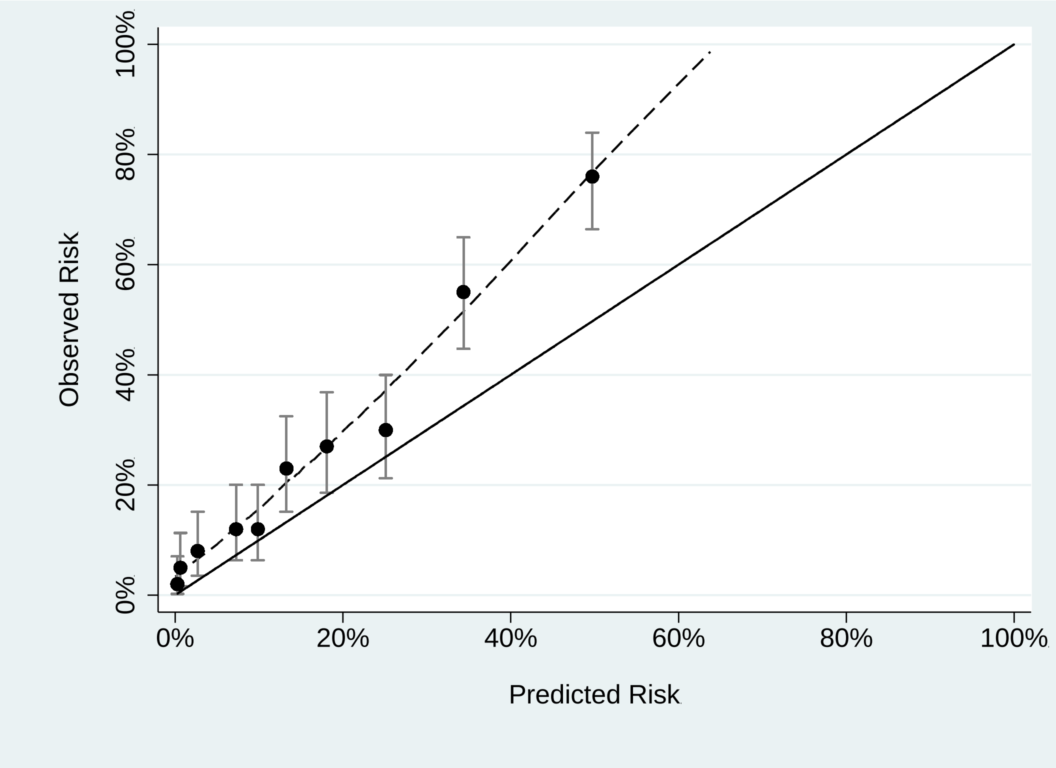
Appendix

Model calibration for the motivating example

The calibration plot for the model given in the paper is shown below. The model underestimates risk, with true risk about 50% higher than predicted risk. For instance, the rate of high-grade cancer is about 15% amongst patients with predicted risk close to 10%. It is not at all clear that a model with this level of miscalibration would be clinically useful even if the AUC, at 0.82, is high. The decision curve is unambiguous that the model would improve clinical decision making.



Examples of the use of decision curves in the medical literature

There are numerous other examples in the literature where decision curve analysis gives a clear answer as to the clinical value of a marker, test or model, where discrimination and calibration are ambiguous. Here we will focus on four cases for illustrative purposes.

Nam et al. evaluated two prediction models for aggressive prostate cancer, the same scenario discussed in the main text of this paper[1]. The "Sunnybrook" prediction model had an AUC of 0.72 compared to 0.67 for the "Prostate Cancer Prevention Trial" model. The Sunnybrook model overestimated risk, although only slightly at lower risks; the Prostate Cancer Prevention Trial model underestimated lower risks and overestimated higher risks. As might be expected, the decision curve shows that of the two models, the Sunnybrook model has higher net benefit. What is not predictable from the AUCs and calibration plots is that neither model has higher net benefit than the strategy of biopsying all men at risk unless threshold probabilities are rather high, well above 10%. Because many doctors and patients would consider biopsy at a risk of aggressive prostate cancer less than 10%, the decision curve does not support clinical use of either model.

