Decision Curve Analysis: A Novel Method for Evaluating Prediction Models

Andrew J. Vickers, PhD, Elena B. Elkin, PhD

Background. *Diagnostic and prognostic models are typically evaluated with measures of accuracy that do not address clinical consequences. Decision-analytic techniques allow assessment of clinical outcomes but often require collection of additional information and may be cumbersome to apply to models that yield a continuous result. The authors sought a method for evaluating and comparing prediction models that incorporates clinical consequences, requires only the data set on which the models are tested, and can be applied to models that have either continuous or dichotomous results.* ***Method***. *The authors describe decision curve analysis, a simple, novel method of evaluating predictive models. They start by assuming that the threshold probability of a disease or event at which a patient would opt for treatment is informative of how the patient weighs the relative harms of a false-positive and a false-negative prediction. This theoretical relationship is then used to derive the net benefit of the model across different threshold probabilities. Plotting net benefit against threshold probability yields the "decision curve." The authors apply the method to models for the prediction of seminal vesicle invasion in prostate cancer patients. Decision curve analysis identified the range of threshold probabilities in which a model was of value, the magnitude of benefit, and which of several models was optimal.* ***Conclusion***. *Decision curve analysis is a suitable method for evaluating alternative diagnostic and prognostic strategies that has advantages over other commonly used measures and techniques.* ***Key words:*** *prediction models; multivariate analysis; decision analysis.* ***(Med Decis Making 2006;26:565–574)***

A typical prediction model provides the probability of an event, such as recurrence after surgery for prostate cancer, on the basis of a set of prognostic factors, such as cancer stage and grade.[1] Such models can be used to predict disease outcome, as in the case of cancer recurrence, or to make a diagnosis, such as whether a patient has appendicitis. Prediction models are usually evaluated by applying the model to a data set and comparing the predictions of the model with actual patient outcome. Results are typically expressed as the area under the receiver operating characteristic (ROC) curve. The area under the curve (AUC) can be interpreted as the probability that in a pair of individuals, one who did and one who did not experience the event, the individual who experienced the event had the higher predicted probability. As such, it is commonly used as a single statistic to compare 2 or more prediction models.[2–4]

The AUC metric focuses solely on the predictive accuracy of a model. As such, it cannot tell us whether the model is worth using at all or which of 2 or more models is preferable. This is because metrics that concern accuracy do not incorporate information on consequences. Take the case where a false-negative result is much more harmful than a false-positive result. A model that had a much greater specificity but slightly lower sensitivity than another would have a higher AUC but would be a poorer choice for clinical use.

Decision-analytic methods incorporate consequences and, in theory, can tell us whether a model is worth using at all or which of several alternative models should be used.[5] In a typical decision analysis, possible consequences of a clinical decision are identified and the expected outcomes of alternative

Received 1 November 2005 from the Department of Epidemiology and Biostatistics (AJV, EBE), Department of Urology (AJV), and Department of Medicine (AJV), Memorial Sloan-Kettering Cancer Center, New York. Dr. Vickers's work on this research was funded by a P50-CA92629 SPORE from the National Cancer Institute. Revision accepted for publication 6 April 2006.

Address correspondence to Andrew J. Vickers, PhD, 1275 York Ave., New York, NY 10021; telephone: (212) 639-6556; fax: (212) 794-5851; e-mail: vickersa@mskcc.org.

DOI: 10.1177/0272989X06295361

clinical management strategies then simulated using estimates of the probability and sequelae of events in a hypothetical cohort of patients. Decision analysis requires explicit valuation of health outcomes, such as the number of complications prevented, life-years saved, or quality-adjusted life-years saved. In a decision analysis of alternative diagnostic or prognostic models, the optimal model is the one that maximizes the outcome of interest. Techniques have been proposed to simplify decision analyses of diagnostic and prognostic tests by using a risk-benefit ratio to summarize the health outcomes associated with the consequences of testing.[6]

There are 2 general problems associated with applying traditional decision-analytic methods to prediction models. First, they require data, such as on costs or quality-adjusted life-years, not found in the validation data set—that is, the result of the model and the true disease state or outcome. This means that a prediction model cannot be evaluated in a decision analysis without further information being obtained. Moreover, decision-analytic methods often require explicit valuation of health states or risk-benefit ratios for a range of outcomes. Health state utilities, used in the quality adjustment of expected survival,[7] are prone to a variety of systematic biases[8] and may be burdensome to elicit from subjects. The second general problem is that decision analysis typically requires that the test or prediction model being evaluated give a binary result so that the rate of true- and false-positive and negative results can be estimated. Prediction models often provide a result in continuous form, such as the probability of an event from 0% to 100%. To evaluate such a model using decision-analytic methods, the analyst must dichotomize the continuous result at a given threshold and potentially evaluate a wide range of such thresholds.

We sought a method for evaluating prediction models that incorporates consequences and so can be used to make decisions about whether to use a model at all or which of several models to use. We hoped to improve upon currently available techniques by developing a method that can be applied directly to a validation data set and does not require the collection of additional information. Moreover, we required a method that could be applied to a model regardless of whether it gave a binary or continuous result. Here we present a novel technique, decision curve analysis, which fulfills these criteria.

