BMC Medical Informatics and Decision Making

RESEARCH ARTICLE

Open Access

Using the weighted area under the net benefit curve for decision curve analysis

Rajesh Talluri[1] and Sanjay Shete[1,2*]

Abstract

Background: Risk prediction models have been proposed for various diseases and are being improved as new predictors are identified. A major challenge is to determine whether the newly discovered predictors improve risk prediction. Decision curve analysis has been proposed as an alternative to the area under the curve and net reclassification index to evaluate the performance of prediction models in clinical scenarios. The decision curve computed using the net benefit can evaluate the predictive performance of risk models at a given or range of threshold probabilities. However, when the decision curves for 2 competing models cross in the range of interest, it is difficult to identify the best model as there is no readily available summary measure for evaluating the predictive performance. The key deterrent for using simple measures such as the area under the net benefit curve is the assumption that the threshold probabilities are uniformly distributed among patients.

Methods: We propose a novel measure for performing decision curve analysis. The approach estimates the distribution of threshold probabilities without the need of additional data. Using the estimated distribution of threshold probabilities, the weighted area under the net benefit curve serves as the summary measure to compare risk prediction models in a range of interest.

Results: We compared 3 different approaches, the standard method, the area under the net benefit curve, and the weighted area under the net benefit curve. Type 1 error and power comparisons demonstrate that the weighted area under the net benefit curve has higher power compared to the other methods. Several simulation studies are presented to demonstrate the improvement in model comparison using the weighted area under the net benefit curve compared to the standard method.

Conclusions: The proposed measure improves decision curve analysis by using the weighted area under the curve and thereby improves the power of the decision curve analysis to compare risk prediction models in a clinical scenario.

Keywords: Decision curve analysis, Clinical decision making, Area under the curve, Net benefit curves, Threshold probabilities

Background

Risk prediction models are used to predict the probability of occurrence of future events for individuals based on several predictors. Predicting the risk of malignant events is of major importance for public health as such information can be used to improve outcomes and personalize clinical care. Risk prediction models have been developed for several cancers [1–3], a variety of conditions and general public health issues (e.g., hypertension, diabetes, cardiovascular disease, smoking experimentation) [4–7]. These risk prediction models are being constantly improved with the identification of new predictors (e.g., genetic markers) associated with the disease or condition of interest. However, assessing the contribution of these predictors in improving risk prediction is challenging.

The area under the receiver operating characteristic curve (AUC) is generally used to determine the predictive accuracy of a model [8]. The AUC provides a natural tool to select optimal models across all thresholds of

* Correspondence: sshete@mdanderson.org
[1] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Dr, FCT4.6002, Houston, TX 77030, USA
[2] Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

sensitivity and specificity. However, in clinical settings, there may be situations in which a model with higher AUC may not be desirable. For example, if a treatment involves high risk, the model with a low false positive rate would be the best model to use despite its lower AUC compared to those for other models. Also, in clinical settings, the models need not be accurate at the extreme ranges. For example, consider a scenario in which, if the predicted probability for the disease were below 0.2, the individual would not be screened/treated, and if the predicted probability was above 0.8, the individual would be screened/treated. Small differences in the predicted probabilities from 2 competing models would not make a significant clinical difference in the decision made by individuals at these extremes. However, the model that is accurate in predicting probabilities between the 0.2 - 0.8 range will be more useful in a clinical setting compared to the model that predicts probabilities well at the extremes. Hence, AUC may be a poor measure of performance for risk prediction models in certain clinical scenarios (Additional file 1: Figure S1). Importantly, the increase in AUC value may not be significant even when the new predictor is statistically associated with the response [9]. To alleviate the problem of low power, the net reclassification index [10] was proposed. However, several concerns have been raised regarding its appropriate use, interpretation, and associated high false positive rates [11, 12], suggesting the need for alternate measures for model comparison. One of the suggested methods was decision curve analysis (DCA) [13, 14].

DCA is used to evaluate the performance of prediction models in clinical decision making. The typical scenario for the application of DCA is when patients have symptoms that suggest a disease but they have not yet been diagnosed with the disease. The clinician has to make a decision regarding whether a biopsy/screening should be performed to diagnose the disease. The biopsy or other screening procedure is associated with various risks or side effects. The decision thus depends on the probability of the disease for that patient, the patient's preferences, the possible side effects and the clinician's experience. If the probability of the disease is too high or too low, the decision is generally clear. DCA provides a way to assess the performance of a model in a specific range of interest. DCA has been extensively used to compare competing methods in several diseases [15–17].

DCA is based on the computation of the net benefit for a model. The decision curve computed using the net benefit details the performance of the model at a given threshold probability or in a range of threshold probabilities that is of interest to the clinician making the decision. In several situations, when the net benefit curves for 2 competing models cross in the range of interest, it is difficult to select the best model. There is no available summary measure that can determine the better model. The key deterrent for using the area under the net benefit curve as the summary measure is the assumption that the threshold probabilities would need to follow a uniform distribution. And, if they do not follow a uniform distribution, additional data such as the exact threshold probabilities and patient preferences need to be collected to estimate the threshold probabilities, which limits the application of DCA to data sets that lack these additional data [18].

