Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests

Andrew J Vickers,[1] Ben Van Calster,[2,3] Ewout W Steyerberg[3]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, New York, NY 10017, USA

[2]KU Leuven, Department of Development and Regeneration, Leuven, Belgium

[3]Department of Public Health, Erasmus MC, 's-Gravendijkwal, Rotterdam, Netherlands

Correspondence to: A J Vickers vickersa@mskcc.org

Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmj.i6)

Cite this as: *BMJ* 2016;352:i6 http://dx.doi.org/10.1136/bmj.i6

Accepted: 08 December 2015

Many decisions in medicine involve trade-offs, such as between diagnosing patients with disease versus unnecessary additional testing for those who are healthy. Net benefit is an increasingly reported decision analytic measure that puts benefits and harms on the same scale. This is achieved by specifying an exchange rate, a clinical judgment of the relative value of benefits (such as detecting a cancer) and harms (such as unnecessary biopsy) associated with models, markers, and tests. The exchange rate can be derived by asking simple questions, such as the maximum number of patients a doctor would recommend for biopsy to find one cancer. As the answers to these sorts of questions are subjective, it is possible to plot net benefit for a range of reasonable exchange rates in a "decision curve." For clinical prediction models, the exchange rate is related to the probability threshold to determine whether a patient is classified as being positive or negative for a disease. Net benefit is useful for determining whether basing clinical decisions on a model, marker, or test would do more good than harm. This is in contrast to traditional measures such as sensitivity, specificity, or area under the curve, which are statistical abstractions not directly informative about clinical value. Recent years have seen an increase in practical applications of net benefit analysis to research data. This is a welcome development, since decision analytic techniques are of particular value when the purpose of a model, marker, or test is to help doctors make better clinical decisions.

SUMMARY POINTS

- Prediction models, diagnostic tests, and molecular markers are traditionally evaluated using statistics such as sensitivity and specificity; such statistics do not tell us whether the model, test, or marker would do more good than harm if used in clinical practice
- Decision analysis attempts to assess clinical value by incorporating clinical consequences, such as the benefit of finding disease early or the harm of unnecessary further testing
- Net benefit is a simple type of decision analysis in which harm is multiplied by an "exchange rate" to place it on the same scale as benefit
- It is relatively straightforward to specify an exchange rate by asking about common medical practice; net benefit can also be plotted against a range of exchange rates in what is called a "decision curve"
- Decision curves are now widely used in the literature to evaluate whether clinical use of prediction models, diagnostic tests, and molecular markers would do more good than harm

Decision making and net benefit

Traditional statistical measures for the evaluation of prediction models, markers, and tests include sensitivity, specificity, area under the curve, and calibration.[1] Such measures do not, however, provide an answer as to whether the model, marker, or test should be used in clinical practice. For instance, it is not clear how high the sensitivity, specificity, or area under the curve needs to be to warrant clinical use. It is similarly unclear what degree of miscalibration would suggest that a prediction model should not be used, or how to chose between two models, one with better calibration and the other with better discrimination.

Decision analysis attempts to tackle such problems by incorporating the clinical consequences of using a model, marker, or test. A key concept in decision analysis is the idea of a trade-off between different endpoints. An obvious example of a trade-off is when a treatment helps reduce one symptom, such as heartburn, but causes a quite different symptom, such as dry mouth, as a side effect. Diagnostic tests are subject to a similar problem: a test may lead to early identification and curative treatment in patients with a disease that has been correctly diagnosed, but unless specificity is 100%, some patients without disease will be subject to unnecessary further diagnostic work-up and interventions. This problem of trade-off makes it difficult to evaluate tests. For instance, imagine a test for cancer that led to biopsy in 25 patients with disease but also in 75 patients who were cancer-free. If a new test

reduces the number of unnecessary biopsies to 50, but only finds 22 cancers, we need to consider whether missing three cancers to avoid 25 biopsies is a good trade-off.

Net benefit is a simple type of decision analysis, with benefits and harms put on the same scale so that they can be compared directly. Net benefit is similar to the idea of net profit in business. Take the case of an importer who buys €1m of wine from France and sells it in the United States for $1.5m. To work out the profit, dollars and euros need to be on the same scale, using the currency exchange rate. If €1 is worth $1.25 then we calculate profit=income in dollars (1.5m)–expenditure in euros (1m)×1.25=$250 000. Net benefit applies a similar methodology to medical research, by specifying an exchange rate between different medical endpoints, such as finding cancer versus unnecessary biopsy.