INTRODUCTORY THEORY

Take the case of a patient deciding whether to undergo treatment for a specific disease. The patient is unsure whether disease is present. A simple decision tree is given in Figure 1: p is the probability of disease, and a, b, c, and d give the value associated with each outcome in terms such as quality-adjusted life-years. Let us imagine that there is a prediction model available. This provides a probability that the patient has the disease: if the probability of disease is near 1, the patient will ask to be treated; if the probability is near 0, he or she is likely to forgo treatment. At some probability between 0 and 1, the patient will be unsure whether to be treated. This threshold probability, p_t, is where the expected benefit of treatment is equal to the expected benefit of avoiding treatment. Solving the decision tree:

$$p_t a + (1-p_t)\,b = p_t c + (1-p_t)d.$$

By some simple algebra:

$$p_t a - p_t c = (1-p_t)d - (1-p_t)\,b$$

$$\Rightarrow p_t(a-c) = (1-p_t)(d-b)$$

$$\Rightarrow \frac{a-c}{d-b} = \frac{1-p_t}{p_t}.$$

Now $d - b$ is the consequence of being treated unnecessarily. If treatment is guided by a prediction model, this is the harm associated with a false-positive result (compared to a true-negative result). Comparably, $a - c$ is the consequence of avoiding treatment when it would have been of benefit, that is, the harm from a false-negative result (compared to a true-positive result). Equation (1) therefore tells us that *the threshold probability at which a patient will opt for treatment is informative of how a patient weighs the relative harms of false-positive and false-negative results.* In this formulation, "harm" is considered holistically as the overall effect of all negative consequences of a particular decision. Our formula has been described previously to derive an optimal threshold for an action such as using a drug or performing a diagnostic test.[9,10] In a typical example, Djulbegovic and others[11] use data from a randomized trial to estimate the benefit and harm of prophylactic treatment for deep vein thrombosis (DVT). They find that if a patient's risk of DVT is 15% or more, he or she should be treated; if it is less than 15%, treatment should be avoided. Our method allows this threshold to vary, depending on uncertainties associated with the likelihood of each outcome and differences between individuals as to how they value outcomes.

	治	不治
病	p : a	p : c
没病	1-p : b	1-p:d

其中a,b,c,d都是数

(a-c)/(d-b) = (1-p)/p

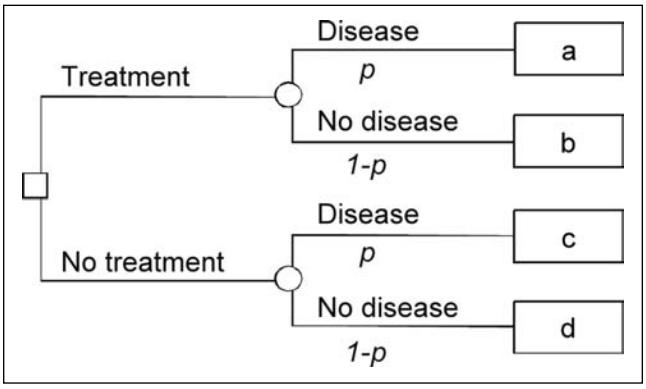

Figure 1 A decision tree for treatment. The probability of disease is given by p; a, b, c, and d give, respectively, the value of true positive, false positive, false negative, and true negative.

PRINCIPAL EXAMPLE

The example that we use to illustrate our methodology comes from a prostate cancer study. Surgery for prostate cancer normally involves total removal of the seminal vesicles as well as the prostate, on the grounds that the tumor may invade the seminal vesicles. The presence of seminal vesicle invasion (SVI) can be observed prior to or during surgery only in rare cases of widespread disease. SVI is therefore typically diagnosed after surgery by pathologic examination of the surgical sample. It has recently been suggested that the likelihood of SVI can be predicted on the basis of information available before surgery, such as cancer stage, tumor grade, and prostate-specific antigen (PSA).[12] Although some surgeons will remove the seminal vesicles regardless of the predicted probability of SVI, others have argued that patients with a low predicted probability of SVI might be spared total removal of the seminal vesicles: most of the seminal vesicles would be dissected, but the tip, which is in close proximity to several important nerves and blood vessels, would be preserved. According to this viewpoint, sparing the seminal vesicle tip might therefore reduce the risk of common side effects of prostatectomy such as incontinence and impotence.[13] Previous investigators have published both binary decision rules[12] and multivariable prediction models[13] to help clinicians identify candidates for tip-sparing surgery. These investigators typically present metrics such as sensitivity, specificity, or AUC to evaluate their models.[13] Accordingly, they are unable to tell us whether their model does more good than harm and therefore should actually be used. To demonstrate the use of decision curve analysis, we used data from an unpublished study of 902 men with prostate cancer who underwent prostatectomy and developed a multivariable model that gave the probability of SVI on the basis of stage, grade, and PSA.

