VOLUME 34 · NUMBER 21 · JULY 20, 2016

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use

Kathleen F. Kerr, Marshall D. Brown, Kehao Zhu, and Holly Janes

Kathleen F. Kerr and Kehao Zhu, University of Washington; and Marshall D. Brown and Holly Janes, Fred Hutchinson Cancer Research Center, Seattle, WA

Published online ahead of print at www.jco.org on May 31, 2016.

Supported by Grants No. 5R01HL085757-09 (K.F.K.) and R01CA152089 (H.J.) from the National Institutes of Health.

Authors' disclosures of potential conflicts of interest are found in the article online at www.jco.org. Author contributions are found at the end of this article.

Corresponding author: Kathleen F. Kerr, PhD, Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195; e-mail: katiek@uw.edu.

0732-183X/16/3421w-2534w/$20.00

DOI: 10.1200/JCO.2015.65.5654

ABSTRACT

The decision curve is a graphical summary recently proposed for assessing the potential clinical impact of risk prediction biomarkers or risk models for recommending treatment or intervention. It was applied recently in an article in *Journal of Clinical Oncology* to measure the impact of using a genomic risk model for deciding on adjuvant radiation therapy for prostate cancer treated with radical prostatectomy. We illustrate the use of decision curves for evaluating clinical- and biomarker-based models for predicting a man's risk of prostate cancer, which could be used to guide the decision to biopsy. Decision curves are grounded in a decision-theoretical framework that accounts for both the benefits of intervention and the costs of intervention to a patient who cannot benefit. Decision curves are thus an improvement over purely mathematical measures of performance such as the area under the receiver operating characteristic curve. However, there are challenges in using and interpreting decision curves appropriately. We caution that decision curves cannot be used to identify the optimal risk threshold for recommending intervention. We discuss the use of decision curves for miscalibrated risk models. Finally, we emphasize that a decision curve shows the performance of a risk model in a population in which every patient has the same expected benefit and cost of intervention. If every patient has a personal benefit and cost, then the curves are not useful. If subpopulations have different benefits and costs, subpopulation-specific decision curves should be used. As a companion to this article, we released an R software package called DecisionCurve for making decision curves and related graphics.

J Clin Oncol 34:2534-2540. © 2016 by American Society of Clinical Oncology

INTRODUCTION

Decision curves[1] are novel and clever graphical devices for assessing the potential population impact of adopting a risk prediction instrument into clinical practice. First proposed in 2006, decision curves have been used in cancer research[2-6] and many other fields.[7-11] For example, decision curves have evaluated models that predict lung cancer in high-risk populations which can then be used as tools for recommending computed tomography (CT) screening.[12] Decision curves were used to assess the utility of a genomic risk model for recommending adjuvant radiation therapy for surgically treated prostate cancer.[6] This article summarizes decision curves and their appropriate use and discusses some ways they can be misunderstood or misconstrued. We also discuss some subtle but important issues that arise when individuals in a population have different costs or benefits from a medical intervention recommended on the basis of predicted risk. We begin by reviewing the concept of risk that is fundamental to decision curves.

THE FUNDAMENTAL CONCEPT OF RISK

A risk prediction instrument is a statistical model or algorithm that calculates an individual's risk of an undesirable outcome (D) in the absence of intervention using the individual's data on biomarkers and/or other characteristics. If the risk is high, the individual may be recommended to undergo some intervention or treatment. As an example from a disease screening context, D might be the outcome of prostate cancer, and the intervention is surgical biopsy. In a therapeutic context, D might be prostate cancer mortality, and the intervention might be radical prostatectomy. In both examples, a model predicting an

individual's risk of D might be used to help guide use of the intervention.

Note however, that an individual's risk is not a single entity. Suppose the prevalence of D in a population (population Q) is 1%. Without any additional information, the only valid risk model assigns every individual a risk of 1%. However, suppose a prognostic biomarker X is available (Fig 1A). If X is measured, we can calculate risks on the basis of X: risk(X) ≡ P(D|X). In this example, if a person's biomarker X is measured as 1, then his risk based on X is risk(X = 1) ≡ P(D|X = 1) = 1.6%.

The common terminology of personal risk has the potential to be misconstrued. Risk is personal in the sense that it depends on an individual's personal measurements. However, risk is not personal in another sense because risk changes given different sets of predictors. Moreover, risk depends on the population in which it is calculated, as we illustrate by extending the example in Figure 1A. Suppose we have a second population (population W) slightly different from population Q. In population W, the biomarker X has exactly the same distribution (Fig 1A). The only difference between populations Q and W is that D is much more common in population W, with 10% prevalence. If a member of population W has biomarker X = 1, then his personal risk is risk(X = 1) = 15.5%. In other words, two individuals with the same biomarker measurement have different personal risks because risks are calculated in the context of each individual's population. This simple example shows that population characteristics, such as the prevalence of D, play a role in personal risk assessment.