In this paper we explain the use of net benefit for the evaluation of prediction models, molecular markers, and diagnostic tests. Models, markers, and tests can be grouped together as forms of "risk prediction" and are subject to similar principles of research design and statistical analysis.

Net benefit and risk prediction

To introduce the concept of net benefit in medical research, we use biopsy for prostate cancer as an example. Men with elevated levels of prostate specific antigen (PSA) are at increased risk of aggressive prostate cancer and often referred for biopsy of the prostate. But most men with high PSA levels have either no cancer or only low grade tumors, which do not need treatment.[2] Biopsy is not only invasive and unpleasant but can cause infection. Researchers have actively sought additional markers to use as a test in men with increased PSA levels to refine the indication for biopsy.[3]

Imagine that we wanted to analyze a study of a new marker for prostate cancer. The study included 100 men, all of whom had increased PSA levels with no obvious benign cause and were therefore candidates for biopsy. We will assume that the results are as for the example in the introduction: high grade disease in 25 patients; when blood samples were analyzed for all 100 patients, 72 had high levels of the new marker, of whom 22 had a high grade tumor.

To analyze the study using a net benefit approach, we need to define the exchange rate by considering the number of men a doctor would biopsy to find one with high grade prostate cancer. A urologist might say that although it is important to find aggressive cancers early, biopsy is uncomfortable and has risks, and most cancers can still be caught at a curable stage even if biopsy is deferred. One reasonable response would be that to find one man with high grade cancer, no more than 10 men should undergo biopsy. This implies that the harm of delaying diagnosis of a high grade cancer is nine times greater than that of an unnecessary biopsy (in 10 men undergoing biopsy, one cancer found equates to nine unnecessary biopsies for each cancer detected). So in our analysis we want to "weight" finding high grade cancer as nine times more important than avoiding unnecessary biopsy. Using a similar principle to calculating profit for importing wine, we can use 1÷9 as the exchange rate. We define net benefit as:

Benefit–(harm×exchange rate)The net benefit for carrying out a biopsy in all men is 25%–(75%×(1÷9))=16.7%; the net benefit if the marker had been used to determine biopsy is 22%–(50%×(1÷9))=16.4%. Because at this particular exchange rate net benefit is lower for the marker than for biopsy in all men, we can conclude that use of the marker to determine biopsy would lead to poorer clinical outcome than the current practice of biopsy in all men with increased PSA levels not clearly due to benign disease.

The unit of net benefit is true positives. So a net benefit of 16.4% means that the marker is equivalent to a strategy that led to biopsy in 164 men per 1000 at risk, with all biopsy results positive for cancer. This is comparable to the concept of profit. Leaving aside the problem of financial risk, a profit of $250 000 for a wine transaction is roughly the equivalent of just being given $250 000 without having to spend money on buying wine.

Another similarity between profit and net benefit is that the rank order is more important than the size of the difference. A wine merchant forced to choose between one of two competing trades would choose the more profitable one, pretty much irrespective of whether profit was higher by $1000 or by $100 000. Similarly, we generally choose the strategy with the highest net benefit, without worrying about the size of the difference in net benefit (though see comments on "test harm").

One obvious criticism of net benefit in such medical applications is that the exchange rate is a subjective variable. But it is straightforward to vary the exchange rate and see how it affects the results. For instance, if a doctor is willing to carry out 20 biopsies to find one patient with high grade cancer, the net benefit of biopsy in all men (that is, 25%–(75%×(1÷19))=21.1%) is still higher than that of using the marker (that is, 22%–(50%×(1÷19))=19.4%).

Net benefit and decision curves

We can go one step further by plotting net benefit for a wide range of exchange rates. However, to do so, we must first bring in an additional idea. One of the ways to decide on whether to have a prostate biopsy or not is to use a statistical prediction model. This might be based on routinely available clinical variables such as age, PSA level, and the results of digital rectal examination,[2] but it might also include novel markers.[4] The result of the model would be expressed in terms of a percentage risk of high grade prostate cancer. To determine whether the model gives a positive result (high risk, biopsy indicated) or a negative result (low risk, biopsy not indicated), we need to use a cut point in terms of a probability threshold. The key concept is that to be consistent, we need to use the same cut point for the statistical model as we do to determine the exchange rate between cancers found and unnecessary biopsies. A cut point of 10 biopsies for each high grade cancer is the equivalent of carrying out a biopsy in men with a

doi: 10.1136/bmj.i6 | *BMJ* 2016;352:i6 | the bmj

risk of ≥10%; if a clinician would be willing to conduct as many as 20 biopsies to find a high grade cancer, the probability threshold would be 5%.