We use this example to illustrate equation (1). Take the case of a surgeon who needs to decide whether to dissect or preserve the seminal vesicle tip in a man scheduled for prostatectomy. The surgeon suspects that total dissection may increase the risk of impotence or incontinence; however, preservation might increase the chance of a cancer recurrence if the patient had SVI and the tumor extended to the seminal vesicle tip. Imagine that the surgeon would definitely opt for total seminal vesicle dissection if the patient's predicted risk of SVI were 30% and that preservation would be chosen if the risk were only 1%, but if the risk were 10%, the surgeon would be uncertain as to the correct approach. By equation (1), we infer that the surgeon feels that, for this patient, failing to remove the tip of a cancerous seminal vesicle (i.e., a false-negative result) is 9 times worse than unnecessary tip dissection (i.e., a false-positive result).

APPLICATION: DECISION CURVE ANALYSIS

As it turns out, our method does not require that we obtain information regarding treatment preferences in this way. We use the *theoretical* relationship between the threshold probability of disease and the relative value of false-positive and false-negative results to ascertain the value of a prediction model. Take a group of patients scheduled for treatment by a surgeon who would be unsure whether to preserve or remove the seminal vesicle tip if the probability of SVI were 10%. We can now calculate each patient's probability of SVI using the multivariable model and classify the result positive if it is equal to or higher than 10% and negative otherwise. Applying these results to the data set yields the data shown in Table 1.

To place a value on this result, we fix $a - c$, the value of a true-positive result, at 1. We then obtain the value of a false-positive result, $b - d$, as $-p_t/(1 - p_t)$. We can now calculate net benefit using the following formula (first attributed to Peirce[14]):

$$\text{Net benefit} = \frac{\text{true-positive count}}{n} - \frac{\text{false-positive count}}{n}\left(\frac{p_t}{1-p_t}\right).$$

Table 1 Relationship between True Seminal Vesicle Invasion (SVI) Status and Result of a Prediction Model with a Positivity Criterion of 10% Predicted Probability of SVI

		SVI	
	***N* = 902**	**Positive**	**Negative**
Prediction model: probability of SVI ≥ 10%	Yes	65	225
	No	22	590

In this formula, true- and false-positive count is the number of patients with true- and false-positive results, and n is the total number of patients. In short, we subtract the proportion of all patients who are false positive from the proportion who are true positive, weighting by the relative harm of a false-positive and a false-negative result. In Table 1, where p_t is 10%, the true-positive count is 65, the false-positive count is 225, and the total number of patients (N) is 902. The net benefit is therefore (65/902) – (225/902) × (0.1/0.9) = 0.0443. A good model will have a high net benefit: the theoretical range of net benefit is from negative infinity to the incidence of disease.

To determine whether this value is a good one—that is, whether the prediction model should be used for a p_t of 10%—we need a comparison. The clinical alternative to using a prediction model is to assume that all patients are positive and treat them—as might be done for individuals possibly exposed to a dangerous infection easily treated with antibiotics—or assume that all patients are negative and offer no treatment, as is done for diseases for which there are no proven screening methods. The true- and false-positive counts for considering all patients negative are both 0, and hence the net benefit for leaving the seminal vesicle tip in all patients is 0. Hence, if the net benefit for the prediction model is positive, it is better to use the model than to assume that everyone is negative. The true- and false-positive counts for the strategy of treating all patients are simply the number of patients with and without SVI, respectively. Calculating net benefit gives (87/902) – (815/902) × (0.1/0.9) = –0.0039 for the strategy of removing seminal vesicles in all patients. This is less than the net benefit of 0.0443 from the prediction model.

At a p_t of 10%, our prediction model is therefore better than both treating no one and treating everyone. However, patients differ as to how they rate possible side effects of surgery. For example, a surgeon might be tempted treat more aggressively a man who was impotent but had many responsibilities. For such a man, the surgeon might use a much lower p_t, say, 2%—that is, the seminal vesicle tip would be removed even if there was only a 2% chance of SVI. At this p_t, the strategies of treating all men and treating using the model are almost identical (net benefit of 0.0780 and 0.0782, respectively). Similarly, there is a difference of opinion between surgeons regarding the increase in recurrence risk associated with preservation of the seminal tip in a patient with SVI: some surgeons feel that even if a patient has SVI, it is unlikely that the seminal vesicle tip will be involved, and even then, it is not clear that preservation will inevitably lead to recurrence; other surgeons feel that leaving any part of a cancerous seminal vesicle will substantively increase recurrence rates. We therefore recommend repeating the above steps for different values of p_t. Hence:

1. Chose a value for p_t.
2. Calculate the number of true- and false-positive results using p_t as the cut-point for determining a positive or negative result.
3. Calculate the net benefit of the prediction model.
4. Vary p_t over an appropriate range and repeat steps 2 and 3.
5. Plot net benefit on the y-axis against p_t on the x-axis.
6. Repeat steps 1 through 5 for each model under consideration.
7. Repeat steps 1 through 5 for the strategy of assuming all patients are positive.
8. Draw a straight line parallel to the x-axis at $y = 0$ representing the net benefit associated with the strategy of assuming that all patients are negative.

