European Radiology (2022) 32:5787–5789
https://doi.org/10.1007/s00330-022-08685-8

EDITORIAL

Decision curve analysis in the evaluation of radiology research

Andrew J. Vickers [1] · Sungmin Woo [2]

Received: 11 January 2022 / Revised: 11 January 2022 / Accepted: 15 February 2022 / Published online: 29 March 2022

The radiology literature consists primarily of diagnostic and prognostic studies. Here, we discuss recent developments in the statistical analysis of such studies. These shift the focus from evaluating the *accuracy* of an imaging test or prediction model to evaluating *outcomes and decision-making*.

As a motivating example, consider the use of magnetic resonance imaging (MRI) for the assessment of parametrial invasion (PMI) in cervical cancer. Suspicion of PMI is important when deciding the primary modality of treatment. If PMI is present, the patient is recommended primary chemoradiation; if not, radical hysterectomy with or without adjuvant therapy is the treatment of choice. MRI is standard of care in evaluation of PMI, but it is not 100% accurate [1]. It has been suggested that MRI results could be combined with clinicopathological variables, such as tumor size or deep stromal invasion, to create a statistical prediction model for the risk of PMI [2]. It might be, for instance, that patients with negative MRI should nonetheless undergo primary chemoradiation if they are at very high risk due to other features, or conversely, avoid such treatment with positive MRI if otherwise low risk.

Assume that two different research groups have proposed statistical models (A and B) and a third research group then independently tests both models on a new data set. Traditionally, the statistical analysis of the data would focus on discrimination and calibration (Fig. 1). Model B has much better discrimination, an area under the curve (AUC) of 0.809 vs. 0.749, but it is difficult to know whether an AUC higher by 0.05 is offset by the poorer calibration of model B compared to model A. Moreover, if we decide that the miscalibration is too much, and that we prefer model A, we are left with the question of whether an AUC of 0.749 is high enough to warrant clinical use of the model. In short, the typical metrics reported by statisticians do not really answer the key questions of whether to use a model and if so, which one.

Decision curve analysis is a statistical technique for the evaluation of tests or models that focuses on decisions and outcomes. We start from the idea that, to know whether the benefits of a model or test outweigh the harms, we must put some numbers on benefit and harm. This is achieved by thinking about the threshold probability of disease, defined as the minimum probability of disease (in this case, PMI) at which a decision-maker—doctor or patient—would opt for an intervention (such as chemoradiotherapy). We call this p_t the threshold probability, a value directly linked to how the consequences of the decision are weighted. Imagine that a patient stated they would opt for primary chemoradiotherapy only if their risk of PMI were 10% or more. A 10% risk of PMI is a 90% chance of no PMI, a 9:1 ratio. Therefore, a patient with a p_t of 10% thinks that the benefits of chemoradiotherapy if they had PMI are worth nine times more than the risks of unnecessary chemoradiotherapy if they did not have PMI.

In decision curve analysis, we estimate the value for each model or test being evaluated across a range of threshold probabilities. The use of a range is to reflect that patients might have different preferences about the relative benefits and harms of aggressive treatment or may differ medically: a healthy young mother might value survival first and foremost and hence have a lower threshold probability, such as 5%; an older patient who has had a previous bad experience with chemotherapy might have a higher threshold probability, such as 25%. Value is calculated as "net benefit." Benefit is "net" because it subtracts harms (false positives, in this case, patients without PMI who receive chemoradiotherapy) from benefits (true positives, in this case, women with PMI who receive chemoradiotherapy) taking into account the different value of benefits and harms. The methods for calculating net benefit have been described in prior methodologic [3] and didactic [4, 5] papers.

