing. Figure 2a illustrates the implications of varying the prevalence in the example above, ceteris paribus, in terms of (net)health for strategies 1, 2 and 4. The (net)health of the strategy that offers best value for each prevalence figure is highlighted in black, and shows that it is worth testing for a range of prevalence values between 8% and 63% (shaded grey area). Figure 2b shows an alternative way of presenting these results, analogous to the graphical displays Phelps and Mushlin proposed(15): instead of the absolute (net)health, it presents the incremental (net)health in relation to the next best non-testing strategy.

Figure 2: Critical aspects of value: disease prevalence (ceteris paribus in relation to Figure 1).

(a) total (net) health



(b) incremental total (net) health of testing in relation to next non-testing strategies -- value of testing

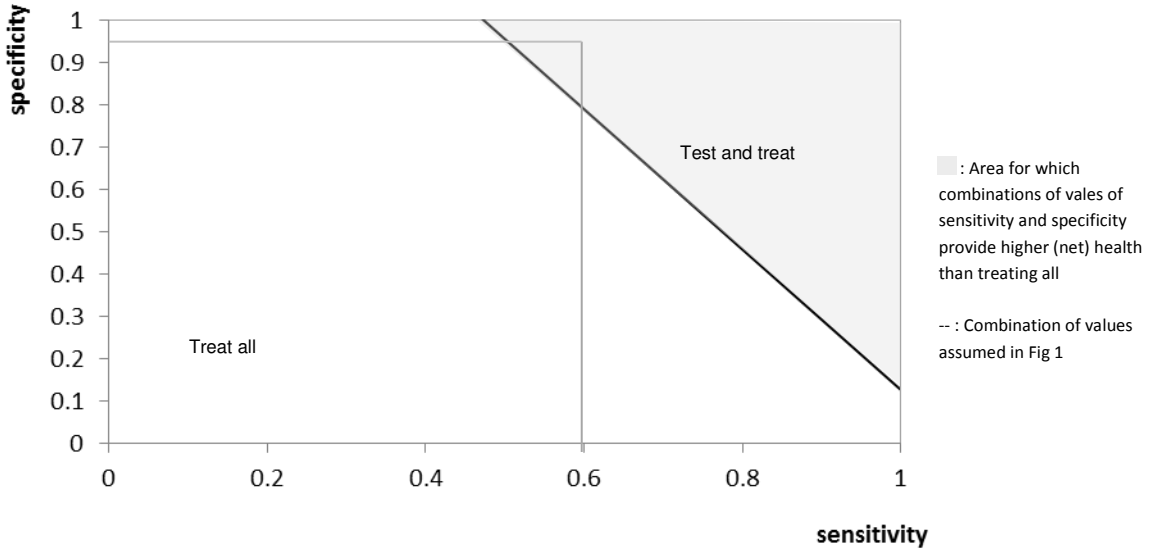


: Range of values of probability of disease which support the use of a testing strategy

Another important driver of value is the performance of the test, i.e. its accuracy. Again using the example in Figure 1, sensitivity and specificity can be varied, ceteris paribus, to show their impact on results. The boundary of acceptance in Figure 3 can be used to identify the combinations of sensitivity and specificity that make testing worthwhile where prevalence is held fixed. The boundary represents combinations of sensitivity (x-axis)/specificity (y-axis) for which we should be indifferent between testing and treating all based on (net)health. Therefore, combinations to the right side of the line lead to recommendations to test, combinations to the left lead to recommendations of treating all.[1] The grey line identifies the sensitivity and specificity values assumed in Figure 1.

[1] Note that testing is compared to treat all because, at the assumed prevalence, this is the best non-testing option.

Figure 3: Critical aspects of value: test performance (ceteris paribus in relation to Figure 1).



The other main key driver of value is the value of the available treatments themselves. That is, the magnitude of health outcomes with each decision option (including the no-treatment option where relevant). For the example in Figure 1, a strategy of testing (strategy 4) is only of value insofar as the losses in health imposed by not treating some of the diseased (imperfect sensitivity) and treating some of the healthy (imperfect specificity) and additional losses due to costs of direct health effects of testing are greater than the losses imposed by treating all of the healthy patients who would be treated under strategy 1. This means any changes to the outcomes of treatment may affect the value of testing.

