Published in final edited form as:

Pharmacoeconomics. 2014 April ; 32(4): 327–334. doi:10.1007/s40273-014-0134-1.

Methods for identifying the cost effective case definition cut-off for sequential monitoring tests: an extension of Phelps and Mushlin

Roberta Longo⁽¹⁾, Paul Baxter⁽²⁾, Peter Hall⁽¹⁾, Jenny Hewison⁽¹⁾, Mehran Afshar⁽³⁾, Geoff Hall⁽³⁾, and Christopher McCabe⁽⁴⁾

⁽¹⁾Leeds Institute of Health Sciences, University of Leeds, UK

⁽²⁾Centre for Epidemiology & Biostatistics, University of Leeds, UK

⁽³⁾Leeds Teaching Hospitals Trust, Leeds, UK

⁽⁴⁾Department of Emergency Medicine, University of Alberta, Canada

Abstract

The arrival of personalized medicine in the clinic means that treatment decisions will increasingly rely on test results. The challenge of limited health care resources means that the dissemination of these technologies will be dependent on their value in relation to their cost; i.e. their cost effectiveness. Phelps and Mushlin have described how to optimize tests to meet cost effectiveness target. However, when tests are applied repeatedly the case mix of the patients tested changes with each administration, and this impacts upon the value of each subsequent test administration. In this paper we present a modification of Phelps and Mushlin's framework for diagnostic tests; to identify the cost effective cut-off for monitoring tests. Using the Ca125 test monitoring for relapse in Ovarian Cancer, we show how the repeated use of the initial cut-off can lead to a substantially increased false negative rate compared to the monitoring cut-off – over 20% higher than in this example – with the associated harms for individual and population health.

Section 1: Introduction

The advent of personalised medicine and whole genome screening seek to place testing at the heart of everyday clinical care. An inevitable result of this potential revolution in health

Corresponding Author: Christopher McCabe PhD, Capital Health Endowed Research Chair, Department of Emergency Medicine, University of Alberta, Suite 736 University Terrace, 8303 112 Street, Edmonton, Alberta, Canada T6G 0B7, mccabe1@ualberta.ca, Tel: +1 780 492 4202.

Conflicts of Interest

Roberta Longo has no conflicts of interest

Paul Baxter has no conflicts of interest

Peter Hall has no conflicts of interest

Jenny Hewison has no conflicts of interest

Mehran Afshar has no conflicts of interest

Geoff Hall has no conflicts of interest

Christopher McCabe has no conflicts of interest

Disclaimer

The views and opinions expressed by the authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the Department of Health or Genome Canada.

care is that reimbursement authorities will have to assess the cost effectiveness of tests and test-directed therapies with increasing frequency. Historically, tests have represented a small component of health technology assessment activity and they have not been the focus of methods development to the same degree as many other technologies. However, reimbursement authorities are increasingly interested in how to appraise these technologies and in some cases have established bespoke processes for just this purpose. [1,2]. Establishing how to incorporate these new technologies into the health technology appraisal and value based pricing frameworks is an increasingly important challenge.

The seminal work on the assessment of the cost effectiveness of tests was published by Phelps and Mushlin in 1988.[3] In this they described how to identify the diagnostic test cut-off that would meet a cost effectiveness target. However, the personalized medicine revolution described by its advocates [4] conceives of the repeated application of these tests to monitor an individual's health, and in people with a diagnosis, the progression of disease. There are already some personalized medicine tests available that are designed for monitoring, including PSA for prostate cancer, Ca125 for Ovarian Cancer and the Enhanced Liver Fibrosis. [5,6,7] When a test is used for monitoring, the results of the first administration of the test and responses to it, change the case mix patients at the second administration of the test and this may impact upon the test performance characteristics and hence the value of the test using the initial test cut-off value.

The objective of this paper is to present an extension of framework developed by Phelps and Mushlin (PM)[3] for assessing diagnostic tests, for monitoring tests. Specifically we aim to describe the factors that may drive a difference in the cost effective cut-off between the first and subsequent applications of the same test; present a formal framework for identifying the cost effective cut-off for sequential monitoring tests and present a worked example of the extended framework using Ca125 monitoring test data in the context of Ovarian Cancer.