A way to avoid overinterpreting personal risk is to remember that an individual has multiple risks, depending on which predictors are chosen as the basis for calculating risk. For example, a woman's 5-year risk of breast cancer can be estimated solely on the basis of her age, solely on the basis of her genetics, or on the basis of both. Each set of risk factors can give a different risk, yet all of these risks are well defined. The fact that a single individual has multiple valid estimates of risk reminds us that personal risk is actually an assessment of the population frequency of D among people who share the same values of the risk factors.

We next describe how decision curves can be used to assess the clinical value of a risk model and to compare different risk models.

DECISION CURVES: A TOOL FOR ASSESSING THE POPULATION PERFORMANCE OF A RISK MODEL FOR TREATMENT RECOMMENDATIONS

The context for decision curves is a situation in which individuals' risks for an outcome D will be assessed, and individuals with sufficiently high risk will be recommended some intervention or treatment. The assumption is that the intervention carries some expected benefit to those who have (or will experience) D in the absence of intervention (cases) and carries some cost or harm to those who do not have (or will not experience) D (controls).

The concept of population net benefit (NB)[13] is fundamental to decision curves. Suppose that the intervention carries some

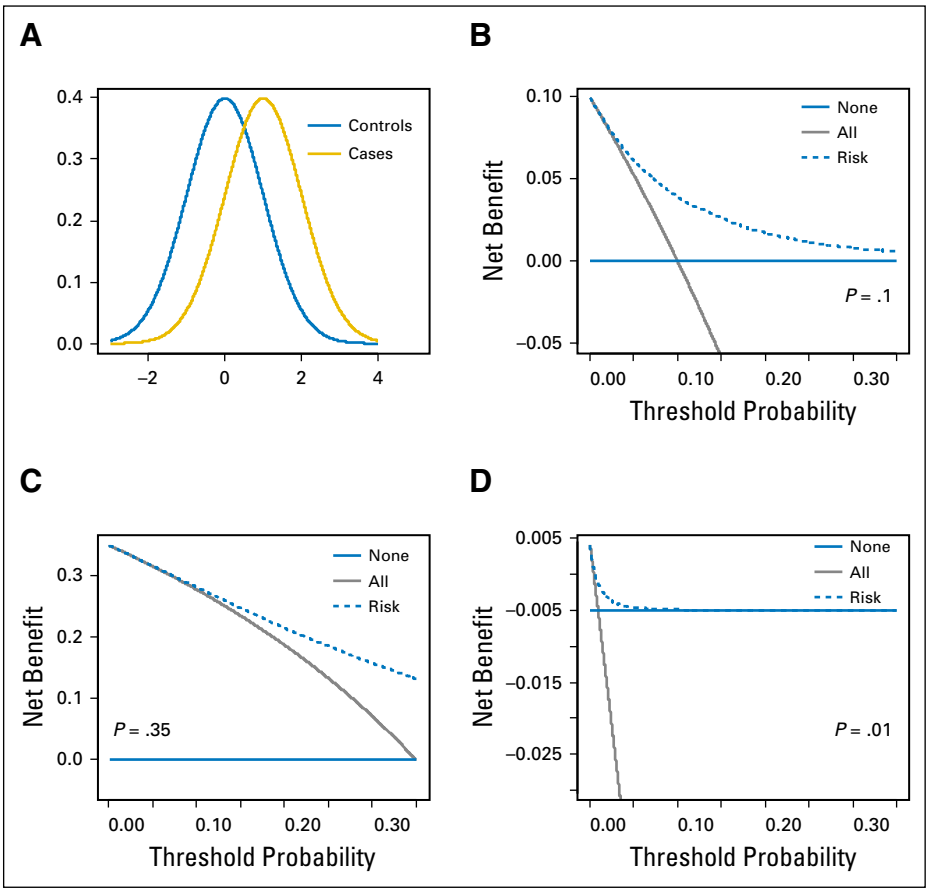

Fig 1. Behavior of decision curves as a function of outcome prevalence. (A) The distribution of a biomarker X in cases and controls. Decision curves for risks based on biomarker X are shown for prevalence (B) 10%, (C) 35%, and (D) 1%.

benefit B > 0 to cases and does not benefit controls but rather carries some harm or cost C > 0. As we will see, decision curves do not require specifying values for B and C but implicitly depend on their existence. Intuitively, if a risk model tends to identify cases as high risk without falsely identifying too many controls as high risk, then the NB of the risk model to the population will be positive. The NB depends on the benefit B, the cost C, the prevalence P of the outcome, and the ability of the risk model to assign high risks to cases and low risks to controls (ie, the model's classification accuracy given a definition of high risk). Suppose high risk is defined as risk above some risk threshold R; such high-risk patients are recommended an intervention. The model's classification accuracy is measured by the true-positive rate (TPR_R), the proportion of cases with risk above risk threshold R; the false-positive rate (FPR_R) is the proportion of controls with risk above risk threshold R. The NB to the population of using the risk model together with high-risk threshold R is:

$$NB_R = TPR_R \; P - \frac{R}{1-R} FPR_R \; (1-P) (**)$$

There are a few important observations about this expression. First, the benefits and costs of the intervention, B and C, do not appear explicitly in (**) but they are included implicitly. This is because the derivation of this expression (Appendix, online only) assumes that the risk threshold R has been chosen rationally, reflecting the costs and benefits of intervention. A classic result in decision theory says that such a rationally chosen risk threshold is a function of B and C: $R = \frac{C}{B+C}$.[1] Equivalently, $\frac{R}{1-R} = \frac{C}{B}$. Additional assumptions embedded in (**) are that every case has the same expected benefit B of intervention, independent of predicted risk, and every control has the same expected cost C, independent of predicted risk. We revisit these assumptions later in this article.