全预测为”治病“，那随着阈值增加TP下降，FP增加benefit 会随着阈值增加而下降

全部预测为“不治病“，这时TP=0,FP=0，benefit == 0

Applying these steps to our data gives Figure 2. We term this a *decision curve*. Note that, as expected, the 2 lines reflecting the strategies of “assume all patients have SVI” (i.e., treat all) and “assume no patients have SVI” (i.e., treat none) cross at the prevalence. 患病率 Also note that the prediction model is comparable to the strategy of treat all at low p_t and comparable to treat none at high p_t. This is because the probability of SVI predicted by the model ranges from a minimum of 1.8% to a maximum of 84.3%. Using the model for $p_t < 1.8\%$ or $p_t > 84.3\%$ therefore gives the same result as treat all or treat none, respectively. Between 50% and 84.3%, the value of the model is sometimes negative: this is due to random noise.

Between these 2 extremes, there is a range of p_t where the prediction model is of value. In the case of SVI prediction, this is between ~2% and ~50%. To

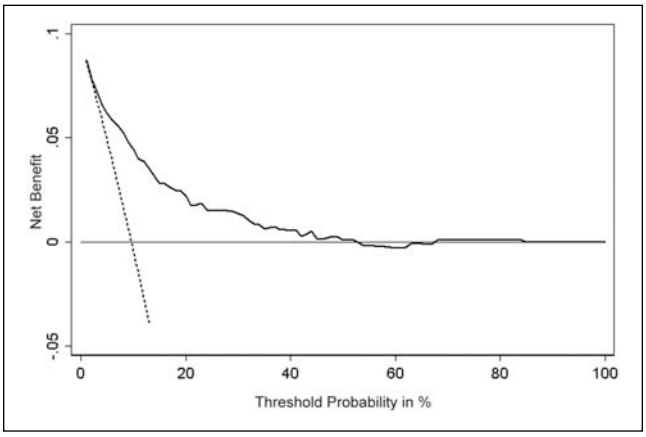

Figure 2 Decision curve for a model to predict seminal vesicle invasion (SVI) in patients with prostate cancer. Solid line: prediction model. Dotted line: assume all patients have SVI. Thin line: assume no patients have SVI. The graph gives the expected net benefit per patient relative to no seminal vesicle tip removal in any patient ("treat none"). The unit is the benefit associated with 1 SVI patient duly undergoing surgical excision of the seminal vesicle tip.

determine whether the model is of clinical value, we need to consider the likely range of p_t in the population, that is, the typical threshold probabilities of SVI at which surgeons would opt for complete dissection of seminal vesicles. If it were the case that all surgeons remove the seminal vesicle tip only if there was at least a 60% to 70% risk of SVI, the model clearly has no clinical role. But it is unlikely that any surgeon would consider removal of a healthy seminal vesicle tip to be worse than failing to remove a potentially cancerous one. If, on the other hand, we assume that the likely range of p_t in the population is between 20% and 30%, we would use the model because it is of clear benefit at these p_ts.

In consultation with clinicians, we estimate that although few if any surgeons would ever have a p_t much above 10% for any patient, some may have p_t approaching 1% or less in certain cases. This means that our prediction model will be of benefit in some but not all cases. Where p_t is less than 2%, the model is no better than a strategy of treating all patients. Hence, where p_t is less than 2%, the model is of no value, and patients should have total seminal vesicle dissection. On the other hand, the model is never worse than the strategy of treating all patients, and because it is based on routinely collected data, it has no obvious downside. Therefore, the model will be of use for clinicians who, at least some of the time, would opt for seminal vesicle tip preservation if a patient's predicted probability of SVI was low.

If the prediction model required obtaining data from medical tests that were invasive or dangerous or involved expenditure of time, effort, and money, we can use a slightly different formulation of net benefit:

$$\text{Net benefit} = \frac{\text{true-positive count}}{n} - \frac{\text{false-positive count}}{n}\left(\frac{p_t}{1-p_t}\right) - \text{test harm.}$$

The harm from the test is a "holistic" estimate of the negative consequence of having to take the test (cost, inconvenience, medical harms, etc.) in the units of a true-positive result. For example, if a clinician or a patient thought that missing a case of disease was 50 times worse than having to undergo testing, the test harm would be rated as 0.02. Test harm can also be thought of in terms of the number of patients a clinician would subject to the test to find 1 case of disease if the test were perfectly accurate.

If the test were harmful in any way, it is possible that the net benefit of testing would be very close to or less than the net benefit of the "treat all" strategy for some p_t. In such cases, we would recommend that the clinician have a careful discussion with the patient and perhaps, if appropriate, implement a formal decision analysis. In this sense, interpretation of a decision curve is comparable to interpretation of a clinical trial: if an intervention is of clear benefit, it should be used; if it is clearly ineffective, it should not be used; if its benefit is likely sufficient for some but not all patients, a careful discussion with patients is indicated.