Sequences of tests

Continuing with the dichotomous test example, but now considering the reference standard test as an alternative test (here denominated test R), its use in isolation or in sequence with the imperfect test (here denominated test A) also become relevant. The reference test is, in this example, associated with significant direct health losses (NH_R = 0.2 QALY) but, by definition, has perfect sensitivity and specificity. In this context, options 7, 8 and 9 were deemed as relevant options for comparison (see Table 1). Strategy 7 uses the reference test alone, and strategies 8 and 9 use it only in patients who tested negative and positive to test A, respectively. Note that other options were omitted, for example those involving treating patients classified as negative[2]. The strategy conferring the most (net)health is strategy 8, and the value of testing when the options are extended in this way is higher at 0.094. In relation to the use of test A in isolation, the sequence represents an almost fourfold improvement in (net) health gains. However, it still falls short of the value of heterogeneity, evaluated at 0.24 QALYs

[2] This option can safely be ignored only if outcomes of treating are undoubtedly better in the diseased, and equal or worse in the non-diseased

in this example, indicating that further developments of the tests involved (e.g. aimed at reducing adverse events of the RS or reducing misclassification with test A) could be of value.

How do prognostics differ from diagnostics?

The above principles on the accrual of value hold true for both diagnostics and prognostics(25). However, prognostics differ from diagnostics in that they aim to predict uncertain events occurring in the future that cannot be known at the stage of testing.(33) The outcomes component cannot be assumed independent of test results. An important consequence is that evidence becomes specific to the context of the primary research informing outcomes, limiting generalisability.(34) The same is true for diagnostic tests without a perfect reference standard. The bottom panel in Figure 1 illustrates how the decision tree diagram would change for a prognostic measure (and a diagnostic test without a perfect reference standard), and lists the inputs required.

Beyond dichotomous

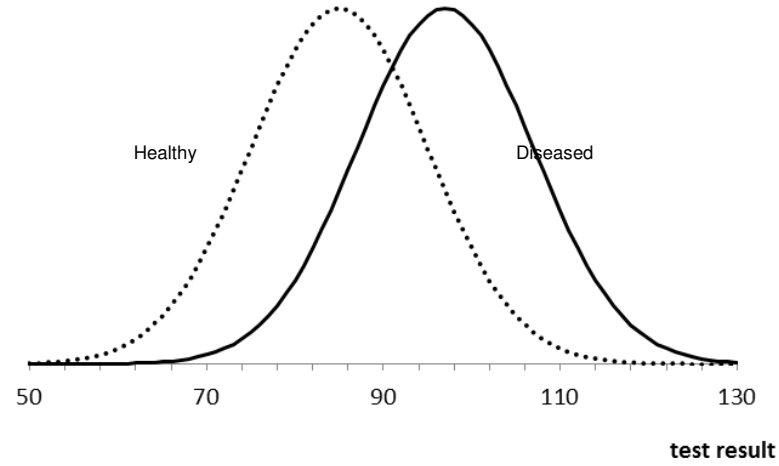
Clinical investigations are increasingly complex, with many moving beyond simply detecting the presence or absence of a disease. One aspect of this complexity is the format of test results, which may range from descriptive information (e.g. signs and symptoms), to continuous results (e.g. physiological quantities such as cholesterol or blood pressure), to complex imaging technologies or composite measures using a series tests (e.g. TNM Classification of Malignant Tumours). Given that clinical policy can only consider as many treatment groups as the number of treatment strategies available (including no treatment), more complex test results will require more nuanced classification rules. In this section, we will examine how value can be established when tests go beyond reporting dichotomous results. For illustrative purposes we will focus on tests reporting continuous results.