A decision curve is a plot of NB as given in expression (**) versus the risk threshold R used to recommend intervention. Figure 1B shows a decision curve for a risk model that uses a marker X (Fig 1A) in a population with prevalence 10%. The horizontal axis is the risk threshold R used to define high risk; the vertical axis is NB as given at (**). The dotted line shows NB for the risk model risk(X). The horizontal line at NB = 0 allows one to compare the NB of using risk(X) to recommend intervention compared with a policy of no intervention in the population (treat none). For the treat none policy, no cases receive benefit B and no controls experience cost C, so the NB of implementing this policy in the population is 0. The gray curve in the plot depicts the NB of recommending the intervention to everyone in the population regardless of risk. Under this policy, TPR = FPR = 1, so $NB = P - \frac{R}{1-R}(1-P)$. Although treat none and treat all are both policies that do not use individual risks and thus do not use a risk threshold, R is still used in calculating the NB of these policies to capture the relative size of B and C.

Decision Curve Operating Characteristics

We can gain some intuition about the behavior of decision curves by varying the outcome prevalence P. Returning to the example of the biomarker in Figure 1A, Figures 1B to 1D show decision curves for the same marker X when P is varied. When prevalence is high (P = 35%; Fig 1C) the decision curve for risk(X) tends to be well above the treat none flat line but may not dominate the treat all curve. It makes sense that a treat all strategy can work well for common outcomes unless the cost of intervention for controls is high. Conversely, when prevalence is low (P = 1%; Fig 1D), the treat all strategy intuitively becomes a poor strategy because most individuals receiving the intervention are controls and experience cost rather than benefit. In line with this intuition, Figure 1D shows that the NB for treat all is negative except for small Rs.

One can show analytically that the risk thresholds at which the NB curves cross are dictated by the prevalence. By equating the NB for treat all and treat none, the expression $P - \frac{R}{1-R}(1-P) = 0$ shows that the decision curves for the treat all and treat none policies cross at R = P.[1] Figures 1B to 1D illustrate this: the treat all and treat none curves cross at R = 10%, 35%, and 0.02%, respectively.

Interpreting NB

A challenge in interpreting decision curves stems from the challenge in interpreting NB itself. A specific difficulty is that the NB is in units of B.

Mathematically, the maximum possible value of NB is achieved when TPR = 1 and FPR = 0; that is, we can never do better than intervening on all cases and no controls. Expression (**) shows that the maximum NB is P. This motivates the metric standardized NB, sNB ≡ NB/P, also known as the relative utility[14-16]:

$$sNB_R = TPR_R - \frac{R}{1-R} \frac{1-P}{P} FPR_R$$

We posit that sNB is slightly easier to interpret than NB. One reason is that sNB always has a maximum value of 1.0, providing a sense of large and small on a percent scale. For the example in Figure 1B and R = 6%, the NB for using the risk model is 0.055. By incorporating the prevalence P = .1, we have sNB = 0.055/0.1 = 55%. From this sNB, we see that the risk model achieves a bit more than half the maximum possible achievable utility. Furthermore, we can say that the risk model offers the same sNB to the population as a policy that resulted in intervention for 55% of cases and no controls. For comparison, sNB for the treat all policy is 0.043/0.1 = 43% and thus offers the same sNB to the population as intervention for 43% of cases and no controls.

APPROPRIATE USE OF DECISION CURVES

Suppose there is consensus on the appropriate risk threshold. For example, suppose there is consensus that an individual with risk greater than 20% for having a cardiovascular event in the next 5 years should be prescribed statins. In this setting, one can evaluate the NB of risk model(s) to assess the population impact of using predicted risks to recommend intervention. A decision curve is not needed because there is consensus on the risk threshold; the NB of interest is the point on the curve at R = 0.20. Decision curves are most useful when there is no such consensus, because the curves allow one to examine risk model performance across a range of

plausible risk thresholds. Equivalently, the value of decision curves is to examine risk models across a range of plausible cost-benefit ratios. If the decision curve for one risk model dominates the curve for another, the relative performance of the two models is clear, even if there is disagreement on the precise appropriate risk threshold. Conversely, if the two curves cross, the preferred model depends on the choice of risk threshold or, equivalently, on the cost-benefit ratio.