The remainder of the paper is structured as follows: Section 2 briefly reviews PM's [3] model for assessing diagnostic technologies. Section 3 describes the model that we used to impose the conditions of differentiability and convexity to the empirically derived Receiver Operator Curve (RoC). Section 4 describes how the monitoring test context is systematically different from the diagnostic test one and presents an extension of the PM model to sequential monitoring tests; which is then applied to the use of the tumour marker Ca125 to monitor disease activity in Ovarian Cancer [6] in Section 5. Section 6 discusses the implications of the results of the case study analysis and presents a new output for test cost effectiveness analyses – the Cost Effectiveness Threshold Curve. Section 7 summarises the arguments for and implications of adopting cost effectiveness criteria in developing new test technologies, such as will be required for the realization of the personalised medicine revolution.

Section 2: Identification of the cost effective case definition cut-off for diagnostic test

PM [3] developed a theoretical framework for determining the expected value of diagnostic information with a two-step approach. In the first step ("hurdle 1") perfect sensitivity and

specificity of the diagnostic technology are assumed. It is required that the expected value of perfect information is higher than the cost of using the new technology for it to access the second step (“hurdle 2”) of evaluation, whereby the test has less than perfect sensitivity and specificity. Likewise, Hurdle 2 would be overcome if the expected value of the imperfect information exceeds the cost of using the diagnostic technology.

The decision theoretic model of PM can also be used to identify the case definition cut-off to meet a pre-specified cost effectiveness threshold. In this section we briefly review PM’s framework.

Consider a diagnostic test that dichotomises the subsequent management strategy dependent upon those results. The actual patient benefits and costs will depend upon true health states and how well the test can identify these (e.g. in terms of characteristics of the RoC, such as area under the curve [8]). Ideally the information from the diagnostic test will be interpreted to best benefit the patient and to be cost efficient, by identifying the case definition cut-off corresponding to the cost-effective combination of sensitivity and specificity at the relevant threshold for cost effectiveness. In PM, this can be done in a general optimization setting, using the same probability, utility and cost data required for calculating the value of diagnostic information and for which the authors define the following notation:

f_d	probability patient presenting for test D is sick (i.e. prevalence of disease)
$1 - f_d$	probability patient presenting for test D is healthy
U_{st}	utility of sick person, treated
U_{sn}	utility of sick person, not treated
U_{ht}	utility of healthy person, treated
U_{hn}	utility of healthy person, not treated
C_{st}	cost of sick person, treated
C_{sn}	cost of sick person, not treated
C_{ht}	cost of healthy person, treated
C_{hn}	cost of healthy person, not treated

It is assumed that $U_{st} > U_{sn}$ (the treatment is efficacious) and $U_{ht} < U_{hn}$ (due to side effects of therapy). We further define cost and utility differentials:

$U_s = U_{st} - U_{sn}$	utility gained by treating a sick person
$U_h = U_{hn} - U_{ht}$	utility gained by not treating a healthy person
$C_s = C_{st} - C_{sn}$	change in cost by treating a sick person
$C_h = C_{hn} - C_{ht}$	change in cost (< 0) from not treating a healthy person

Then we define the cost effectiveness threshold that we wish to optimize the test for. Consider the health gains achievable if the resources needed for the proposed test were put to use in the least favourable procedure that was still acceptable, i.e. the first activity that would be dropped to provide others in a resource constrained environment [9]. We denote the effectiveness to cost ratio (in QALYS/£) of this marginal activity as g .

The sensitivity of the test can be increased at the cost of reduced specificity, or vice versa, by varying the cut-off, which we denote as k_d , used to establish the diagnosis.

In general, we wish to maximise the net benefit of the test (NB_d) for the patient with respect to p_d and q_d ; where p_d is sensitivity and q_d is $1 - \text{specificity}$.

From Eqn. (A11) in PM [3] it can be shown that:

$$\begin{aligned}
 NB_d = & f_d [p_d U_{st} + (1 - p_d) U_{sn}] \\
 & + (1 - f_d) [(1 - q_d) U_{hn} + q_d U_{ht}] \\
 & - g \{f_d [p_d C_{st} + (1 - p_d) C_{sn}] \\
 & + (1 - f_d) [(1 - q_d) C_{hn} + q_d C_{ht}]\} \quad (1)
 \end{aligned}$$

We can find combinations of p_d and q_d in which the net benefit remains the same, by taking the differential of Eqn. (1), allowing p_d and q_d to vary jointly, but holding the total change in NB_d ($d NB_d$) to zero.