Diagnostics for which there is a reference standard

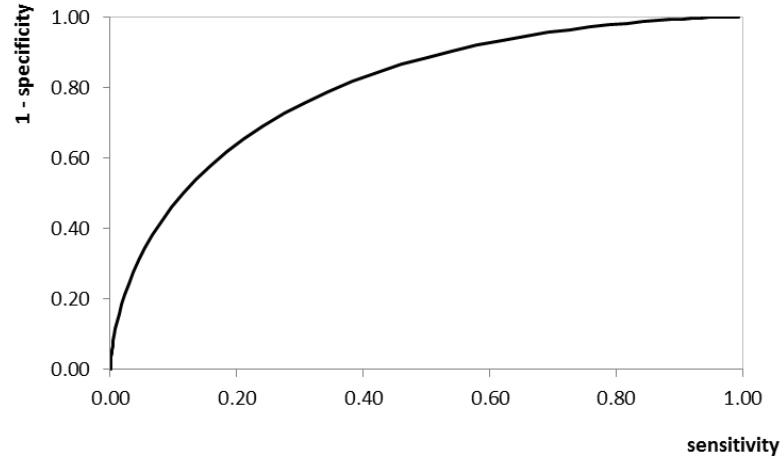
The first example considered is a test used to restrict the use of a single treatment. The test aims to identify two subgroups (healthy and diseased) but does so imperfectly as the distributions of test results in the two subgroups may overlap—for certain values of the test one cannot be sure whether the patient is diseased or healthy. A hypothetical example is shown in plot (a) in Figure 4. In this case, a single cut-off value may be used as classification rule but, in changing the cut-off, sensitivity and specificity will vary. The Receiver Operating Characteristic (ROC) curve(32) depicts the relationship between these two quantities as the cut-off changes—plot (b) in Figure 4, with sensitivity on the y-axis and 1-specificity on the x-axis. The ROC curve, however, cannot inform which point of the curve generates the highest level of (net)health for the population tested. It is more useful to present directly the (net)health attained by using alternative cut-off points as illustrated in plot (c) in Figure 4. Superimposed in this plot is the boundary of acceptance of the test, which shows that there is only value to testing if the cut-off is between 69 and 94 and that the maximum (net)health is attained at a cut-off of 85.

Figure 4 : Diagnostics: continuous results and two treatments available

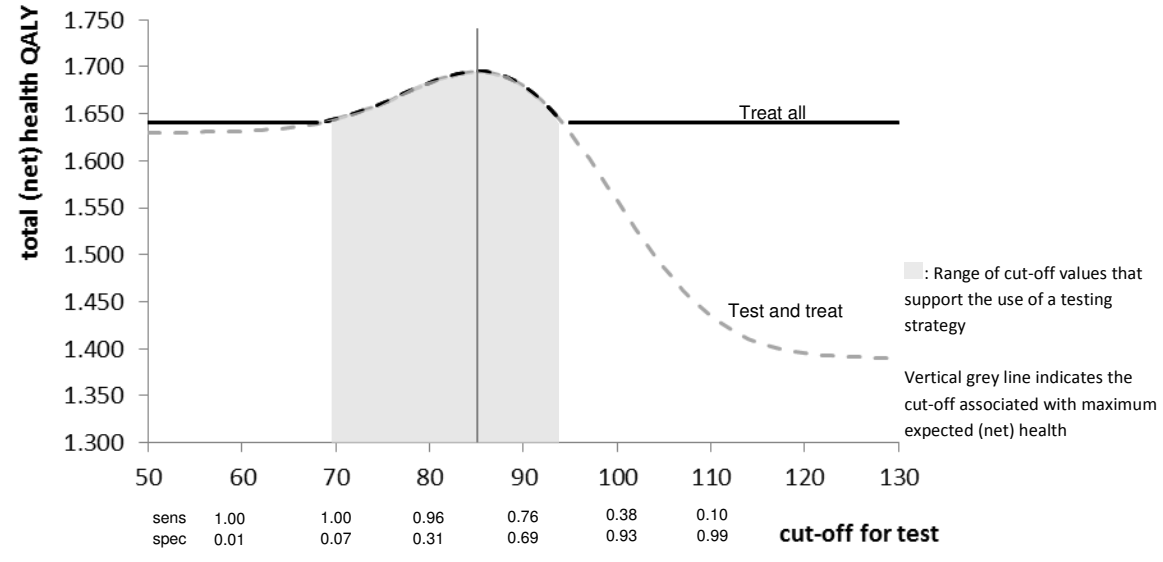
(a) Distribution of test results in the two populations, healthy and diseased