Decision Curves Cannot Be Used to Choose a Risk Threshold

There is a strong temptation to use decision curves to choose a risk threshold to maximize NB. Because a decision curve is typically decreasing, this approach tends to lead to low risk thresholds. However, this approach is incorrect: the risk threshold R must be selected from other considerations and then used to evaluate the relative merits of policies. To make this explicit, suppose we had data on a population that included only individuals' risks and true D status absent intervention. Such data are enough to construct a decision curve. However, such data include no information about the benefit and cost of intervention, which are properties of the clinical context and are not specific to any risk model. Data that contain no information on benefits and costs of intervention clearly cannot inform choice of a risk threshold to maximize NB.

A different misconception is to examine the decision curve plot to identify areas in which there is a large difference between the risk model and other policies such as treat all and treat none. There is then a suggestion that high-risk individuals are those whose risks are in this region of the plot.[4] This erroneous approach tends to lead to high risk thresholds, at least when a single risk model is compared with treat all and treat none.

In summary, a decision curve is best interpreted by reading vertically: given a chosen risk threshold, the curve displays the NB of using the risk model with that risk threshold, assuming that the risk threshold accurately summarizes the costs and benefits of intervention.

Decision Curves, Rational Decision Making, and Uncalibrated Risk Models

Suppose a risk model is not calibrated in a population, meaning that P(D|X) differs from the model-predicted risk(X). For example, a risk model may systematically overestimate or underestimate individuals' risks. Does it make sense to evaluate the NB of a miscalibrated risk model? Is it appropriate?

The answer to the first question is "yes": the NB of a policy using the miscalibrated risk model is well defined and interpretable. For example, Lughezzani et al[17] used decision curves to compare the population NB of candidate risk prediction instruments for decision making in prostate cancer. One of the models was miscalibrated in the study population but was included as-is to reflect the way it is used in clinical practice. The decision curve for this model is interpretable as describing the NB of using the (miscalibrated) model.

Is it appropriate to use a decision curve with a miscalibrated risk model? Although NB—and thus the decision curve—for a miscalibrated risk model is interpretable, we argue that it is not rationally consistent to choose a risk threshold R on the basis of the benefits and costs of intervention but then to make decisions on the basis of a miscalibrated risk model. To be concrete, suppose an assessment of costs and benefits leads to R = 6%. Suppose the risk model systematically overestimates risk, so those assigned risks of 6% actually have an event rate of 3% and those assigned risks of 12% actually have an event rate of 6%. Individuals assigned risks between 6% and 12% will be recommended the intervention using risk threshold R = 6%. However, their event rate will be less than 6%, and they should not be recommended the intervention because their expected NB from the intervention is negative.

It is known theoretically[18] and has been demonstrated empirically[19] that miscalibration reduces NB. In other words, a calibrated risk model has equal or greater NB compared with a miscalibrated version. The simple example in the previous paragraph gives some reasons for why this is so. If risks are overestimated, too many controls will experience the costs of intervention. If risks are underestimated, too few cases will experience the benefits of intervention.

In summary, the NB, and thus the decision curve, of a miscalibrated risk model is well defined and interpretable. However, if one aspires to rational decision making and chooses a risk threshold R accordingly, one should not use R with a miscalibrated model. Moreover, miscalibration can only diminish NB, so calibrated models are preferred from a strictly practical point of view.

Variation in Risk Threshold

It has been suggested that decision curves are useful when patients differ in their expected costs and/or benefits of intervention.[1,20] For example, if the intervention is surgery, older patients may be more likely to suffer from adverse surgical outcomes (such as death) and thus have a higher C:B ratio than younger patients. A specific suggestion has been to compare the decision curve at different risk thresholds on the horizontal axis to examine NB for subpopulations with different costs or benefits. However, when a population can be divided into subpopulations with subpopulation-specific risk thresholds, the best course of action is a decision curve analysis performed separately for each subpopulation. It can be misleading in this context to use a decision curve computed for the whole population to assess risk model performance in subpopulations, as the next example illustrates.

Consider the example in Figure 2. Two risk models have nearly identical decision curves for the population POP (Fig 2A, solid curves), with risk model I slightly higher than risk model II for most risk thresholds. From this, one might conclude that the two risk models offer about the same utility to the population, and this is true if the whole population shares the same risk threshold. However, suppose the population is composed of two subpopulations, S1 and S2. For model I, S1 and S2 have the same distributions of predicted risks (Fig 2B). Therefore, the decision curve I showing NB of model I in each of S1 and S2 is the same as the decision curve of model I for the whole population. We can, in fact, use the population decision curve to evaluate the utility of risk model I for both S1 and S2. For example, if $R_1 = 0.4$ for S1 and $R_2 = 0.6$ for S2, then we can use the population decision curve to note that the NB of risk model I for S1 is 0.202 and the NB of risk model I for S2 is 0.140.