To see how net benefit varies by different exchange rates, we use an algorithm:

1. Choose a threshold probability (p_t) to define when a patient is positive
2. Count the number of patients with a positive result (risk ≥p_t) who have the disease (true positives) versus those who have a positive result but are disease-free (false positives)
3. With N the total sample size, calculate the net benefit (see equation):

$$\text{Net benefit} = \frac{\text{True positives}}{N} - \frac{\text{False positives}}{N} \times \frac{p_t}{1-p_t}$$

Details of equation

4. Repeat steps 2 and 3 for a reasonable range of threshold probabilities.
5. Repeat all steps for each marker, model, or test in the study, as well as the "default" strategies of treating all men or no men as if the result is positive.

Figure 1 shows a graph obtained from applying this algorithm to the marker described above and to a hypothetical statistical prediction model. This type of graph is known as a "decision curve"[5] and shows the net benefit of the marker and the statistical model as well the two clinical alternatives, carrying out a biopsy in all men or no men. The first thing to note is the range of threshold probabilities on the x axis, which has an upper limit of 20% (for illustrative purposes, figure 2 shows the decision curve across all threshold probabilities). We chose an upper limit of 20% because though doctors (taking into account patient preferences) might vary in their values for finding cancer compared with avoiding unnecessary biopsy, it is unrealistic that any doctor or patient would need more than a 20% risk of high grade disease before biopsy is recommended. Thus the initial step in creating a decision curve involves determining a reasonable range of threshold probabilities for the specific decision informed by the marker, test, or model.

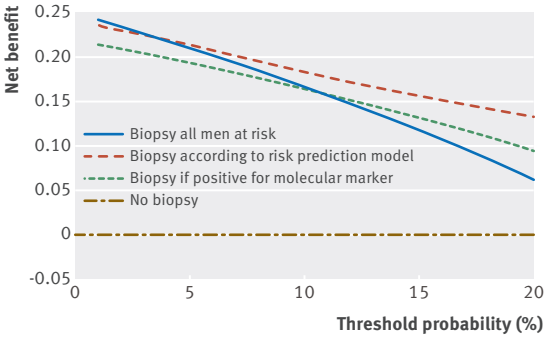

Fig 1 | Decision curve showing net benefit for carrying out biopsy in men at risk for aggressive prostate cancer

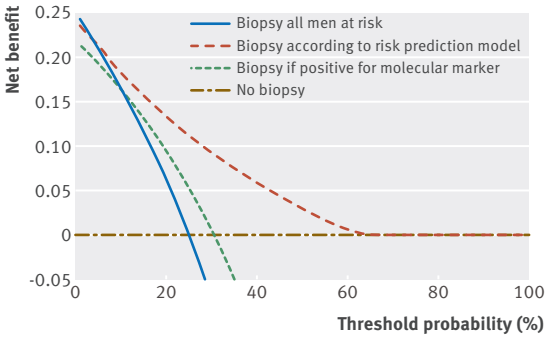

Fig 2 | Decision curve as in figure 1 shown for illustrative purposes across all threshold probabilities

The basic interpretation of a decision curve is that the strategy with the highest net benefit at a particular threshold probability has the highest clinical value. We note that the net benefit for the marker is lower than that for the strategy of "biopsy all" for threshold probabilities below about 11%. For the sort of risk averse doctors or patients who have a low threshold probability, say 5%, this means that the best clinical outcome—in terms of the number of unnecessary biopsies conducted and cancers found—would be achieved by conducting the biopsy irrespective of the marker results. At a threshold of say, 20%, net benefit for the marker is higher than for carrying out a biopsy in everyone, so the optimal clinical strategy would be to carry out a biopsy in only those men positive for the marker. The key point is that the marker is only helpful for a subset of preferences. What we would really like is for the marker to be better than any alternative strategy across a wide range of reasonable preferences. This is exactly what we see for using the statistical model evaluated in the study. Owing to slight miscalibration,[6] the statistical model is worse than just carrying out a biopsy of all men at very low threshold probabilities. If we assume that no urologist would routinely carry out a biopsy in a man with less than a 5% risk of high grade cancer, using the model to determine whether or not biopsy would lead to the best clinical outcomes independent of individual preference.