(b) ROC curve



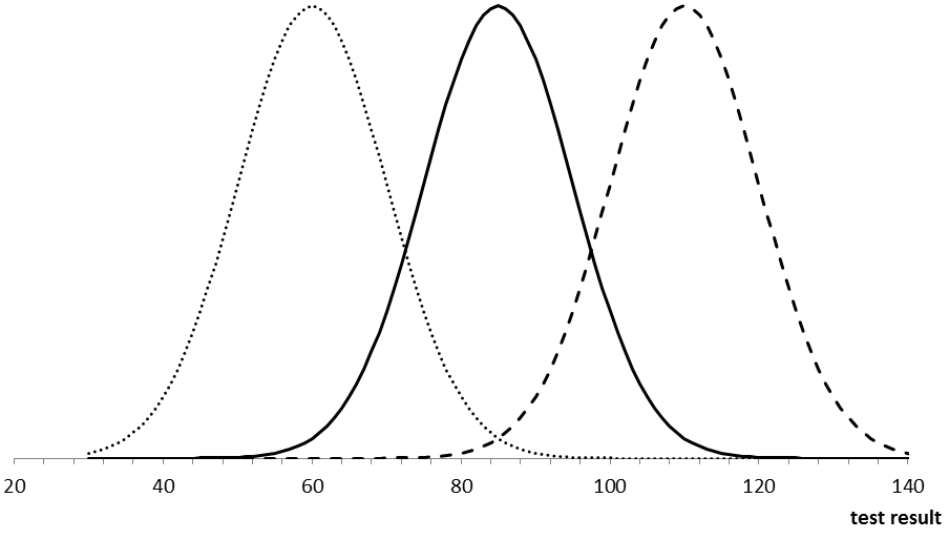
(c) total (net) health associated with relevant test/treat strategies -- acceptance boundary



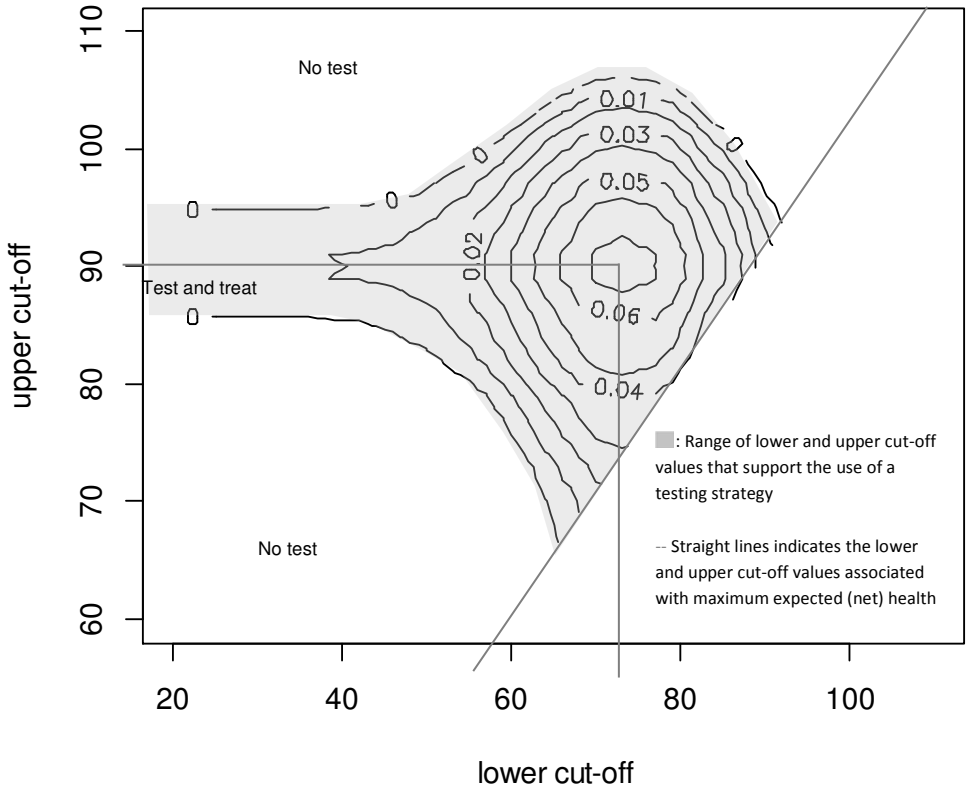
The second example extends the previous one to consider three subgroups, three treatment options (e.g. two available treatments and no treatment), and two cut-off values for the test results. Plot(a) in Figure 5 shows an example of the distribution of test results within each the three subgroups. Misclassification exists as the distributions of test results in each subgroup overlap but, given there are three subgroups, sensitivity and specificity cannot be used as these are only used for dichotomous tests. Instead, analogous measures can be used: probability of a patient in a given group j={1,…,3} being identified by the test as belonging to group k={1,…3}. For example, for those in group 1 (D=1), two quantities are needed to describe the accuracy of the test: P[T=1|D=1] (correctly identified) and P[T=2|D=1], with the remainder summing to 1 (i.e. P[T=3|D=1] = 1- P[T=2|D=1] - P[T=2|D=1]). Additionally, evidence is needed on the outcomes of all three treatment options when used in each of the three patient groups. With such evidence, one can identify combinations of the cut-offs that return the most health for the population. To display this information, plot (b) in Figure 5 uses contour lines, where each line represents combinations of cut-offs that lead to equal (net)health gains when compared to the next best alternative (which in this example is to treat none). These contour lines are only shown for the acceptance region for testing (shaded in the figure), i.e. where incremental (net)health as a result of testing is equal to, or above, 0. The plot shows that a lower cut-off of 73 and an upper cut-off of 90 lead to highest gains.