In contrast, for model II, S1 and S2 have different distributions of predicted risks (Fig 2C). The population decision curve at

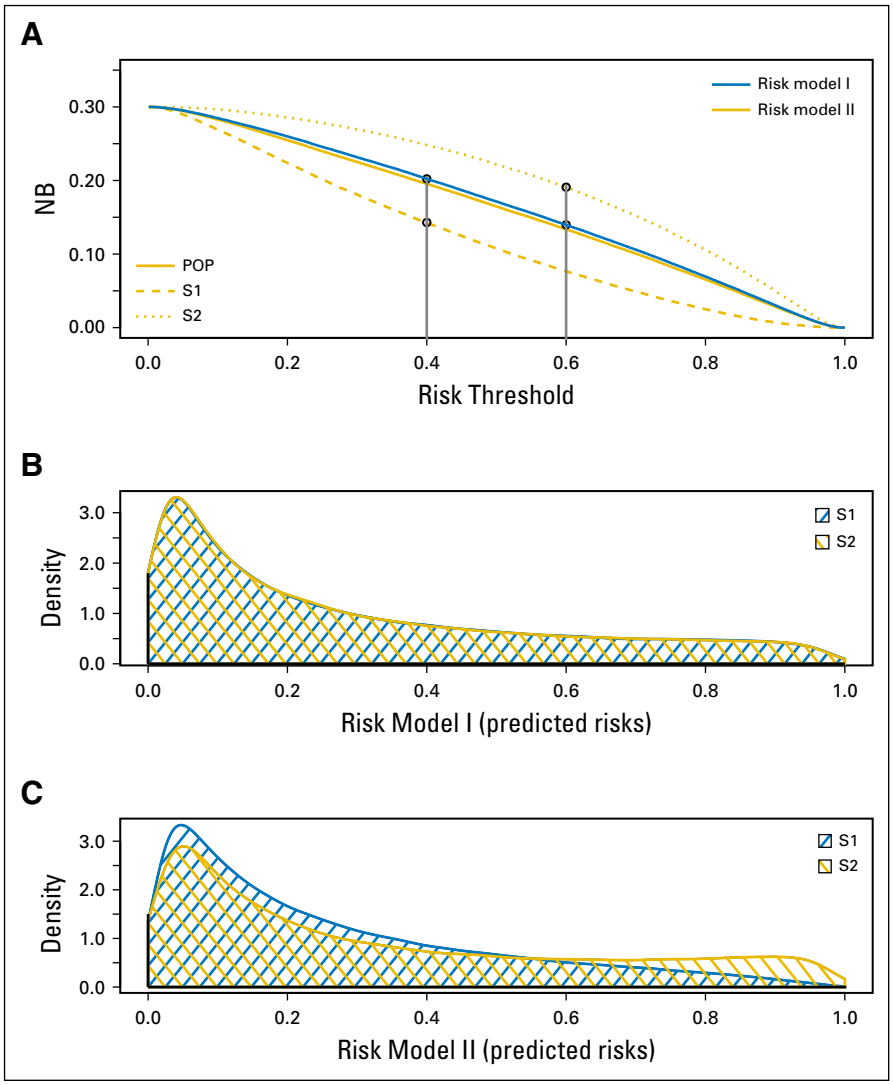

Fig 2. The performance of risk models in a population (POP), which comprises two subpopulations S1 and S2. (A) Solid curves give the decision curves for risk models I and II in POP. Dashed and dotted lines give the decision curves for model II in subpopulations S1 and S2; model I has the same performance in both subpopulations. The population decision curves show that the two risk models are similar in the population. However, for subpopulation S2, the subpopulation net benefit (NB) is higher for risk model II (dotted line). For subpopulation S1, NB is higher for risk model I (dashed line). (B) For risk model I, the distribution of predicted risks is the same for S1 and S2. Therefore, if S1 and S2 have different risk thresholds, the subpopulation NB for risk model I is the same as the NB in POP. (C) For risk model II, the distributions of predicted risks differ for S1 and S2. This explains the divergence of the population and subpopulation-specific decision curves for risk model II in (A).

$R_1 = 0.4$ is 0.196 but the NB of model II in S1 is really 0.143. The population decision curve at $R_2 = 0.6$ is 0.134 but the NB of model II in S2 is really 0.191. The population decision curve does not show subpopulation-specific NB.

On the basis of the population decision curves shown in Figure 2, models I and II had similar performance. But on the basis of the subpopulation-specific decision curves, we see that the NB of using risk model II in S2 is higher than the NB of using model I, and the NB of using risk model II in S1 is lower than the NB of using model I. Therefore, better intervention policies can be formed for S1 using model I; better policies can be formed for S2 on the basis of model II. This would not be apparent by examining only the population decision curves for models I and II in POP.

In summary, interpreting points on decision curves at different risk thresholds as representing NB in different subpopulations is valid only if the distribution of predicted risks among cases and controls in each subpopulation is the same as in the whole population, and the prevalence of D is constant across subpopulations. These are strong conditions; if subpopulations have different risk thresholds, it seems likely they also have different distributions of risk.

