Published in final edited form as:

Med Decis Making. 2008 ; 28(1): 146–149. doi:10.1177/0272989X07312725.

Decision curve analysis: a discussion

Ewout W. Steyerberg, Ph.D. and

Dept of Public Health, Erasmus University, Rotterdam, Netherlands. E.Steyerberg@ErasmusMC.nl

Andrew J. Vickers, Ph.D.

Dept. of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, USA. vickersa@mskcc.org

Shortly after publication of “Decision curve analysis: a novel method for evaluating prediction models” in *Medical Decision Making* late last year¹, Ewout Steyerberg, wrote to the first author, Andrew Vickers, with some comments and questions. Their discussion is reproduced below.

Ewout Steyerberg: I enjoyed your paper introducing the concept of Decision Curve Analysis. I particularly like the way that the method can be directly applied to a data set, without needing to get the sort of information normally required for decision analysis, such as patient utilities or drug costs. I have previously published a similar idea, a weighted accuracy metric², and was wondering whether you could comment on the differences between this metric and the decision curve.

Andrew Vickers: Thanks for the reference: this was not something I had previously seen. There are several similarities between your method and ours, in particular, you use the threshold probability from a predictive model both to classify patients as positive or negative and to assign a relative weight to the cost of false negatives versus false positives. I think this supports a point that I have made elsewhere, namely that what underpins decision curve analysis is tried-and-tested decision theory, and that elements of our technique can be found in many previously developed decision analytic applications. I think there are two key differences between decision curve analysis and your weighted accuracy metric. First, decision curve analysis allows one to vary the threshold probability over an appropriate range. This is important because, often, either a) there are insufficient data on which to calculate a rational threshold, or b) patients can reasonably disagree about the appropriate threshold, due to different preferences for alternative health states. Indeed, in one of your examples in the cited paper², you state that there is no agreement on the cutoff and so use a “hypothetical cutoff” for illustrative purposes. Second, the results of a decision curve analysis – the net benefit of a model – can easily be stated in clinically applicable terms: either as the net increase in the proportion of appropriately treated patients or the net decrease in the proportion of patients treated unnecessarily. Your clinical usefulness metric, which is a percentage, has no such directly applicable interpretation.

ES: I think a main point of confusion concerns the fact that the threshold probability, p_t , can vary between patients. Are you saying that we should ask individual patients to give us probability thresholds, then work out where they are on the decision curve and choose a model accordingly? It will not sure be easy to get patients to tell you a threshold probability of disease above which they would take action, but below which they would not.

AV: I agree entirely that it is not easy to get threshold probabilities from individual patients. However, for decision curve analysis, *you don't have to get the threshold probability from a patient at all*. What the decision curve tells you is the range of threshold probabilities for which the prediction model would be of value. Once you have this range, you then need to consider (perhaps by informal discussions with clinicians) whether all patients would fall within the

range, all fall outside the range or whether some patients might fall in the range and some outside.

As an example, see the following decision curves from three separate prostate cancer biopsy data sets (figures 1, figures 2, 3). In each case, we created a statistical model including age, prostate specific antigen (PSA) and then an additional molecular marker: urokinase, free PSA and what I'll call "PSA X" (the main results haven't been published yet so I corrupted the data set slightly and am using the name of an imaginary marker). In the urokinase example (figure 3), the curve for the prediction model is only superior to the curve for "treat all" (i.e. biopsy everyone) for thresholds between 40% and 80%. For PSA X it is 15 – 35%; for free PSA it is 10 – 75%. To interpret these results, let's think about the sort of probability for prostate cancer that men would need before they would decide to have a biopsy. Missing prostate cancer is obviously something you want to avoid, although it is not a fast growing cancer, and it is unlikely that delaying diagnosis for a few months would lead to important harm. On the other hand, a biopsy is unpleasant (it requires an ultrasound probe to be placed in the rectum, and the prostate to be punctured 12 times with needles) and can cause side-effects such as infection and bleeding. A very risk averse man might opt for biopsy even if he had only a 10% risk of cancer. Someone less risk averse, but a little more concerned about the biopsy procedure, might want to have a 30 – 40% chance of cancer before agreeing to biopsy. However, I don't think that many men would demand, say, a 50% risk of cancer before they had a biopsy; this threshold would imply that an unnecessary biopsy is just as bad as a missed cancer. So one estimate for the range of p_t 's in the community might be 10 – 40%. So we can now see that while the urokinase model is totally useless, the free PSA model should help everyone. The results for the marker "PSA X" mean that it would help some patients, but not others. I would interpret these results as providing evidence that free PSA is a useful marker, that urokinase is not a useful marker and that the "PSA X" marker of is benefit to some, but not to others.

ES: If a patient has a p_t where a model has limited value, is that a problem? The likelihood that the model will change the decision for a specific patient is low, but calculating a predicted probability from a model is not that much work usually.

AV: In general, I would agree, and indeed this is what happened for the example we used in the original paper: we thought that a plausible range of threshold probabilities would be from 1% – 10%, the model was of value for p_t 's between 2 – 50%, however, the model was no worse than "treat all" at p_t 's <2% and, because it was based on routinely collected data, we recommended the use of the model. But the urokinase decision curve is a counter-example: why go to the trouble of analyzing urokinase if it won't help you make a decision?