The optimal choice, i.e. that which maximizes the expected net benefit of the test with respect to $\hat{p}_d$ and $\hat{q}_d$, is then found (see [3]) by setting:

$$\frac{dp_d}{dq_d} = \frac{(1 - f_d) [\Delta U_h - g \Delta C_h]}{f_d [\Delta U_s - g \Delta C_s]} = \beta_d \quad (2)$$

In Eqn. (2) we are selecting the cut-off k_d that identifies a point on the RoC curve with slope β_d as shown in Fig. 1 below.

Thus, given a specified cost effectiveness threshold, such as that used by the United Kingdom National Institute for Health and Care Excellence [10], knowledge of the costs and outcomes of treatments for true and false negatives and positives, as well as the RoC for a diagnostic test, it is possible to identify the case definition cut-off that will lead to the test being a cost effective addition to the portfolio of services provided.

Section 3: Fitting a Receiver Operator Curve (RoC) model

In order to identify a unique point on a RoC from its slope, it is necessary for the curve to be both smooth (differentiable) and convex. Moreover, it can be shown that non-convexity in an ROC curves implies the use of an irrational decision strategy [11]. Empirically derived RoC do not necessarily satisfy these conditions and hence we impose them by fitting a model to the empirical curve. The model is then used to derive the slope for the optimal sensitivity, specificity and cut-off as identified by PM methods (suitably modified when applied in a monitoring context).

The most common RoC model is the binormal [12]. This model assumes biomarker values are normally distributed in both the population of patients that are in remission and the population of patients whose cancer has recurred, but allows the means and standard deviations in the two populations to differ. The binormal model has the advantage that its slope for a particular sensitivity, specificity pairing can be derived analytically [13]. However, in many applications the normality assumptions are not satisfied [14] and the

model curve, although differentiable, is not convex (and hence a slope value does not identify a unique point on the curve).

As a refinement to the binormal model, Metz et al [15] proposed a more flexible, latent variable approach whereby it is assumed that some monotonic transformation of the biomarker values yield normal distributions. Maximum likelihood fitting of this latent variable binormal model is implemented in the ROCKIT software provided by the Metz lab (<http://metz-roc.uchicago.edu/>). Although the latent variable model does not impose convexity on the fitted curve, in many applications where large samples of data are available for continuous biomarkers, the convexity property holds [16]. The ROCKIT software can be used to identify both the sensitivity, specificity pairings that form the model curve and the underlying biomarker cut-off values for a particular pairing used in predicting remission/recurrence status. Although no analytical method is available to determine the slope of latent variable RoC, numerical derivatives are easily obtained (e.g. by finite difference see, for example, Burden and Faires [17]).

As a further extension of the binormal model Pesce & Metz [18] describe a “proper” binormal model, whereby convexity is imposed on the curve. This is of particular importance when models are fitted to small samples or where biomarkers are ordinal rather than continuous. Current software to fit the “proper” binormal model provided by the Metz lab can be used to identify sensitivity and specificity pairings that form the model curve for use, e.g. in identifying performance measures such as the area under the RoC (AUC). However, the software does not yield the underlying biomarker cut-off values for a particular pairing to predict remission/recurrence status. Therefore it cannot be used to identify optimal cut-offs for the PM method.

Given that, for the CA125 data example (described in section 5), we have a continuous biomarker and a large sample size we adopt the latent variable model of Metz et al [15]. Convexity and model fit has not proved problematic in this application, although we acknowledge that in other data sets further work on software to identify cut-off values for the “proper” binormal model may be required.

Section 4: Sequential monitoring tests

Developing on the framework described in Section 2, consider the utilization of the same test for monitoring. There are a number of circumstances in which a test may be used for monitoring purposes. The first is when the result from the first administration of the test, is negative. In patients who have a clear predisposition for the condition of interest, periodic re-application of the same test is an obvious management strategy. Clinicians may also choose to re-administer the test in patients who test positive on the first administration. If the natural history of the disease allows for remission or if treatment is expected to achieve remission then monitoring is also a sensible strategy. Tests may also be re-administered when there are concerns about the test-re-test reliability. However, this is not a monitoring activity as it is concerned with the performance characteristics of the test rather than capturing changes in the patient’s clinical condition. An illustration of the process is shown in Figure 2 below, If the result of D_1 , which is the monitoring test at time 1, is positive, a

patient receives monitoring test D_{2+ve} at time 2.. If the result of D_1 is negative, a patient receives monitoring test D_{2-ve} . Note that the subsequent tests are not different from the initial test, they are just administered to different subgroups in a subsequent point in time. The key observation is that the patients presenting for monitoring tests D_{2+ve} and D_{2-ve} have a different probability of being sick (prevalence) compared to the general population of patients presenting for the initial test; whilst the costs, utilities and health gains of treatment remain the same.