ILLUSTRATION AND SOFTWARE

We have developed software to estimate and interpret decision curves, an R package called "DecisionCurve" (https://cran.r-project.org/web/packages/DecisionCurve/). Our software has several features not available in existing software, which we illustrate on simulated data. The simulated data mimic clinical and biomarker data on men age 40 to 89 years, all of whom have prostate cancer status (D) determined by prostate biopsy. Suppose we are interested in assessing the performance of a risk model that uses clinical predictors and an expanded model that additionally uses two biomarkers to identify men at high risk of prostate cancer who should be sent for definitive diagnosis with biopsy. Figure 3A shows the basic plot of model performance for the two risk models. There are two notable differences between Figure 3A and the more typical decision curve plots presented in Figure 1. First, the vertical axis in Figure 3A is sNB rather than NB. As discussed in this article, we find sNB slightly easier to interpret; our software allows for a choice of either NB or sNB. Second, the horizontal axis in Figure 3A is labeled in terms of both risk threshold and cost:benefit

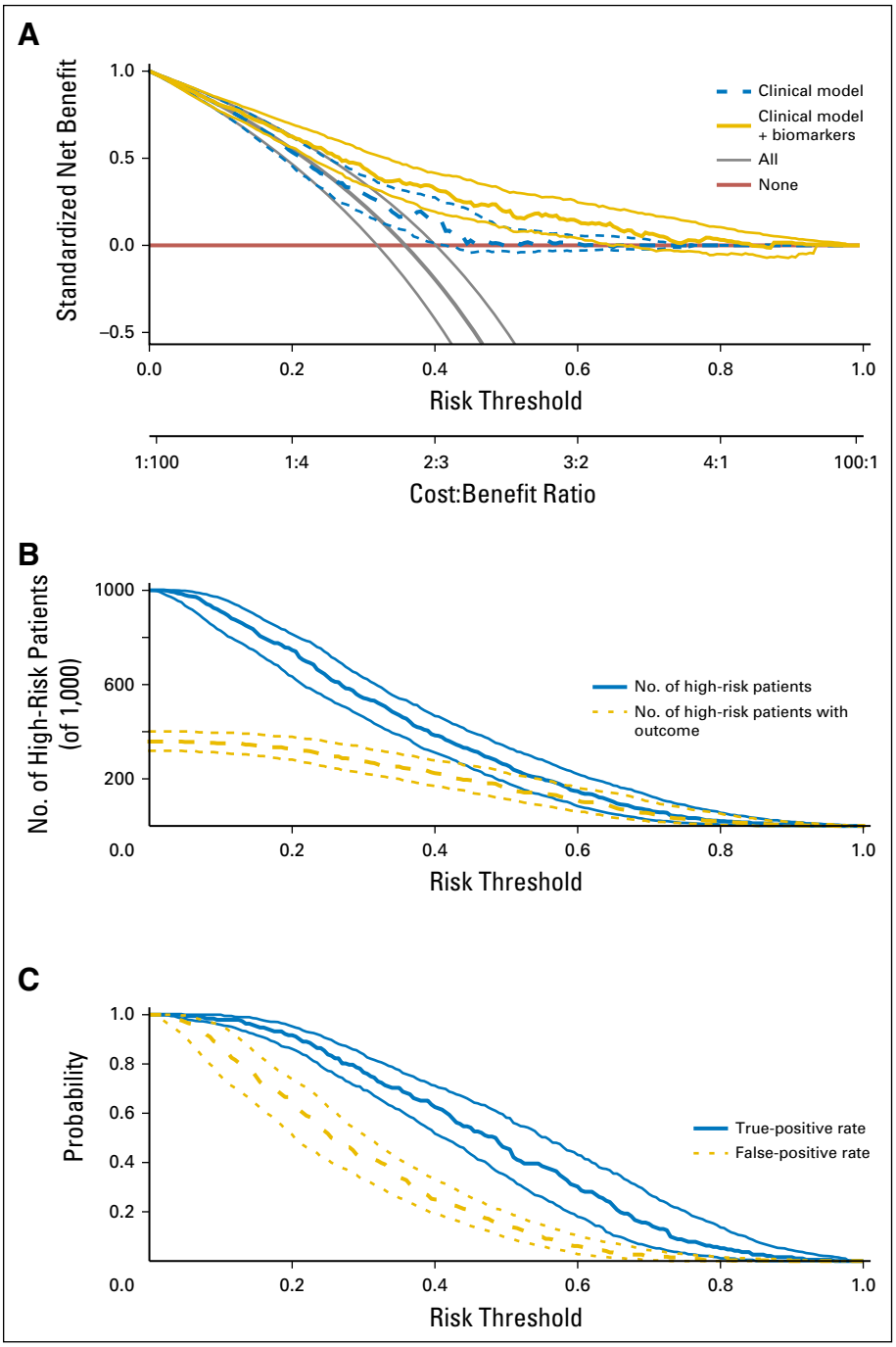

Fig 3. (A) Decision curves for two risk models for prostate cancer produced with DecisionCurve software. The vertical axis displays standardized net benefit. The two horizontal axes show the correspondence between risk threshold and cost:benefit ratio. (B) Clinical impact curve for the biomarker-based risk model. Of 1,000 patients, the heavy blue solid line shows the total number who would be deemed high risk for each risk threshold. The gold dashed line shows how many of those would be true positives (cases). (C) True- and false-positive rates as functions of the risk threshold, for the biomarker-based risk model. The figure shows information similar to that of a receiver operating characteristic curve and also shows the risk threshold corresponding to each true- and false-positive rate. Bands on all plots represent pointwise 95% CIs constructed via bootstrapping.

ratio. The two axes remind readers of the correspondence between risk threshold and cost:benefit ratio. Moreover, for the treat all and treat none policies, cost:benefit ratio is a more natural axis because those policies do not use a risk threshold.