Another example from cancer screening is an evaluation of a lung cancer prediction tool[2]. The tool has been proposed to determine which patients should be considered for lung CT screening and was evaluated in three cohorts, although for the sake of simplicity, we will focus on just one, the Harvard cohort. The prediction model had an AUC of 0.76, which is reasonable, although not much better than the AUC of 0.74 for smoking duration. Calibration was problematic, with about 30% more events occurring than predicted by the model. The clinical value of a model with modest discrimination and some miscalibration might well be questionable. However, the decision curve clearly demonstrated that across the complete range of reasonable threshold probabilities, the model had higher net benefit than smoking duration alone, or recommending all patients for screening. The authors concluded that the study gave evidence of "benefits for stratifying patients for lung cancer CT screening."

A study on a model to predict six-month mortality in elderly patients provides an example where a net benefit approach can come to different conclusions about the value of a model depending on its clinical role[3]. The mortality prediction model was found to have higher net benefit than both assuming all patients die within six months and assuming no patient dies within six months across risk thresholds 5% - 25%. The authors concluded that while the model could be of value for determining which patients should be counseled about advance care planning, it should not be used for referral to hospice care. This is on the grounds that whereas a discussion about advance care is indicated for patients with even a relatively low probability of mortality, a hospice referral will generally require more than a 25% risk of death within six months.

Lughezzani et al. provide an example where a decision curve helped choose between two models, one with better discrimination, the other with better calibration[4]. The paper also provides an example where net benefit is used to compare a binary decision rule with a prediction model.

Two high-profile editorials on decision curve analysis have been published in JAMA[5] and the *Annals of Internal Medicine[6].*

Uncertainty of net benefit estimates

One concern is whether estimates of net benefit should include a measure of uncertainty, such as a 95% confidence interval (C.I.). Although bootstrap resampling methods have been published to calculate a 95% C.I. for net benefit[7], these are not widely used. This may be because uncertainty is a problematic concept in decision theory. An argument that is common among decision analysts is that when we are forced to make a choice between limited options – such as biopsying vs. not biopsying vs. measuring a marker and then deciding – we should choose based on our best guess as to the right choice, irrespective of uncertainty[8]. If we predict that outcome will be best if we biopsy, it is irrelevant whether the chance that we are wrong is 1% or 49%. Others might argue that if we are recommending some change in clinical practice, we ought to be pretty sure that it will do more good than harm.

The table shows one approach to incorporating 95% C.I. into net benefit approaches, based on the data in figure 1 for a study with 1000 patients. At some lower threshold probabilities, the lower bound of the 95% C.I. does include no net benefit of the model compared to the strategy of biopsying all men at risk.

Table. An example of how statistical uncertainty can be incorporated into decision curve analysis. As the model clearly has better net benefit than the both "biopsy none" and "marker", the table shows the 95% C.I. for the difference in net benefit between the model and "biopsy all".

Threshold probability	**Difference in net benefit between model and biopsy all**	**95% C.I.**
5%	0.002	-0.005 to 0.008
7.5%	0.006	-0.003 to 0.015
10%	0.014	0.001 to 0.025
12.5%	0.024	0.009 to 0.038
15%	0.036	0.018 to 0.052
20%	0.063	0.041 to 0.086

References

1. Nam RK, Kattan MW, Chin JL, et al. Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2011;**29**(22):2959-64 doi: 10.1200/jco.2010.32.6371.
2. Raji OY, Duffy SW, Agbaje OF, et al. Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study. Annals of internal medicine 2012;**157**(4):242-50 doi: 10.7326/0003-4819-157-4-201208210-00004.
3. Duarte CW, Black AW, Murray K, et al. Validation of the Patient-Reported Outcome Mortality Prediction Tool (PROMPT). Journal of pain and symptom management 2015;**50**(2):241-7 e6 doi: 10.1016/j.jpainsymman.2015.02.028.
4. Lughezzani G, Zorn KC, Budaus L, et al. Comparison of three different tools for prediction of seminal vesicle invasion at radical prostatectomy. European urology 2012;**62**(4):590-6 doi: 10.1016/j.eururo.2012.04.022.
5. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015;**313**(4):409-10 doi: 10.1001/jama.2015.37.
6. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Annals of internal medicine 2012;**157**(4):294-5 doi: 10.7326/0003-4819-157-4-201208210-00014.
7. Vickers AJ, Cronin AM, Elkin EB, et al. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC medical informatics and decision making 2008;**8**:53 doi: 10.1186/1472-6947-8-53.
8. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. Journal of health economics 1999;**18**(3):341-64