The probability that a patient presenting for tests D_{2+ve} and D_{2-ve} is sick depends on the positive predictive value and negative predictive value of test D_1 , with an adjustment term to allow for disease progression. To address this analytically we need to define some additional terms. Define:

$$f_{2+ve} = r_1 + (1 - r_1) \rho \quad (3)$$

$$f_{2-ve} = (1 - s_1) + s_1 \rho \quad (4)$$

Where f_{2+ve} and f_{2-ve} represent the prevalence of disease of patients presenting for subsequent tests; ρ is a direct function of time and is the probability a patient is sick when presenting for test D_{2+ve} or D_{2-ve} given they were not sick when presenting for test D_1 (i.e. the probability that a healthy person presenting for test D_1 becomes sick due to disease progression before presenting for test D_{2+ve} or D_{2-ve}); r_1 and s_1 are respectively the positive predicted value and the negative predicted value of test D_1 . Thus Eqn. (3) represents the probability that a person presenting for test D_{2+ve} will be sick and Eqn. (4) represents the probability a person presenting for test D_{2-ve} will be sick.

By an analogous argument to that presented in Section 2, it follows that optimal cut-offs, k_{2+ve} and k_{2-ve} can be identified by choosing points on the ROC curve with slopes β_{2+ve} and β_{2-ve} , where:

$$\beta_{2+ve} = \frac{(1 - f_{2+ve}) [\Delta U_h - g \Delta C_h]}{f_{2+ve} [\Delta U_s - g \Delta C_s]} \quad (5)$$

$$\beta_{2-ve} = \frac{(1 - f_{2-ve}) [\Delta U_h - g \Delta C_h]}{f_{2-ve} [\Delta U_s - g \Delta C_s]} \quad (6)$$

If the arguments presented above are accepted, it is clear that cut-off values k , for further monitoring tests proceeding D_{2+ve} and D_{2-ve} can be identified. To do so, the proportion of sick patients presenting for further investigation need only be updated to reflect the *positive predictive value* and *negative predictive value* of tests D_{2+ve} and D_{2-ve} etc., with appropriate modification for further disease progression up to the point in time when the subsequent tests would be performed. In fact, being ρ a direct function of time, if the test is performed at an earlier or later point in time, the estimate of true negatives who are expected to become true positives in the interval between test administrations would change as well.

Furthermore, PM identify the circumstances under which a test creates net value (see Eqns. (4-6). in [3]). If the clinician's default choice prior to testing is to treat, then the test should be employed if:

$$-(1 - p_{\bullet}) f_{\bullet} [\Delta U_s - g \Delta C_s] + (1 - q_{\bullet}) (1 - f_{\bullet}) [\Delta U_h - g \Delta C_h] > C_t \quad (7)$$

Similarly, if the clinician's default choice prior to testing is not to treat, then the test should be employed if:

$$f_{\bullet} p_{\bullet} [\Delta U_s - g \Delta C_s] - (1 - f_{\bullet}) q_{\bullet} [\Delta U_h - g \Delta C_h] > g C_t \quad (8)$$

We suggest that Eqns. (7-8) can be used to identify the point in a testing sequence when further tests are not cost effective. We also argue that (7-8) offer a framework which can account for quality of life and costs accrued over an extended time horizon.

Section 5: Identifying the Optimum Monitoring Test Cut-offs for the Ca125 test in the management of Ovarian Cancer

For the purposes of this example we constructed a stylized cost effectiveness model of monitoring ovarian cancer patients with Ca125, to identify relapse and trigger chemotherapy. The model uses a 12 month time horizon as there was no difference in mortality between Ca125 guided and symptom guided therapy observed in the MRC OV05 trial [6]. The outcome used in the model was the Quality Adjusted Life Year with utility weights for sick patients, both treated and untreated taken from the NICE Clinical Guideline for Ovarian Cancer [19]; and utility weights for healthy patients taken from Kind et al [20]. The costs of care are also taken from the NICE Clinical Guideline [19]. Model parameters are summarized in Table 1.