Figure 3B is another type of plot produced by DecisionCurve that we call a clinical impact plot. For a single-risk model, Figure 3B shows the estimated number who would be declared high risk for each risk threshold and visually shows the proportion of those who are cases (true positives). In this example, if a 20% risk threshold were used, then of 1,000 men screened, about 746 would be deemed high risk and sent for biopsy, with about 328 of these being true prostate cancer cases. Similar plots have been used in the literature.[21] Finally, Figure 3C shows the constituents of NB: the true- and false-positive rates. Figure 3C is an alternative to a traditional receiver operating characteristic (ROC) curve for a risk model. It is more informative than an ROC curve because the true- and false-positive fractions are displayed as functions of the risk threshold, whereas the risk threshold is suppressed in the ROC plot.

DecisionCurve uses the empirical distribution of risks to estimate true- and false-positive rates. It includes methodology for constructing CIs via bootstrapping, which are displayed on all of the plots in Figure 3. The user has the option of using cross-validation instead of

bootstrapping. The package also implements methodology for constructing decision curves by using data from case-control studies.

CONCLUDING REMARKS

In conclusion, decision curves, related metrics, and graphical devices are tools for evaluating risk prediction models in a clinical context. We endorse the move away from a purely mathematical assessment of risk model performance, such as assessing the area under the ROC curve.[16] At the same time, we caution against misinterpreting decision curves.

In the current era in which the personalization of medicine is championed, one can imagine that every patient has his own benefit and cost and thus his own personal risk threshold for choosing intervention. However, we caution against the notion that decision curves are useful for individuals to use along with their personal risk thresholds. Decision curves are useful for examining the impact of treatment policies in populations, but they cannot teach us about individual benefits of risk model–guided intervention. The curves display population NB, which is not a well-defined concept for individuals.[22]

In many clinical contexts, it is reasonable to consider that different subpopulations have distinct risk thresholds. In this setting, a stratified decision curve analysis is required to examine NB in subpopulations. A stratified analysis consists of calculating decision curves separately for each subpopulation, as in Figure 2. Stratified analysis properly accounts for subpopulation-specific distributions of risks.

We have largely discussed decision curves for recommending intervention in a context in which no intervention is the standard of care. An analogous situation is one in which intervention is the standard of care, and foregoing that intervention would be considered the exceptional course of action. Decision curves, and our software, can be used in either context. The difference in context manifests in the relevant comparison curve: it is the treat all curve when intervention is standard and the treat none line when no intervention is standard. The points we make regarding proper use and interpretation of decision curves apply to both contexts.

We provide free and publically available software for making decision curves and related graphical devices. Our software offers greater flexibility than existing software and includes features that facilitate correct interpretation of results.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at www.jco.org.