✉ Andrew J. Vickers
vickersa@mskcc.org

[1] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave, 2nd Floor, New York, NY 10017, USA

[2] Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

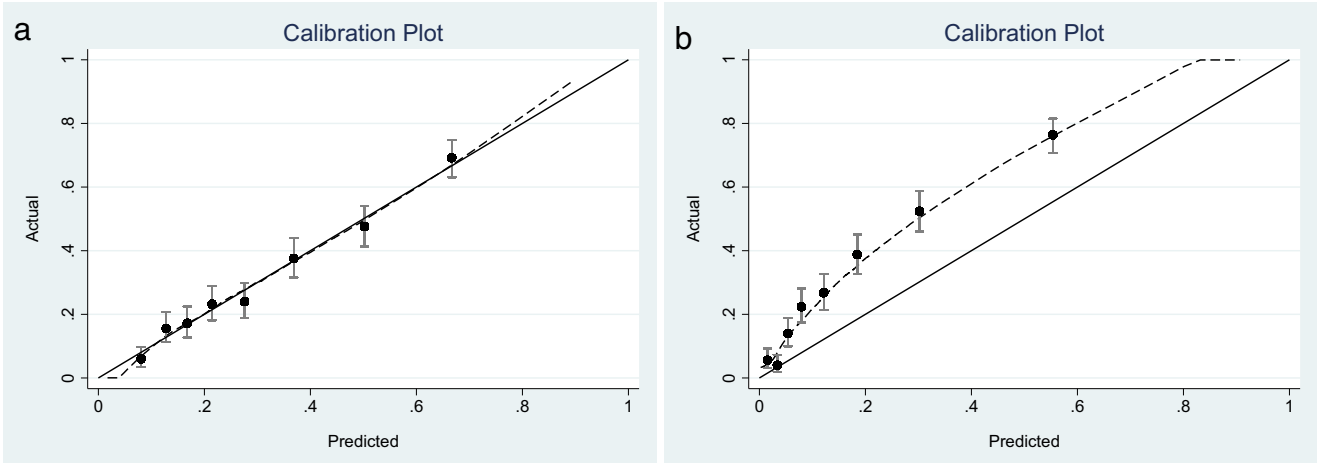

Fig. 1 Calibration plots for model A (left-hand side) and model B (right-hand side). Model B has better discrimination (AUC 0.809 vs. 0.749) but has some miscalibration

Figure 2 shows the decision curve analysis for our study of PMI models. Net benefits of the two models are shown as well as the default approaches of treating all or no women with chemoradiotherapy. Model A has the highest net benefit across the full range of threshold probabilities. This means that we can recommend model A to help the decision of whether to undergo chemotherapy irrespective of patient preferences or medical condition. Of interest, model B sometimes has a lower net benefit compared to a strategy of chemoradiotherapy for all patients: the miscalibration in the model means that patients will be given incorrect risk estimates and would make bad decisions as a result.

It is not always the case that models with good calibration and discrimination calibration are of clinical benefit. As an example, men with PIRADS 4 or 5 on prostate MRI have about a 65% risk of high-grade prostate cancer and so normally receive a biopsy. Imagine some researchers propose that using other features (e.g., prostate-specific antigen levels, prostate volume) might help some men with high PIRADS scores avoid biopsy. The model, shown in Fig. 3, has good calibration and discrimination (0.760). That might tempt us to recommend the use of the model. The decision curve is shown in Fig. 4, with a wide range of threshold probabilities shown for the sake of illustration. The model is of little use clinically because the net benefit is the same as for biopsy all—the current strategy for men with high PIRADS scores—at the sort of lower thresholds that would typical for prostate biopsy. Figure 4 also shows the net benefit for a highly accurate binary marker (sensitivity 90%, specificity 85%). This is to demonstrate that decision curves can be calculated for categorical markers and that such curves may show that even excellent tests may not have clinical benefit in the context of a given clinical scenario.

In sum, metrics of accuracy, such as calibration and discrimination, may be of interest to statisticians, but do not

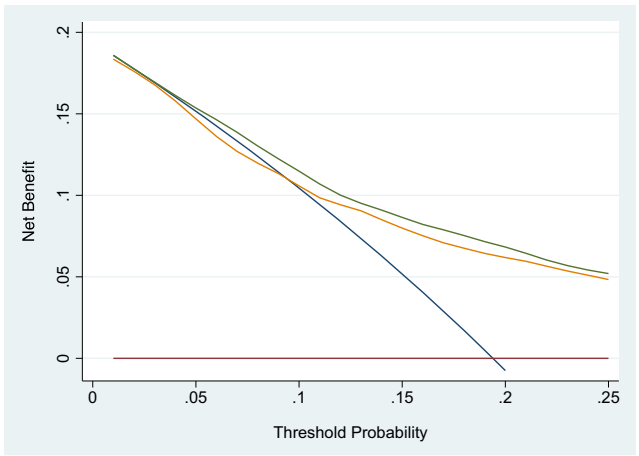

Fig. 2 Decision curve analysis comparing deciding treatment according to model A (green line), model B (orange line), a strategy of treating all patients with chemoradiotherapy (blue line) and no chemoradiotherapy for any patient (red line)