5.1 Ca125 Test Data

The anonymized clinical data used for this analysis was extracted from the central clinical database used by oncology services in Leeds Teaching Hospitals NHS Trust (LTHT). Six hundred consecutive patients with a diagnosis of ovarian cancer after 1st January 2004 were included in the analysis. Patients of all stages at diagnosis were included, irrespective of initial or subsequent treatments. The most common management course consists of initial primary treatment with surgery or chemotherapy or both treatments. The majority of patients' cancers will respond either completely or partially to this initial treatment, following which they enter a period of monitoring; monitoring consists of clinical assessment, Ca125 measurement (kiloUnits per liter in our clinical data) and Computer Tomography (CT) scanning.

After entering a monitoring period, further treatment with chemotherapy is usually initiated on the basis of a clinical decision that there is sufficient evidence of cancer progression based on CT and clinical assessment. A rising Ca125 will typically prompt a CT scan, but will only rarely, in isolation, result in treatment without confirmatory CT or clinical evidence of progression. Progression status is accurately coded in this clinical dataset. We

refer to our clinical outcome as ‘disease progression’ or ‘disease control’. The term disease progression is used for consistency where cancer has grown, recurred or relapsed. Disease control is used as a summary term representing tumour shrinkage, stability, remission or possible cure. A true positive Ca125 test is defined as a Ca125 measurement followed in less than 30 days by a CT scan showing progression. A true negative Ca125 measurement is defined as a marker measurement that is not followed within 30 days or preceded within 30 days by a diagnosis of progression (from CT scan or commencement of treatment).

5.2 Cost effective Case Definition Cut-off

To calculate the cost effective cut off for the test at its first administration we used Equation 2 above (from PM) and solved using the values in Table 2.

The prevalence of progressive ovarian cancer in this dataset was 33% at the time of the first administration of the test, using the definition described above. The utility gain from not treating a healthy person is $U_h = 0.1$, i.e. a person not in need of treatment benefits from not receiving chemotherapy, which would cause them extensive side effects. On the other hand, the utility gain from treating a sick person is $U_s = 0.18$, i.e., the treatment is effective.

With regard to costs, given that everyone in the sample is administered the diagnostic test and given the simplified framework described above, the change in cost from treating a sick person shows just the chemotherapy cost, $C_s = £316.29$; whilst the change in cost from not treating a healthy one is the chemotherapy cost avoided, $C_h = -£316.29$.

Finally, $g = 0.00005$ is the inverse of NICE’s willingness to pay threshold, i.e. the minimum QALY gain the NHS is willing to accept when spending £20,000 on an intervention.

The optimum slope identified with these parameter values is 1.44, which corresponds to a case definition cut-off of 127.94 kU/L (Table 2); i.e. this is the cut-off that delivers improvement in patient outcomes with an Incremental Cost Effectiveness Ratio (ICER) that is equal to the Cost Effectiveness Threshold. With this case definition cut-off, Sensitivity = 67% and Specificity = 79%. This means that, with an underlying prevalence of 33% at time 1, 22% of ill patients will be correctly identified as progressed; whilst the remaining 11% will be incorrectly identified as under control. As the underlying prevalence of the condition increases, the optimal test cut-off decreases. Likewise, as the cost of missed diagnoses increases the optimal test cut-off decreases.

For ease of exposition of the issues around test optimization for monitoring, we narrow the discussion of sequential testing to a subset, i.e. the re-administration of a test following an initial negative test result, the ‘control disease’ group in our example. In order to identify the cost effectiveness cut-off for the subsequent application of the test, we need to specify the prevalence of the disease in the group that the initial test defined as having controlled disease (testing negative on D_1). At a cut-off of 127.94 kU/L, the negative predictive value of the test is 0.611, and the positive predictive value is 0.831. When this is combined with a disease progression rate of 20% - i.e. with 20% of those initially under control experiencing recurrence between surveillance timepoints - the prevalence of disease in the controlled

group is 50%. The new cut-off identified is significantly lower than at the first administration, 42.53 kU/L, with Sensitivity = 86% and Specificity = 61%. With this case definition cut-off 43% of ill patients will be correctly identified as progressed; whilst the remaining 7% will be incorrectly identified as under control.