Figure 5: Diagnostics: continuous results and three treatments available

(a) Distribution of test results in the three populations



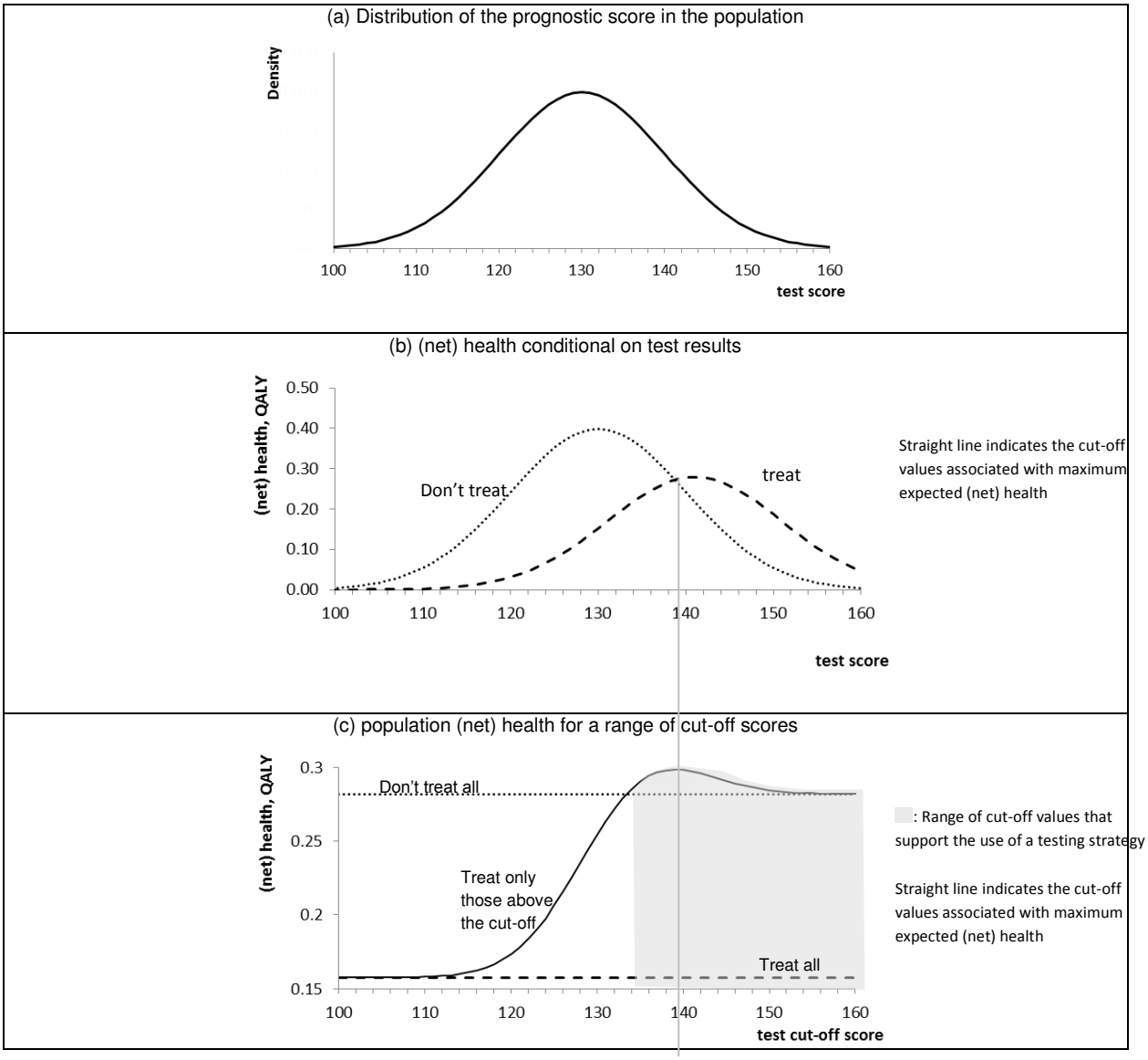
b) incremental total (net) health of testing in relation to next non-testing strategies -- contour plot with shaded acceptance region. Contour lines represent the combinations of cut-offs that lead to equal net health gain compared to the next best alternative. So, the 0.01 line represents the combinations associated with a net health gain of 0.01 QALYs.