EXTENSIONS OF DECISION CURVE ANALYSIS

Decision curve analysis has 2 important additional advantages. First, the benefit of using a prediction model can be quantified in simple, clinically applicable terms. Table 2 gives the results of our analysis for p_ts between 1% and 10%. The net benefit of 0.062 at a p_t of 5% can be interpreted in terms that use of the model, compared with assuming that all patients are negative, leads to the equivalent of a net 6.2 true-positive results per 100 patients without an increase in the number of false-positive results. In terms of our specific example, we can state that if we perform surgeries based on the prediction model, compared to tip preservation in all patients, the net consequence is equivalent to removing the tip of affected seminal

Table 2 Net Benefit for Removing the Tip of the Seminal Vesicles from All Patients or According to a Prediction Model, Using a Threshold of p_t

	Net Benefit		Advantage of Model	
p_t (%)	Treat All	Prediction Model	Net Benefit	Reduction in Avoidable Tip Surgeries per 100 Patients
1	0.087	0.087	0	0
2	0.078	0.078	0	0
3	0.069	0.072	0.004	13
4	0.059	0.066	0.007	17
5	0.049	0.062	0.013	25
6	0.039	0.059	0.020	31
7	0.028	0.056	0.027	36
8	0.018	0.053	0.035	40
9	0.007	0.048	0.041	41
10	−0.004	0.044	0.048	43

Note: The reduction in the number of unnecessary surgeries removing the seminal vesicle tip per 100 patients is calculated as follows: (net benefit of the model − net benefit of treat all)/(p_t/(1 − p_t)) × 100. This value is net of false negatives and is therefore the equivalent to the reduction in unnecessary surgeries without a decrease in the number of patients with seminal vesicle invasion who duly have tip surgery.

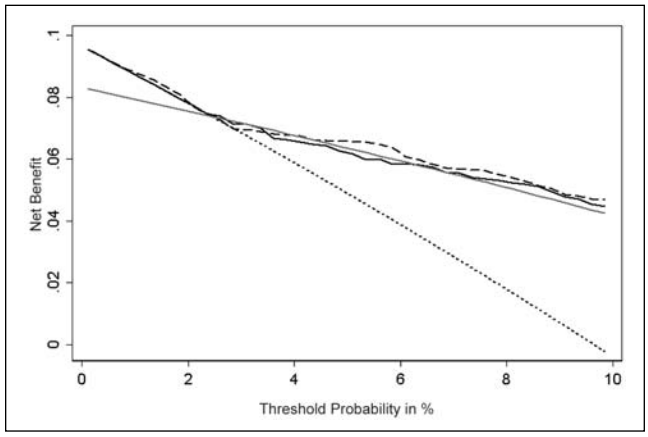

Figure 3 Decision curve for seminal vesicle invasion (SVI): comparison of 3 models. Dotted line: assume all patients have SVI. Grey line: binary decision rule. Solid line: basic prediction model. Dashed line: expanded prediction model incorporating additional biomarkers. The graph gives the expected net benefit per patient relative to no seminal vesicle tip removal in any patient ("treat none"). The unit is the benefit associated with 1 SVI patient duly undergoing surgical excision of the seminal vesicle tip.

vesicles in 6.2 patients per 100 and treating no unaffected patients. Moreover, at a p_t of 5%, the net benefit for the prediction model is 0.013 greater than assuming all patients are positive. We can use the net benefit formula to calculate that this is the equivalent of a net $0.013 \times 100/(0.05/0.95) = 25$ fewer false-positive results per 100 patients. In other words, use of the prediction model would lead to the equivalent of 25% fewer tip surgeries in patients without SVI, with no increase in the number of patients with an affected seminal vesicle left untreated.

A second advantage of decision curve analysis is that it can be used to compare several different models. To illustrate this, we compare the basic prediction model with an expanded model and with a simple clinical decision rule. The expanded model includes all of the variables in the basic model as well as some additional biomarkers. The clinical decision rule separates patients into 2 risk groups based on Gleason grade and tumor stage: those with grade greater than 6 or stage greater than 1 are considered high risk. To calculate a decision curve for this rule, we used the methodology outlined above except that the proportions of true- and false-positive results remained constant for all levels of p_t. Figure 3 shows the decision curve for these 3 models in the key range of p_t from 1% to 10%. There are 3 important features to note. First, although the expanded prediction model has a better AUC than the basic model (0.82 v. 0.80), this makes no practical difference: the 2 curves are essentially overlapping. Second, the basic model has a considerably larger AUC than the simple clinical rule, yet for p_ts above 2%, there is essentially no difference between the 2 models. Third, at some low values of p_t, using the simple clinical rule actually leads to a poorer outcome than simply treating everyone, despite a reasonably high AUC (0.72). In addition to illustrating the use of decision curves to compare multiple prediction models, Figure 3 also demonstrates that the methodology can easily be applied to a test or model with an inherently binary outcome, such as the simple clinical decision rule.