In the controlled disease group, only a very low test result is required to confirm the presence of progressive disease in a cost effective manner.

Section 6: Using cost effective test cut-offs

The choice of g , the cost effectiveness threshold used in the optimization analysis is a key decision. Conceptually, it is likely that the cost effectiveness threshold will differ between health systems because the threshold is a function of the health system's budget, the portfolio of the current therapies it provides and the demand for those therapies from the population it serves [9]. Given the complexity of the determinants of the cost effectiveness threshold it is perhaps unsurprising that there is substantial disagreement about the appropriate choice of threshold value even within a health system [21; 22; 23]. As a result decision makers are likely to be interested in the impact of changing the cost effectiveness threshold on the optimal case definition cut-off value.

Figure 3 illustrates the Cost Effectiveness Threshold Curve for the Ca125 test in Ovarian Cancer. The optimum Case Definition Cut-off is plotted on the vertical axis and the cost effectiveness threshold, sometimes called the Willingness to Pay for Health, is plotted on the horizontal axis. We can see that in this case, as the willingness to pay for health increases from £5,000 to £10,000 both the initial and sequential cut-off for patients assessed with controlled disease drop substantially, but any further increases have only a moderate impact.

For the exemplar test, most health systems are likely to use a willingness to pay value that is in the less sensitive range. However, it is credible that for other tests different health systems might operate very different test cut-off scores, reflecting the large differences in their willingness to pay. This ability to identify a cut-off value at which a test is cost effective may be attractive to budget holders who wish to avoid simple accept or reject decisions for personalized medicine technologies, by allowing them to utilize companion diagnostics to identify the sub-group of patients in whom treatment is cost effective.

Section 7: Discussion

In this paper we have shown how the model of PM for identifying the cost effective case definition cut-off for a diagnostic test can be extended to monitoring tests; and how the test cut-off that is cost effective in the investigation at time 2 may be quite different from the one that is cost effective at time 1. The notion of prevalence varying under repeat testing necessitating a difference in cut-offs is discussed (though not developed) in the recent paper [24]. We show that the extent of this difference, and hence its importance to clinicians, patients and decision makers depends upon the underlying prevalence of the condition, test performance characteristics of the test at its first administration, rate of disease progression and the interval between test administrations, as well as the target cost effectiveness threshold. In our example that difference in cut-off levels means that in the application at

time 2 the test returns a higher true positive rate and a lower false positive rate than the test at time 1. In fact we have a higher proportion of ill patients correctly identified as such; 43% identified using the subsequent cut-off, compared 22% using the initial cut-off. We also observe a lower proportion of ill patients incorrectly identified as under control; 7% using the subsequent cut-off compared to 11% using the initial cut-off. Specifically, the repeated use of the initial cut-off leads to increased false negative rate – 4% higher in this example – with the associated harms for individual and population health.

The results presented in this paper consider monitoring in patients whose initial test result is negative. Monitoring after a positive test result can occur when remission is part of the natural history of the disease or when treatment is aimed at achieving remission. In principal the framework described in Section 4 could be used in these circumstances by incorporating a probability of disease remission. The impact of remission on costs and outcomes of care are unlikely to be a simple reversal of the costs and outcome impact of progression. Including these forms of monitoring in the exemplar analysis would have increased the complexity without adding substantially to its value.

As health care systems increasingly turn to cost effectiveness based decision criteria, including value based pricing, to determine which innovative technologies are reimbursed, developers of these test dependent technologies will need to consider the relationship between their choice of cut-off and the number of markets that are likely to view their technology as a good use of limited health care resources, and thus the likely sales revenue. Alternatively, developers may wish to simply provide health care systems with the Receiver Operating Curve characteristics for their technologies and allow each system to identify its own case definition cut-off, or cut-offs in the case of monitoring tests, at which the test is good value.

Conclusions

The potential tsunami of diagnostic tests that the proposed personalized medicine revolution is bringing forth, will drive health care budget holders away from accepting conventional individual patient level clinical utility as a sufficient justification for the choice of case definition. Net health benefit at the population level will be an important part of the justification of any particular cut-off. The framework developed by PM, and extensions, such as the ones we have described here, offer a formal decision analytic approach identifying the cut-off that optimizes the contribution of these technologies to population health in a given health care system.