Prognostics and diagnostics without reference test

In this subsection we illustrate how a hypothetical prognostic measure with continuous results (risk score, for example) can be evaluated within HTA. A possible distribution of test results in the population of interest is shown in plot(a) in Figure 6. Panel (b) of Figure 6 shows the (net)health of treating versus not treating patients with a given test result—this is no more than a marker-by-treatment predictiveness curve(34) or a treatment effect pattern plot(35), using (net)health as the outcome of interest. This plot can be used directly (i.e. independently of panel a) to define threshold scores above which treatment should be recommended—in this example the threshold should be set at 139 where the (net)health of treating becomes higher than that of not treating.

Figure 6: Prognostics: two treatments, 1 cut-off value.



However, a cut-off lower than 139 can still prove beneficial in relation to not treating. This is because the gains from treating those that benefit the most (at very high risk scores) still compensate for the losses imposed on those treated at a lower score than the optimal. Plot (c) in Figure 6 shows the average (net)health of the population when treating only those above a cut-off, and varies the cut-off values considered. The use of the prognostic test for clinical decision making is worthwhile for cut-off values above 134, despite the maximum (net)health still being generated at the cut-off of 139. Note that the distribution of test results in the population of interest (panel a) matters for this cut-off.

DISCUSSION

This paper lays out a coherent framework for the assessment of diagnostic and prognostic tests for HTA using a linked-evidence(5), or decision modelling, approach. It is solidly grounded on the indirect mechanism of value accrual for these health technologies that can be summarised using three interlinked components: classification (using test results to define treatment groups), choice (in terms of treatment) and outcomes. Importantly, this paper proposes a series of innovative graphical displays aiming to better inform decision making.

Implications for appraisal processes

The indirect and complex mechanisms of value for diagnostic and prognostic tests means that decisions over the 'upstream' tests that are used to restrict treatment to particular groups cannot be separated from decisions over 'downstream' treatments. It is, however, seldom the case that these decisions are considered simultaneously. Diagnostic decisions reached by NICE, for example, cannot issue recommendations on treatments. Also, the typical evaluation question 'which patients to treat' should be broadened to a question of 'when to treat' which is closer to clinical practice where treatment decisions are reviewed longitudinally over time.(31)

To inform decision making, all possible combinations of whether and when to test, test(s) to use, classification rules and treatment choices should be evaluated and compared. This may mean the evaluation becomes very complex, with many options to compare, but such complexity is necessary to identify the combination generating the most health for the population. It may prove to be analytically challenging, and therefore removing some of the options from the analyses may be desirable. A first consideration may be *relevance,* which can be established a priori based on specific characteristics of the tests or its purpose/role.(11) For example, if treatment cannot be undertaken without a test, e.g. tumor location in breast cancer surgery, a strategy of not testing and treating cannot be considered. However, there may still remain a multitude of options left for evaluation. With a value-based approach, options that do not retain significant possibility of being effective or cost-effective (given the existing uncertainties) can also be confidently excluded from the results. For example, Colbourn et al(36) in evaluating prenatal testing for group B streptococcal infection, alongside antibiotic treatment and vaccination, retained for analysis only combinations of interventions

that had more than a 1% probability of being cost effective, and discarded those remaining. Complex evaluations involving multiple strategies could also explore the use of operations research methods(37), such as optimisation where the number of combinations is large, or Markov decision processes for sequential testing [example in (38)].