Compare this conclusion from figure 1 with statistics such as the sensitivity and specificity of the marker (88% and 33%, respectively), the area under the curve or Brier score of the model (0.822 and 0.150), or the calibration plot (see supplementary appendix). It is not at all clear how we could know whether the model's discrimination or calibration was sufficient to justify clinical use.

An example of a decision curve, published in *The BMJ*, concerned external validation of models to predict cardiovascular disease in the United Kingdom.[7] The authors found that the QRISK model, developed on a UK population, had a higher net benefit than the well established Framingham model. Interestingly, net benefit for Framingham at one widely used cut point was zero or negative. This allowed the authors to state that, while QRISK was a "useful model" in the UK population, Framingham "has no clinical benefit" at some

Urologist discussing a range of threshold probabilities to be used for a decision curve in a study of biopsy for prostate cancer

In my own practice, I would not want to do more than 10 biopsies to find one high grade prostate cancer, so my own personal threshold is about 10% on average. But some of my colleagues are a little more aggressive, and I can imagine them biopsying at risks of 6% or 7%. However, I do not think anyone should be biopsied if they have a risk less than 5%. This is about the prevalence of high grade disease in 70 year olds, and it is not as if we are biopsying almost every 70 year old. Also, the infection rate for biopsy is around 4%, and I think your risk of high grade cancer has to be higher than your risk of infection. Now at the upper end, I can imagine that some older patients, or those who do not like medical procedures, might have a higher threshold, something like 15%. But I can't imagine many patients refusing a biopsy if they had more than a 20% risk of high grade cancer: that's about the risk of someone with a PSA of 25 ng/mL, which is normally seen as being pretty much off the charts

widely used thresholds. (The supplementary appendix gives some further examples of decision curves published in the recent literature.)

Choice of threshold probabilities

Determining a reasonable range of threshold probabilities is a critical aspect of net benefit approaches. This is much simpler than it might seem because the use of a range of values means that detailed information about any individual's personal preferences is not needed. All we need are general ideas about what might be considered more or less reasonable. The box gives an example of how a doctor might justify the range of thresholds (5-20%) used in figure 1. Note that we can derive threshold probabilities by thinking in terms of "numbers needed" (that is, how many biopsies would I carry out to find one aggressive cancer?). The high and low ends of the range are justified in terms of typical patients who unquestionably would and would not undergo biopsy, and in terms of the risk of side effects. The key point is that the investigator does not have to find the "correct" threshold, just have an idea of the sort of thresholds that would and would not make sense. Note also that the doctor in this particular case does not consider financial cost, but another might choose to do so (for example, "biopsy is very expensive, so we don't want to do too many to find one cancer").

What happens if doctors do not have any idea about appropriate thresholds, such that an appropriate range cannot be determined? Net benefit is a tool for evaluating the clinical implications of models, markers, and tests. A model gives a predicted probability directly; markers and tests give positive and negative predictive values. If such probabilities cannot be compared against some kind of threshold to aid a decision, then they have no clinical use and the question of clinical evaluation by net benefit is moot.

Another question about threshold probabilities concerns the role of patients. They should generally be involved in decision making about their own care, such as whether or not to have a biopsy for cancer. However, that does not mean that they need to be directly involved in considerations of the appropriate range for threshold probability in a research study. The range of threshold probabilities considered captures differences in patient preferences. In the biopsy example, for instance, the investigator setting the range thinks about patients who "don't like medical procedures." Net benefit may take a clinical perspective and incorporate differences in preferences between individuals, but the research technique gives a result at the population level: should doctors use this model, marker, or test in their practice?

Extensions to net benefit methods

Net benefit methods are flexible and can be adapted to correct for statistical overfit,[8] incorporate time to event data with and without competing risk,[8] and provide estimates in terms of the reduction in unnecessary interventions.[5] Methods have also been published to calculate a 95% confidence interval for net benefit[8] (see supplementary appendix). Net benefit can also incorporate the harm of a test, such as if an invasive or expensive procedure is required.[5] In brief, investigators are asked not only about threshold probabilities in the usual way but also to specify the maximum number of tests they would do to find one true case, assuming that the test was perfect. For instance, if a prediction model for biopsy of the prostate required magnetic resonance imaging, investigators might state that they would do no more than about 30 scans to find one high grade cancer. The test harm for magnetic resonance imaging would then be 0.033 (1/30), and this amount would be subtracted from the net benefit of the prediction model across all threshold probabilities.