Acknowledgements

This paper presents independent research partially funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (RP-PG-0707-10101: Evaluating the benefits for patients and the NHS of new and existing biological fluid biomarkers in liver and renal disease). Additional support was received through the Medical Technologies Innovation and Knowledge Centre funded by the UK Research Councils, EPSRC and BBSRC, and the Technology Strategy Board. We wish to acknowledge helpful discussions with colleagues on the NIHR Programme Grant award, especially Doug Altman, Jon Deeks, Walter Gregory and Peter Selby. Prof. McCabe's research is made possible by the Capital Health Research Chair Endowment to the University of Alberta Faculty of Medicine and Dentistry. He also received funding from the Genome Canada Large

Scale Applied Research Program on Personalized Medicine for some of this work. The work reported in this paper and any errors remain the responsibility of the authors.

All individuals who made substantive contributions to the ideas presented in this article are included as authors. The initial concepts for this article were proposed by Christopher McCabe. Roberta Longo and Paul Baxter undertook the modelling work to provide the example and prepared the draft for the manuscript. Peter Hall, Geoff Hall and Mehran Afshar provided the clinical data and contributed to the manuscript writing. Christopher McCabe and Jenny Hewison supervised the preparation of the draft and contributed on the manuscript writing.

References

- [1]. National Institute for Health and Clinical Excellence. [accessed 29th July 2013] Diagnostics Assessment Programme. <http://www.nice.org.uk/diagnostics>
- [2]. Department of Health and Aging. [accessed 29th July 2013] Co-dependent and Hybrid Technologies. <http://www.health.gov.au/internet/hta/publishing.nsf/Content/co-1>
- [3]. Phelps CE, Mushlin A. Focussing Technology Assessment using Decision Theory. Medical Decision Making. 1988; 8(4):279–289. [PubMed: 3185180]
- [4]. Guttmacher AE, Collins FS. Genomic Medicine – a primer. NEJM. 2002; 347(19):1512–1520. [PubMed: 12421895]
- [5]. Carroll PR, Whitson JM, Cooperberg MR. Serum Prostate-Specific Antigen for the Early Detection of Prostate Cancer: Always, Never, or Only Sometimes? Journal of Clinical Oncology. 2011; 28(4):345–346. [PubMed: 21189396]
- [6]. Rustin GJ, van der Burg ME, Griffin CL, Guthrie D, Lamont A, Jayson GC, Kristensen G, Mediola C, Coens C, Qian W, Parmar MK, Swart AM, MRC OV05; EORTC 55955 investigators. Early versus delayed treatment of relapsed ovarian cancer (MRC OV05/EORTC 55955): a randomised trial. Lancet. 2010; 376(9747):1155–63. [PubMed: 20888993]
- [7]. Parkes J, Guha IN, Roderick P, Harris S, Cross R, Manos MM, Irving W, Zaitoun A, Wheatley M, Ryder S, Rosenberg W. Enhanced Liver Fibrosis (ELF) test accurately identifies liver fibrosis in patients with chronic hepatitis C. Journal of Viral Hepatitis. Jan; 2011 18(1):23–31. doi: 10.1111/j.1365-2893.2009.01263.x. [PubMed: 20196799]
- [8]. Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. 2nd ed. John Wiley & Sons, Inc.; 2000. p. 156-164.
- [9]. McCabe C, Claxton K, Culyer AJ. The NICE Cost Effectiveness threshold: what it is and what that means? Pharmacoeconomics. 2008; 26(9):733–744. [PubMed: 18767894]
- [10]. NICE. Guide to the Methods of Technology Appraisal. London: 2008.
- [11]. Pesce LL, Metz CE, Berbaum KS. On the convexity of ROC curves estimated from radiological test results. Academic Radiology. 2010; 17(8):960–968. [PubMed: 20599155]
- [12]. Grey DR, Morgan BJT. Some aspects of ROC curve-fitting: Normal and logistic models. Journal of Mathematical Psychology. 1972; 9:128–139.
- [13]. Gu W, Pepe MS. Estimating the diagnostic likelihood ratio of a continuous marker. Biostatistics. Jan; 2011 12(1):87–101. [PubMed: 20639522]
- [14]. Hanley JS. The robustness of the binormal assumptions used in fitting ROC curves. Medical Decision Making. 1988; 8:197–203. [PubMed: 3398748]
- [15]. Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. Statistics in Medicine. 1998; 17:1033–1053. [PubMed: 9612889]
- [16]. Metz CE, Pan X. Proper binormal ROC curves: theory and maximum-likelihood estimation. Journal of Mathematical Psychology. 1999; 43:1–33. [PubMed: 10069933]
- [17]. Burden, RL.; Faires, JD. Numerical Analysis. 7th Ed. Brooks/Cole; 2000.
- [18]. Pesce LL, Metz CE. Reliable and computationally efficient maximum-likelihood estimation of “proper” binormal ROC curves. Academic Radiology. 2007; 14:814–829. [PubMed: 17574132]
- [19]. National Institute for Health and Clinical Excellence. NICE Full Guideline. London: 2011. Ovarian cancer: the recognition and initial management of ovarian cancer.
- [20]. Kind P, Dolan P, Gudex C, Williams A. Variations in population health status: results from a United Kingdom national questionnaire survey. British Medical Journal. 1998; 316:736–41. [PubMed: 9529408]