A final critical implication for policy of the interlinked nature of the components of value is that a change in any element generates the need for re-appraisal. For example, if a new treatment emerges, or the price of one of the treatments changes, the value of upstream tests should be re-assessed, with classification rules and optimal treatment choices reviewed. Within a linked evidence approach re-appraisal is, in principle, simple as the evidence on components that are unchanged can still be used.

End-to-end studies

Clinical research on diagnostic tests has recently moved from exclusively focussing on accuracy evidence to recently highlighting the need for end-to-end, or outcome, studies(39)– clinical trials randomising patients to testing strategies (which may be a testing strategy and one not involving testing) to detect differences in health outcomes. Some policy makers in HTA have explicitly stated a preference for such end-to-end studies(3). These studies, however, embed pre-specified choices on classification and treatment, not allowing for the mechanisms of value to be made explicit and compromise on generalisability and adaptability to other settings. Importantly, where a new treatment emerges and re-appraisal is necessary, evidence from an original end-to-end study may become irrelevant. Additionally, given the aim of detecting overall outcome differences, the sample sizes for such studies may need to be unfeasibly high (especially for less prevalent diseases) with the associated costs potentially not worth the limited evidence they produce. Perhaps because of their costs, the availability of end-to-end trials is as yet very limited.(40)

Uncertainty

In the context of HTA, uncertainty refers specifically to uncertainty in knowledge that can be resolved through further research; uncertainty arises from sampled data, the need for extrapolation or the use of judgements. However, in the diagnostics literature, the term uncertainty has, confusingly, also been used to reflect imperfect performance of tests in identifying diseased patients. Accuracy cannot be improved with further research but only by developing a ‘better’, less imperfect, test (or by using a sequence of complementary tests).

It is important that uncertainty in the evidence base is explicitly considered, so that decision makers can be confident in making recommendations where a lower level of decision uncertainty presents. But, perhaps more importantly, a careful analysis of uncertainty will help determine whether further research is worthwhile and whether to condition the use of a particular technology only in the context of research.(39, 40) Given that regulatory processes for these technologies do not have as strict requirements for clinical evidence to be presented, it is especially important that any implications of

uncertainty are scrutinised. Uncertainty can be reflected through using both deterministic and probabilistic sensitivity analysis,(43) Careful consideration should be given to the analyses conducted with, for example, the use of bivariate distributions for considering uncertainty in test accuracy given the clear correlation between sensitivity and specificity.(44)

REFERENCES

1. World Health Organisation. Report of global survey on Health Technology Assessment. 2015.
2. National Institute for Health and Care Excellence. Diagnostic Assessment Programme [Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance.
3. Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. Assessing personalized medicines in Australia: a national framework for reviewing codependent technologies. Med Decis Making. 2013;33(3):333-42.
4. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ. 2013;346.
5. National Institute for Health and Care Excellence. Diagnostics Assessement Programme Manual. 2011.
6. Department of Health, Australian Government. Co-dependent and Hybrid Technologies, Health Technology Assessment Access Point 2011 [Available from: http://www.health.gov.au/internet/hta/publishing.nsf/content/co-1.
7. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6.
8. Meckley LM, Neumann PJ. Personalized medicine: factors influencing reimbursement. Health Policy. 2010;94(2):91-100.
9. 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:e686.
10. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making. 2009;29(5):E30-8.
11. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332(7549):1089-92.
12. Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Defining the clinical pathway in cochrane diagnostic test accuracy reviews. BMC Med Res Methodol. 2016;16(1):153.
13. Fugel HJ, Nuijten M, Postma M. Stratified medicine and reimbursement issues. Front Pharmacol. 2012;3:181.
14. Garrison LP, Jr., Austin MJ. Linking pharmacogenetics-based diagnostics and drugs for personalized medicine. Health Aff (Millwood). 2006;25(5):1281-90.
15. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Med Decis Making. 1988;8(4):279-89.
16. Longo R, Baxter P, Hall P, Hewison J, Afshar M, Hall G, et al. Methods for identifying the cost-effective case definition cut-off for sequential monitoring tests: an extension of Phelps and Mushlin. Pharmacoeconomics. 2014;32(4):327-34.
17. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Making. 1998;18(2 Suppl):S68-80.
18. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991;11(2):88-94.