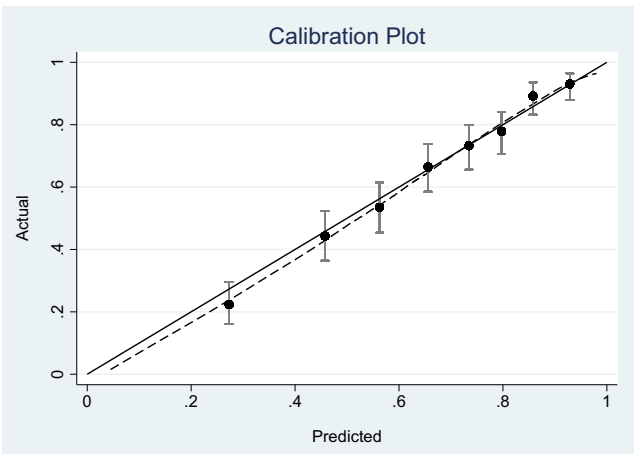

Fig. 3 Calibration plot of a model to predict high-grade cancer on prostate biopsy in men with high PIRADS scores (AUC 0.760)

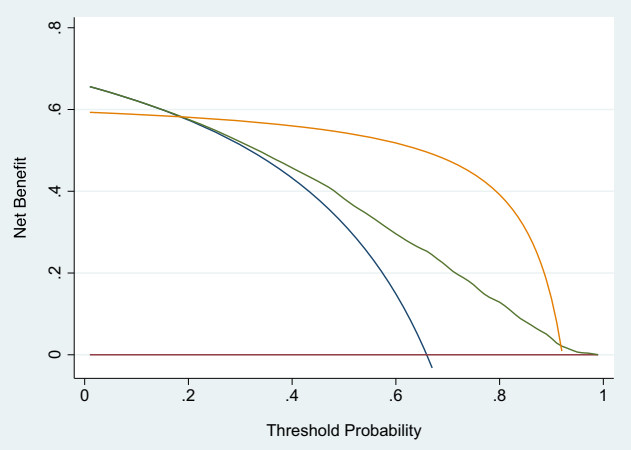

Fig. 4 Decision curve analysis of the prostate cancer model. Net benefit of the model (green line) or accurate binary test (orange line) is not higher than that of the current strategy of biopsying all men with high PIRADS scores (blue line) for the sort of threshold probabilities reasonable for prostate biopsy

evaluate clinical value. Decision curve analysis gives us clear answers as to whether a model (or test) does more good than harm. Increased use of this decision-analytic methodology is recommended for radiology research.

Further reading Links to papers, tutorials, data sets, and code can be found at www.decisioncurveanalysis.org.

Funding This work was supported by a Cancer Center Support Grant to Memorial Sloan Kettering Cancer Center [P30 CA008748]

Declarations

Guarantor The scientific guarantor of this publication is Dr. Andrew Vickers.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and Biometry One of the authors (AV) has significant statistical expertise.

Informed consent Not applicable

Ethics approval Not applicable

Methodology
• Editorial

References

1. Woo S, Atun R, Ward ZJ, Scott AM, Hricak H, Vargas HA (2020) Diagnostic performance of conventional and advanced imaging modalities for assessing newly diagnosed cervical cancer: systematic review and meta-analysis. Eur Radiol 30:5560–5577
2. Li C, Yang S, Hua K (2021) Nomogram predicting parametrial involvement based on the radical hysterectomy specimens in the early-stage cervical cancer. Front Surg 8:759026
3. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26:565–574
4. Vickers AJ, Van Calster B, Steyerberg EW (2016) Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 352:i6
5. Vickers AJ, van Calster B, Steyerberg EW (2019) A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 3:18