ADDITIONAL EXAMPLE: PROGNOSIS

In addition to evaluating diagnostic models, such as the model for SVI in men with prostate cancer, decision curves can be used to assess the value of prognostic models. For example, a number of models have been developed to obtain a patient's preoperative probability of prostate cancer recurrence at 5 years based on PSA, clinical stage, and grade of cancer at biopsy.[15,16] These models are used to counsel patients but also to inform decision making because a patient with a high risk of recurrence might be asked to consider adjuvant hormonal therapy. We use an observational cohort of men treated for prostate

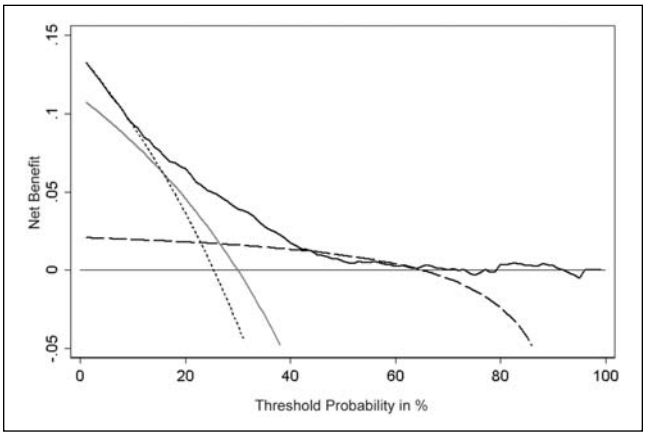

Figure 4 Decision curve for prediction of recurrence after surgery for prostate cancer. Thin line: assume no patient will recur. Dotted line: assume all patients will recur. Long dashes: binary decision rule based on cancer grade ("Gleason rule"). Grey line: binary decision rule based on both grade and stage of cancer ("stage rule"). Solid line: multivariable prediction model. The graph gives the expected net benefit per patient relative to no hormonal therapy for any patient ("treat none"). The unit is the benefit associated with 1 patient who would recur without treatment and who receives hormonal therapy.

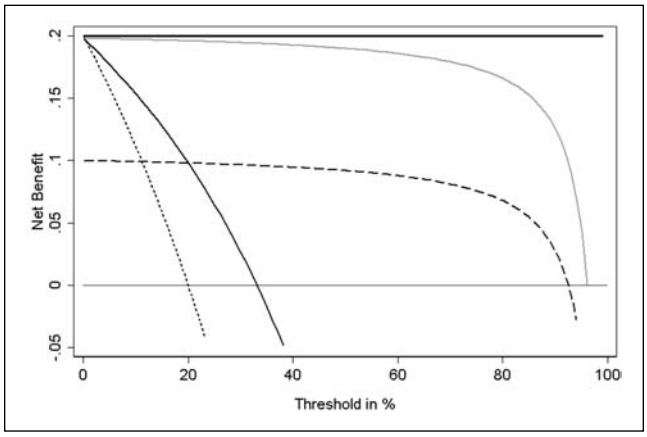

Figure 5 Decision curve for a theoretical distribution. In this example, disease incidence is 20%. Thin line: assume no patient has disease. Dotted line: assume all patients have disease. Thick line: a perfect prediction model. Grey line: a near-perfect binary predictor (99% sensitivity and 99% specificity). Solid line: a sensitive binary predictor (99% sensitivity and 50% specificity). Dashed line: a specific binary predictor (50% sensitivity and 99% specificity).

cancer to examine 3 models that predict recurrence within 5 years: 1 based on a multivariable model ("model") and 2 clinical rules: "assume that any man with Gleason grade 8 or above will recur" ("Gleason rule") and "assume that any man with Gleason grade 8 or above, or stage 2 or above, will recur" ("stage rule"). We assess these models to determine how well they aid the decision to undergo or not undergo adjuvant hormonal therapy.

Figure 4 shows the decision curves. Although cancer recurrence is a serious event, the benefit of adjuvant hormonal therapy is somewhat unclear, and the drugs have important side effects such as hot flashes, decreased libido, and fatigue. Due to differences in opinion about the value of hormonal therapy, as well as differences between patients in the importance attached to side effects, the probability used as a threshold to determine hormonal therapy varies from case to case. A typical range of p_ts is 30% to 60%—that is, if you took a group of clinicians and patients and documented the probability of recurrence at which the clinician would advise hormonal therapy, this would vary between 30% and 60%. For much of this range, the simple Gleason rule is comparable to the multivariable model, even though it has far less discriminative accuracy (e.g., AUC 0.56 compared to 0.73). Moreover, the Gleason rule is much better than the stage rule in the key 30% to 60% range even though it has a lower AUC (0.56 v. 0.58). This is no doubt because the stage rule is highly sensitive and the Gleason rule more specific. But although sensitivity and specificity give a general indication as to which test is superior in which situation, decision curve analysis can delineate precisely the conditions under which each test should be preferred.

As a further illustration, Figures 5 and 6 show the decision curves for some theoretical distributions. Note that the results of the decision curve analysis accord well with our expectations. For example, a sensitive predictor is superior to a specific predictor where p_t is low—that is, where the harm of a false negative is greater than the harm of a false positive. The situation is reversed at high p_t: the curves for the sensitive and specific predictor cross near the incidence; a near-perfect predictor is of value except where p_t is close to 1, that is, where the patient or the clinician has to be nearly certain before taking action. A predictor that is 2 standard deviations higher in patients who have the event is superior across nearly the full range of p_t.