AUTHOR CONTRIBUTIONS

Conception and Design: Kathleen F. Kerr, Holly Janes
Data Analysis: Kathleen F. Kerr, Marshall D. Brown, Kehao Zhu
Other: Marshall D. Brown [Software development]
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Vickers AJ, Elkin EB: Decision curve analysis: A novel method for evaluating prediction models. Med Decis Making 26:565-574, 2006
2. Secin FP, Bianco FJ, Cronin A, et al: Is it necessary to remove the seminal vesicles completely at radical prostatectomy? Decision curve analysis of European Society of Urologic Oncology criteria. J Urol 181:609-613, 2009; discussion 614
3. Ang SF, Ng ES, Li H, et al: Correction: The Singapore Liver Cancer Recurrence (SLICER) score for relapse prediction in patients with surgically resected hepatocellular carcinoma. PLoS One 10: e0128058, 2015
4. Li J, Liu Y, Yan Z, et al: A nomogram predicting pulmonary metastasis of hepatocellular carcinoma following partial hepatectomy. Br J Cancer 110: 1110-1117, 2014
5. Scattoni V, Lazzeri M, Lughezzani G, et al: Head-to-head comparison of prostate health index and urinary PCA3 for predicting cancer at initial or repeat biopsy. J Urol 190:496-501, 2013
6. Den RB, Yousefi K, Trabulsi EJ, et al: Genomic classifier identifies men with adverse pathology after radical prostatectomy who benefit from adjuvant radiation therapy. J Clin Oncol 33:944-951, 2015
7. Labidi M, Lavoie P, Lapointe G, et al: Predicting success of endoscopic third ventriculostomy: Validation of the ETV Success Score in a mixed population of adult and pediatric patients. J Neurosurg 123:1447-1455, 2015
8. McMahon PJ, Panczykowski DM, Yue JK, et al: Measurement of the glial fibrillary acidic protein and its breakdown products GFAP-BDP biomarker for the detection of traumatic brain injury compared to computed tomography and magnetic resonance imaging. J Neurotrauma 32:527-533, 2015
9. Matignon M, Ding R, Dadhania DM, et al: Urinary cell mRNA profiles and differential diagnosis of acute kidney graft dysfunction. J Am Soc Nephrol 25:1586-1597, 2014
10. Slankamenac K, Beck-Schimmer B, Breitenstein S, et al: Novel prediction score including pre- and intraoperative parameters best predicts acute kidney injury after liver surgery. World J Surg 37: 2618-2628, 2013
11. Baart AM, de Kort WL, Moons KG, et al: Zinc protoporphyrin levels have added value in the prediction of low hemoglobin deferral in whole blood donors. Transfusion 53:1661-1669, 2013
12. 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. Ann Intern Med 157:242-250, 2012
13. Peirce CS: The numerical measure of the success of predictions. Science 4:453-454, 1884
14. Baker SG: Putting risk prediction in perspective: Relative utility curves. J Natl Cancer Inst 101: 1538-1542, 2009
15. Baker SG, Cook NR, Vickers A, et al: Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc 172:729-748, 2009
16. Baker SG, Kramer BS: Evaluating prognostic markers using relative utility curves and test trade-offs. J Clin Oncol 33:2578-2580, 2015
17. Lughezzani G, Zorn KC, Budäus L, et al: Comparison of three different tools for prediction of seminal vesicle invasion at radical prostatectomy. Eur Urol 62:590-596, 2012
18. Pepe MS, Fan J, Feng Z, et al: The Net Reclassification Index (NRI): A misleading measure of prediction improvement even with independent test data sets. Stat Biosci 7:282-295, 2015
19. Van Calster B, Vickers AJ: Calibration of risk prediction models: Impact on decision-analytic performance. Med Decis Making 35:162-169, 2015
20. Vickers AJ, Cronin AM: Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: Towards a decision analytic framework. Semin Oncol 37:31-38, 2010
21. Bryant RJ, Sjoberg DD, Vickers AJ, et al: Predicting high-grade cancer at ten-core prostate biopsy using four kallikrein markers measured in blood in the ProtecT study. J Natl Cancer Inst 107:7, 2015
22. Bossuyt PM, Reitsma JB, Linnet K, et al: Beyond diagnostic accuracy: The clinical utility of diagnostic tests. Clin Chem 58:1636-1643, 2012

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or jco.ascopubs.org/site/ifc.

Kathleen F. Kerr
No relationship to disclose

Marshall D. Brown
No relationship to disclose

Kehao Zhu
No relationship to disclose

Holly Janes
No relationship to disclose

Appendix

Mathematical Derivation of Net Benefit and Decision Curves

Notation generally follows Baker and Kramer.[16]

Case: a patient with a negative outcome
Control: a patient without a negative outcome
P: the prevalence or rate of the negative outcome
B: A positive constant quantifying the expected benefit of intervention for a case
C: A positive constant quantifying the expected cost of intervention to a control, interpreted broadly (money, time, stress, negative health effects)
R: risk threshold for recommending treatment

Consider a policy in which individuals with risk of a negative outcome greater than some threshold R are considered high risk and are recommended an intervention, and individuals with risks lower than R are not considered high risk and are not recommended the intervention. A proportion of cases, true-positive rate (TPR_R), will be recommended the intervention and receive expected benefit B. However, a proportion of controls, false-positive rate (FPR_R), will also be recommended the intervention and have expected costs C. The expected net benefit (NB) of the policy to the population is the expected benefit to cases minus the expected costs to controls:

$$B\ TPR_R\ P - C\ FPR_R(1-P) = B\left[TPR_R\ P - \frac{C}{B}\ FPR_R(1-P)\right]$$

The first mathematical trick is to consider this NB to be measured in units of B. This leads to the following expression for population NB that is the basis for decision curves:

$$NB = TPR_R\ P - \frac{C}{B}\ FPR_R(1-P)$$

The second mathematical trick uses a classic result from decision theory that says the optimal risk threshold R satisfies

$$\frac{C}{B} = \frac{R}{1-R}$$

This equivalence leads to another expression for NB,

$$NB_R = TPR_R\ P - \frac{R}{1-R}\ FPR_R(1-P),$$

which expresses NB as a function of the TPR, FPR, the prevalence P of the outcome, and the risk threshold R used to recommend intervention. Although this expression for NB does not explicitly involve C and B, for this expression to be valid, one must assume that R has been chosen rationally, that is, R corresponds to the costs and benefits as given above. Additional important assumptions are that every case has the same expected benefit B of treatment, independent of predicted risk, and every control has the same expected cost C, independent of predicted risk.