ES: If patients truly have varying p_t , then we may want some kind of summary measure of clinical usefulness, e.g. an integral over the range of individual p_t (i.e. area-under-the-curve)?

AV: An integral would assume that there is a uniform distribution of threshold probabilities among patients, for example, that just as many men would opt for a biopsy with a 10% risk of cancer as would require a 40% probability before they would consent to biopsy. This is unlikely to be true: my guess is that most men would ask for biopsy if they had 20% or higher probability of prostate cancer, but not if their probability of cancer was less than 20%-indeed, this is close to the positive predictive value of the current PSA test – and that fewer men would have values of p_t close to 10% or 40%. So to get a summary measure, you'd have to go out and get some data on the distribution of patient preferences, either by asking about threshold probabilities directly, or getting, say, health state utilities for unnecessary biopsy and missed cancers and calculating threshold probabilities accordingly. Doing so would subvert the key advantage of decision curve analysis, which is that decision analytic methods can be applied directly to a data set without obtaining additional data. So we prefer to take the following approach: obtain

an estimate for a plausible range of p_t 's; if the decision curve is superior for one model across this range, use of the model would be of clinical benefit for all; if a model has the highest net benefit for some but not all p_t 's, the model predictions are useful for some patients but not for others. In the latter case, other considerations come into play to determine whether or not to use the model, such as whether the information needed for the model is expensive or time consuming to acquire, whether the net benefit for the model is actually worse than an alternative at any point, and whether the range of threshold probabilities for which the model is useful is thought to reflect all but a few patients, or conversely, a large segment of the population.

ES: I have often noted that the distribution of predicted probabilities from a model is important to their usefulness: models which don't separate risks that well probably aren't that useful. How is this incorporated in the decision curve? Also, I am wondering about the relationship between the decision curve graph and other graphs that have been proposed, which have disutility on the y axis.

AV: You can get some idea of the distribution of risks by examining where the decision curve for your model overlaps with "treat all" and "treat none". Look at figure 3 for example. The lowest probability from the urokinase model is 9%, although the lowest centile is around 20%. This is close to where you start to see a difference between the different decision curves. The highest predicted probability from the model is close to 100%, which is why the curve never touches the "treat none" line on the x axis. In figure 2, the 99th centile of probability for the PSA X model is 50%, which is where the curve is equivalent to "treat none". There is a very slight bump again, for a few outlying high probabilities. The reason for all this is straightforward: if the lowest predicted probability from your model is, say, 9%, then a strategy of using the model will obviously be identical to a strategy of "treat all" for threshold probabilities of 9% or less; similarly, if the highest predicted probability is 63%, using the model will give identical results to a strategy of treating none for all threshold probabilities of 63% or greater.

As regards the y axis, it is easy to convert to disutility, you would just change the formula from: true positive - false positives($p_t/(1-p_t)$), that is, good stuff minus bad stuff, to false negatives + false positives($p_t/(1-p_t)$), that is, add up the bad stuff. The axes would change, but there would be no difference in our conclusions about which model was best. I prefer the net benefit formulation to disutility because you fix the value of doing nothing at zero.

ES: It would be interesting for me to try out some decision curve analysis on some of my own data. I have a testis cancer model I am using as an example for my forthcoming book on clinical prediction models and it would be interesting to see the decision curve for this model. Is there software available that I can use to run these analyses?

AV: Code for implementing decision curve analysis in both R and Stata are available from <http://www.mskcc.org/mskcc/html/74366.cfm>. The R code saves threshold probabilities with the net benefit for each model; this can then be used to graph the decision curve. The Stata code produces a graph directly, optionally saving net benefits at each threshold as a data set.

ES: Here is what I got when I ran your code on my data set (figure 4). The issue here is whether to undergo additional resection based on the probability of having residual tumor. I have marked in where I think the optimal threshold should be (30%), but I guess it would not be unreasonable for others to disagree and have either slightly lower or higher thresholds³. The decision curve shows that the model is not of benefit in the sample of patients considered, in line with 0% weighted accuracy we have previously reported². All these patients should have resection, since risk predictions are all above the threshold. This is interesting because the model has good calibration and discrimination (area under the receiver operating characteristic curve 0.79). So the results of the decision curve analysis are important for understanding the

clinical value of the model, beyond what standard statistical performance measures may suggest.

References

1. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* Nov-Dec;2006 26(6):565–574. [PubMed: 17099194]
2. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? *Semin Urol Oncol* May;2002 20(2):96–107. [PubMed: 12012295]
3. Steyerberg EW, Marshall PB, Keizer HJ, Habbema JD. Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. *Cancer* Mar 15;1999 85(6):1331–1341. [PubMed: 10189139]

Figure 1. Decision curve analysis for prostate biopsy. Black line: model including age, PSA and free-to-total PSA ratio. Dashed line: Model including age and PSA only.

Figure 2. Decision curve analysis for prostate biopsy. Black line: model including age, PSA and the marker “PSA X”. Dashed line: Model including age and PSA only.

Figure 3. Decision curve analysis for prostate biopsy. Black line: model including age, PSA and urokinase. Dashed line: Model including age and PSA only.

Figure 4.
Decision curve for testis cancer model