DISCUSSION

Given the exponential increase in molecular markers in medicine and the integration of information

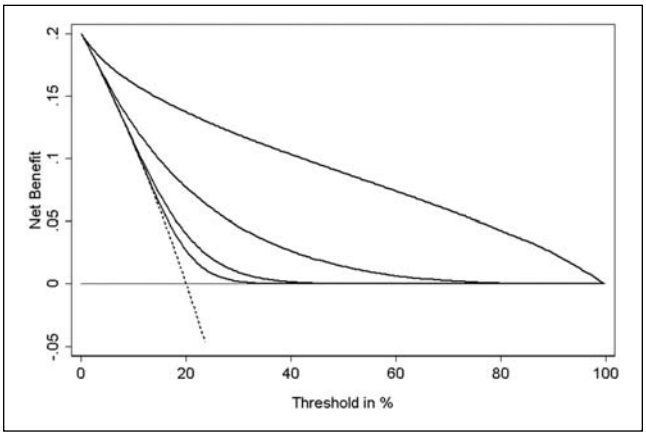

Figure 6 Decision curve for a theoretical distribution. In this example, disease incidence is 20%; the predictor example is a normally distributed laboratory marker. Thin line: assume no patient has disease. Dotted line: assume all patients have disease. Solid lines: prediction model from a single, continuous laboratory marker: from left to right, the lines represent a mean shift of 0.33, 0.5, 1, and 2 standard deviations in patients with disease.

technology into clinical management, the development of prognostic and diagnostic models is likely to increase.[1] This lends urgency to the search for appropriate methods to determine the value of such models.

We have introduced a novel method for the evaluation of prediction models. This method is decision analytic in nature and can therefore inform the decision of whether to use a model at all or which of several models is optimal. Applying our method to several prediction models on 2 data sets confirms the general principle that metrics of accuracy, such as sensitivity, specificity, and AUC, do not address the clinical value of a model. Although a model with a higher AUC is likely to be more valuable than one with a lower AUC, we have shown that, as would be expected from decision theory, models with very different AUCs can be comparable and that models with higher AUCs can sometimes lead to inferior outcomes.

Decision curve analysis can be applied to both multivariable prediction models that give the probability of an event and to standard diagnostic tests that produce a simple binary result. Moreover, this method does not require information on the costs or effectiveness of treatment or how patients value different health states. We see this as beneficial because the method can be directly applied to a model validation data set. A search of the medical literature suggests that the number of studies on diagnostic and prognostic markers using metrics of accuracy, such as the AUC, dwarfs the number of those using decision-analytic methods. This is likely due to the burden of obtaining additional data, especially if health state utilities are required.

Nonetheless, *interpretation* of a decision curve analysis requires some understanding of the likely range of patients' values. We believe that this is comparable to interpretation of clinical trial results. To determine the value of a treatment tested in a trial, a clinician needs to have some idea of just how effective the treatment needs to be before patients would be prepared to take it, bearing in mind its side effects. If the benefits of treatment are either much larger or smaller than most patients would require, the clinician can either give or withhold treatment without a detailed understanding of a patient's preferences. If, on the other hand, effects of the treatment might be large enough for some patients but not for others, a careful discussion with the patient is indicated, perhaps with formal elicitation of health state preferences and a decision analysis. In short, we do not propose decision curve analysis as a substitute for existing decision-analytic methods, though it may help indicate where such methods may be of benefit.

Similarly, we are not suggesting that decision curve analysis can replace measures of accuracy such as sensitivity and specificity. First, such measures are vital in the early stages of developing diagnostic and prognostic strategies—for example, when determining whether a biomarker shows any evidence of value on a convenience sample or when calibrating instruments or techniques.[17] Second, although it is possible for a model to be accurate but useless, the converse is not true: those proposing diagnostic or prognostic methods must show that they are reasonably accurate as well as demonstrating that they improve decision making.

As pointed out when describing our methods, several previous workers have used the relative benefit and harm associated with true- and false-positive results to determine a single, optimal threshold for a diagnostic test.[10,11,18] Determining a single threshold is only possible under 2 conditions: first, the benefits and harms of action must be well understood; second, how benefits and harms are valued must be similar between individuals. As an illustration, compare one of the examples given by Djulbegovic and others[11]—prevention of DVT—with our prostatectomy example. In the DVT case, precise estimates for the treatment benefit (reduction in rate of DVT) and harm (major bleeding) are available from a randomized trial; moreover, there are no important differences between individuals as to the relative harm of a DVT and a bleed: the authors use the assumption that "the avoidance of DVT and bleeding complications represents

approximately the same value to the patient."[11] In the prostatectomy example, conversely, estimates of the benefits (improved urinary and erectile function) and harms (increased rate of cancer recurrence) of the seminal vesicle tip–sparing approach are available only from observational studies of moderate size[13] and are subject to considerable disagreement. Moreover, how different individuals value potency and continence compared to cancer recurrence varies greatly. Other investigators have used the relative benefits and harms associated with different test outcomes in net benefit or loss functions to compare predictive models.[19,20] However, this requires investigators to specify a value for harms and benefits. For example, Habbema and Hilden[20] describe a method to assess a prediction tool for management of acute abdominal pain by ascribing losses to outcomes such as failure to diagnose appendicitis (36 units), appendectomy for a patient with a healthy appendix (10 units), and intensive follow-up in a patient with nonspecific abdominal pain (2 units). Such methods can be adapted for sensitivity analysis using different values for such outcomes, but this is generally complex and does not clearly maintain the inherent relationship between values and probability thresholds. Our method combines values and thresholds in a simple, parsimonious method to determine whether a predictive model should be used clinically while allowing each to covary appropriately.