- [21]. Mason H, Jones-Lee M, Donaldson C. Modelling the monetary value of a qaly: a new approach based on uk data. *Health Economics*. 2009; 18(8):933–950. [PubMed: 18855880]
- [22]. Towse A. Should NICE's threshold range for cost per QALY be raised? Yes. *British Medical Journal*. 2009; 338
- [23]. Claxton, K.; Martin, P.; Soares, M., et al. Methods for the estimation of the NICE Cost effectiveness threshold. Draft Final Report October 2012. <http://www.york.ac.uk/media/che/documents/reports/Methods%20for%20the%20Estimation%20of%20the%20NICE%20Cost%20Effectiveness%20Threshold%20%28Draft%20Final%20Report%29%20%281%29.pdf>
- [24]. 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(4):536–41. [PubMed: 23796287]

Key Points

- By applying Phelps and Mushlin's framework to monitoring tests, we have shown that test cut-off that is cost effective in monitoring may be quite different from the one that is cost effective in the first administration of the test.
- We show that the extent of this difference, and hence its importance to clinicians, patients and decision makers depends upon the underlying prevalence of the condition, test performance characteristics of the test at its first administration, rate of disease progression and the interval between test administrations, as well as the target cost effectiveness threshold.
- Decision makers are likely to be interested in the impact of changing the cost effectiveness threshold on the optimal case definition cut-off value; hence we present the Cost Effectiveness Threshold Curve which is able to identify the optimal case definition cut-off value corresponding to each cost-effectiveness threshold.
- The framework developed by Phelps and Mushlin, and extensions, such as the ones we have described here, offer a formal decision analytic approach identifying the cut-off that optimizes the contribution of these technologies to population health in a given health care system.

Figure 1. Identifying the point k_d on the ROC curve

Figure 2. Monitoring tests subsequent to an initial test D₁

Figure 3. Cost Effectiveness Threshold Curve

Table 1
Parameters values used to identify the cost effective threshold at the first administration

Parameters	Values	Source
f_{dt}	0.328	From the clinical data
$g = 1/£20,000$	0.00005	–
U_{st}	0.81	[19]
U_{sn}	0.63	[19]
U_{ht}	0.7	Assumption
U_{hn}	0.8	[20]
$U_s = U_{st} - U_{sn}$	0.18	–
$U_h = U_{hn} - U_{ht}$	0.1	–
C_{st}	£ 339.29	[19]
C_{sn}	£ 23.00	[19]
C_{ht}	£ 339.29	[19]
C_{hn}	£ 23.00	[19]
$C_s = C_{st} - C_{sn}$	£ 316.29	–
$C_h = C_{hn} - C_{ht}$	–£ 316.29	–

Table 2
Test performance characteristics

	Initial Test D_I	Subsequent Test D_{2-ve}
Optimum slope	1.44	0.675
Cut-off value	127.94 kU/L	42.53 kU/L
False positive (1 – specificity)	0.208	0.393
True positive (sensitivity)	0.670	0.856
PPV	0.831	0.896
NPV	0.611	0.515