Conclusions

Simple decision analytic approaches can provide a clear answer to the question about which of two models, markers, or tests would lead to better clinical outcomes on average among suitable patients, and whether either would be better than a default strategy of treating all patients or none. Net benefit is one such decision analytic technique.

Several explanatory papers give further information about net benefit[9 10 11 12]; simple to use software, along with tutorials and example data, are available at http://decisioncurveanalysis.org/. Several similar measures, including "relative utility" and "weighted net reclassification improvement," have been proposed that provide results consistent with net benefit.[13]

Net benefit approaches assume that doctors and patients will act rationally in accordance with their preferences. In the prostate biopsy example, for instance, we assume that if a doctor and patient engage in shared decision making and decide on a threshold probability of 10%, then the patient will indeed undergo biopsy if the prediction model gives a probability of 10% or more but not if the predicted probability is less than 10%. The real world of actual clinical practice might be somewhat messier, such as if an anxious doctor recommended a biopsy to a patient even though the predicted risk turned out to be low. In some cases it can be useful to complement net benefit with "impact studies" that empirically evaluate the effect of a marker, model, or test on clinical decision making and patient outcomes.[14]

doi: 10.1136/bmj.i6 | *BMJ* 2016;352:i6 | the bmj

Although this decision analytic approach is relatively novel, net benefit is increasingly used in practical applications, including high profile publications such as *The BMJ*.[7] Given that the purpose of statistical analysis is often to help doctors make better decisions, wider use of net benefit, a sound decision analytic technique, will better match the clinical aim of much medical research. We advocate reporting net benefit alongside measures of discrimination and calibration to provide a statistic of immediate clinical interpretability.

Contributors: AJV conceived the paper and wrote the initial draft. EWS and BVC advised on and added new material. All authors edited and approved the final manuscript. AVJ is the guarantor.

Funding: This study was supported in part by funds from David H Koch provided through the Prostate Cancer Foundation, the Sidney Kimmel Center for Prostate and Urologic Cancers, P50-CA92629 SPORE grant from the National Cancer Institute to H Scher (principal investigator of a programme grant on prostate cancer), the P30-CA008748 NIH/NCI Cancer Center Support Grant to Memorial Sloan Kettering Cancer Center, and R01 CA179115 to AJV. Also supported in part by Internal Funds KU Leuven (grant C24/15/037), Research Foundation–Flanders (grant G0B4716N) to BVC, and U award (AA022802, value of personalized risk information), and by a FP7 grant (602150, CENTER-TBI) to EWS.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Steyerberg EW. *A practical approach to development, validation, and updating. Clinical prediction models*. Springer, 2009.
2 Ankerst DP, Hoefler J, Bock S et al. Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low- vs high-grade prostate cancer. *Urology* 2014;83: 1362-7. doi:10.1016/j.urology.2014.02.035. 24862395
3 Liang Y, Ankerst DP, Ketchum NS et al. Prospective evaluation of operating characteristics of prostate cancer detection biomarkers. *J Urol* 2011;185: 104-10. doi:10.1016/j.juro.2010.08.088. 21074193
4 Vickers A, Cronin A, Roobol M et al. Reducing unnecessary biopsy during prostate cancer screening using a four-kallikrein panel: an independent replication. *J Clin Oncol* 2010;28: 2493-8. doi:10.1200/JCO.2009.24.1968. 20421547
5 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26: 565-74. doi:10.1177/0272989X06295361. 17099194
6 Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making* 2015;35: 162-9. doi:10.1177/0272989X14547233. 25155798
7 Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012;344: e4181. doi:10.1136/bmj.e4181 22723603
8 Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak* 2008;8: 53. doi:10.1186/1472-6947-8-53. 19036144
9 Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making* 2008;28: 146-9. doi:10.1177/0272989X07312725. 18263565
10 Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *Am Stat* 2008;62: 314-20. doi:10.1198/000313008X370302. 19132141
11 Steyerberg EW, Vickers AJ, Cook NR et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21: 128-38. doi:10.1097/EDE.0b013e3181c30fb2. 20010215
12 Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2012;42: 216-28. doi:10.1111/j.1365-2362.2011.02562.x. 21726217
13 Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Med Decis Making* 2013;33: 490-501. doi:10.1177/0272989X12470757. 23313931
14 Steyerberg EW, Moons KG, van der Windt DA et al. PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10: e1001381. doi:10.1371/journal.pmed.1001381. 23393430

Appendix: supplementary information