One assumption of our method is that the predicted probability and threshold probability are independent. This is true for the examples we give in this article: there is no relationship between, say, the appearance of a cancer cell under the microscope (Gleason grade) and how a patient values sexual and urinary function relative to cancer recurrence. It is possible that a third variable, such as age, might influence both the probability of recurrence and treatment preferences; however, in our data set, the correlation between age and SVI was very low (0.04). We think that an important correlation between predicted probability and threshold probability will be very much the exception rather than the rule. One possible example would be gender. If gender was indeed correlated with both outcome and threshold probability, the analyst might consider constructing a decision curve separately for men and women.

In the examples presented here, we have not considered the uncertainty associated with model predictions and their possible impact on the decision curve. We are currently evaluating methods to characterize uncertainty, including confidence bands and metrics such as the probability that the net benefit of a model is superior to a comparator.

Hilden[21] has written of the schism between what he describes as "ROCographers," those who are interested solely in accuracy, and "VOIographers," who are interested in the clinical value of information (VOI). He notes that although the former ignore the fact that their methods have no clinical interpretation, the latter have not agreed on an appropriate mathematical approach. We feel that decision curve analysis may help bridge this schism by combining the direct clinical applicability of decision-analytic methods with the mathematical simplicity of accuracy metrics.

REFERENCES

1. Freedman AN, Seminara D, Gail MH, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. J Natl Cancer Inst. 2005;97:715–23.

2. Das SK, Baydush AH, Zhou S, et al. Predicting radiotherapy-induced cardiac perfusion defects. Med Phys. 2005;32:19–27.

3. Hendriks DJ, Broekmans FJ, Bancsi LF, Looman CW, De Jong FH, Te Velde ER. Single and repeated GnRH agonist stimulation tests compared with basal markers of ovarian reserve in the prediction of outcome in IVF. J Assist Reprod Genet. 2005;22: 65–73.

4. Cindolo L, Patard JJ, Chiodini P, et al. Comparison of predictive accuracy of four prognostic models for nonmetastatic renal cell carcinoma after nephrectomy. Cancer. 2005;104:1362–71.

5. Hunink M, Glasziou P, Siegel J. Decision-Making in Health and Medicine: Integrating Evidence and Values. New York: Cambridge University Press; 2001.

6. Djulbegovic B, Desoky AH. Equation and nomogram for calculation of testing and treatment thresholds. Med Decis Making. 1996;16:198–9.

7. Loomes G, Mckenzie L. The use of QALYs in health care decision making. Soc Sci Med. 1989;28:299–308.

8. Van Osch SM, Wakker PP, Van Den Hout WB, Stiggelbout AM. Correcting biases in standard gamble and time tradeoff utilities. Med Decis Making. 2004;24:511–7.

9. Weinstein MC, Fineberg HV. Clinical Decision Analysis. Philadelphia: W. B. Saunders; 1980.

10. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109–17.

11. Djulbegovic B, Hozo I, Lyman GH. Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. Medgenmed. 2000;2:E6.

12. Zlotta AR, Roumeguere T, Ravery V, et al. Is seminal vesicle ablation mandatory for all patients undergoing radical prostatectomy? A multivariate analysis on 1283 patients. Eur Urol. 2004;46:42–9.

13. Guzzo TJ, Vira M, Wang Y, et al. Preoperative parameters, including percent positive biopsy, in predicting seminal vesicle involvement in patients with prostate cancer. J Urol. 2006;175: 518–21.

14. Peirce CS. The numerical measure of the success of predictions. Science. 1884;4:453–4.

15. Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst. 1998;90: 766–71.

16. Steuber T, Karakiewicz PI, Augustin H, et al. Transition zone cancers undermine the predictive accuracy of partin table stage predictions. J Urol. 2005;173:737–41.

17. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst. 2001;93: 1054–61.

18. Moons KG, Stijnen T, Michel BC, et al. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. Med Decis Making. 1997;17:447–54.

19. Parmigiani G. Modeling in Medical Decision Making. New York: John Wiley; 2002.

20. Habbema JDF, Hilden J. The measurement of performance in probabilistic diagnosis: IV. Utility considerations in therapeutics and prognostics. Meth Inform Med. 1978;17:238–46.

21. Hilden J. Evaluation of diagnostic tests: the schism. Soc Med Decis Making Newsletter. 2004;16:5–6.