19. AHRQ. Methods guide for medical test reviews. Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; 2010.
20. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care. 2013;29(3):343-50.
21. Medical Services Advisory Committee. Technical Guidelines for preparing assessment reports for the Medical Services Advisory Committee – Service Type: Investigative (Version 2.0) Department of Health, Australian Government; 2016.
22. Schaafsma JD, van der Graaf Y, Rinkel GJ, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. J Clin Epidemiol. 2009;62(12):1248-52.
23. Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nat Rev Drug Discov. 2007;6(4):287-93.
24. Grutters JP, Sculpher M, Briggs AH, Severens JL, Candel MJ, Stahl JE, et al. Acknowledging patient heterogeneity in economic evaluation : a systematic literature review. Pharmacoeconomics. 2013;31(2):111-23.
25. Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. Pharmacoeconomics. 2008;26(9):799-806.
26. Briggs AH, Claxton K, Sculpher MJ. Decision modelling for health economic evaluation. Oxford: Oxford University Press; 2006. x, 237 p. p.
27. Espinoza MA, Manca A, Claxton K, Sculpher MJ. The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. Med Decis Making. 2014;34(8):951-64.
28. van Gestel A, Grutters J, Schouten J, Webers C, Beckers H, Joore M, et al. The role of the expected value of individualized care in cost-effectiveness analyses and decision making. Value Health. 2012;15(1):13-21.
29. Hamburg MA, Collins FS. The path to personalized medicine. N Engl J Med. 2010;363(4):301-4.
30. Kalow W. Pharmacogenetics and personalised medicine. Fundam Clin Pharmacol. 2002;16(5):337-42.
31. Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? Health Economics. 2006;15(7):677-87.
32. Zhou X-h, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. 2nd ed. Hoboken, N.J.: Wiley; 2011. xxx, 545 p. p.
33. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008;54(1):17-23.
34. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med. 2011;154(4):253-9.
35. Lazar AA, Cole BF, Bonetti M, Gelber RD. Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: subpopulation treatment effect pattern plot. J Clin Oncol. 2010;28(29):4539-44.
36. Colbourn TE, Asseburg C, Bojke L, Philips Z, Welton NJ, Claxton K, et al. Preventive strategies for group B streptococcal and other bacterial infections in early infancy: cost effectiveness and value of information analyses. BMJ. 2007;335(7621):655.
37. Sainfort Fo, Brandeau ML, Pierskalla WP. Operations research and health care : a handbook of methods and applications. Boston, Mass.: Kluwer Academic; 2004. viii, 872 p. p.
38. Alagoz, O., Hsu, H., Schaefer, A. J., & Roberts, M. S. (2010). Markov decision processes: a tool for sequential decision making under uncertainty. Medical Decision Making, 30(4), 474-483.
39. Lijmer JG, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. Journal of Clinical Epidemiology. 2009;62(4):364-73.

40. 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(3):282-7.
41. Claxton K, Palmer S, Longworth L, Bojke L, Griffin S, McKenna C, et al. Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. Health Technol Assess. 2012;16(46):1-323.
42. Conti R, Veenstra DL, Armstrong K, Lesko LJ, Grosse SD. Personalized medicine and genomics: challenges and opportunities in assessing effectiveness, cost-effectiveness, and future research priorities. Med Decis Making. 2010;30(3):328-40.
43. Briggs A. Probabilistic analysis of cost-effectiveness models: statistical representation of parameter uncertainty. Value Health. 2005;8(1):1-2.
44. Shinkins B, Yang Y, Abel L, Fanshawe TR. Evidence synthesis to inform model-based cost-effectiveness evaluations of diagnostic tests: a methodological review of health technology assessments. BMC Med Res Methodol. 2017;17(1):56.