In this manuscript, we propose a novel way to estimate the distribution of threshold probabilities without collecting additional data by using only a binary clinical decision made by the clinician (e.g., whether screening is performed or not based on the disease probability) that is readily available for most of the data sets. Using the estimated distribution of threshold probabilities, we propose the weighted area under the net benefit curve in the range of interest as a summary statistic for model comparison. We performed several simulation studies to demonstrate the improvement in model comparison for the weighted area under the net benefit curve statistic compared to the standard method that uses confidence intervals to assess whether one model is statistically better than another [13].

Methods

Clinical scenario for decision curve analysis

Our guiding example will be the same clinical scenario used by Vickers and Elkin for DCA [14]. Individuals with prostate cancer face the possibility that the cancer could invade either one or both of their seminal vesicles, a condition described as seminal vesicle invasion (SVI). However, SVI is not officially diagnosed until after surgery, following an examination of the surgical sample by a pathologist. Hence, the surgeon has to make a decision regarding the removal of seminal vesicles before prostate surgery, based on the predicted probability of SVI. Several models have been proposed for assessing the probability of SVI prior to prostate surgery, based on predictors such as prostate specific antigen (PSA) and Gleason score (GS) [19, 20]. After estimating the probability of SVI using one of the risk prediction models, the clinician or the patient has a decision to make regarding whether or not the seminal vesicles will be removed during surgery. If the probability of SVI is low, the tip of the seminal vesicles is preserved in surgery to prevent long-term loss of urinary continence [21]. If the probability of SVI is high, and the seminal vesicles are not removed, there is a risk of recurrence of prostate cancer. The decision to remove the seminal vesicles is made using a threshold probability p_t, which depends on many factors such as the preference of the patient and the clinician and other covariates such as the age of the patient. If the predicted probability of SVI is greater than the threshold probability, p_t, then the seminal vesicles are removed.

Data simulation

The data for the simulation were based on the clinical scenario for predicting SVI in prostate cancer. We considered a cohort of n patients with prostate cancer. For simulation purposes, we assumed the risk of SVI depends on GS [22, 23], PSA [22, 23], and a generic covariate labeled X_1 (additional covariates such as age, body mass index, and ethnicity can be added to this model). We simulated PSA using an exponential distribution because, typically, PSA levels in prostate cancer patients are heavily skewed towards larger values, which can be simulated using a heavy tailed distribution. We chose a rate parameter of 0.1 to correspond closely to prostate cancer cohorts [24]. We simulated primary and secondary grades for tumors using binomial distributions. The final GS values in the range of 2 to10 were obtained by adding the primary and secondary grades (GS = 2+ binomial $(n = 4,\ \ p = 0.5)$ + binomial $(n = 4,\ \ p = 0.5)$). The mean for the GS score was 6, which corresponds closely to the average GS score for prostate cancer cohorts [25]. Finally, we simulated X_1 using a normal distribution, with a mean of 27 and a standard deviation of 6. We modeled X_1 based on the values for the mean and standard deviation of the BMI from the 2010 US census. Let $Y = 1$ and $Y = 0$ correspond to the presence and absence of SVI, respectively, in the cohort of n prostate cancer patients, and let p_d denote the probability of SVI ($p_d = P[Y = 1]$). The simulation model is as follows:

$$logit(p_d) = -10 + 0.1PSA + 0.2X_1 + 0.5GS \qquad (1)$$

We calculated the probability of SVI in individuals with prostate cancer using this model. We needed to simulate another decision indicator $Z = [0,1]$, which indicates whether the clinician or the patient decided to have the seminal vesicles removed. This is dependent on the threshold probability p_t. The distribution of p_t in the population is generally unknown. However, p_t is likely to be on the lower side for diseases with serious consequences and to be higher for diseases with minimal consequences. For simulation purposes, we simulated p_t using a beta distribution, *Beta* (2,7). If the disease probability p_d was greater than or equal to the threshold probability p_t, the decision would be made to remove the seminal vesicles during surgery ($Z = 1$), otherwise the seminal vesicles would not be removed ($Z = 0$). The above simulation process was used for all the data simulations we report here, with changes to the distribution of p_t and the addition of new predictors for SVI based on the simulation scenario. The number of needed replicates was determined by using a method proposed in [26]. Using the desired precision (half width of the 95 % confidence interval) to be 5, we needed 743 replicates. Therefore, all simulation results are based on 1000 replicates of a cohort of 10000 individuals.

Data analysis

The purpose of DCA in this study was to compare 2 competing models for the prediction of SVI. We assume 2 models M1 and M2 for predicting the probability of SVI. We used the net benefit to compute decision curves for the 2 competing models. The net benefit [27] was defined as

$$Net\,Benefit = \frac{True\,Positives}{n} - \frac{False\,Positives}{n}\left(\frac{p_t}{1-p_t}\right),$$

where, p_t is the threshold probability and n is the total number of individuals.

The general approach for DCA involves computing the net benefit for the 2 models and selecting the model that has higher net benefit at a particular threshold p_t or in a range of thresholds [14]. The standard method uses confidence intervals for the net benefit curve to assess whether one model is statistically better than another [13]. We implemented this approach as described below. To evaluate the confidence intervals, we first resampled the data set with replacement K times. We then used these K data sets to estimate K corresponding net benefit curves. The confidence interval for the net benefit curves at each probability threshold p_t was estimated using α/2 and 1−α/2 percentiles for the bootstrap distribution for a confidence interval coverage of 1−α.

When comparing 2 models using DCA, it is recommended to use the same bootstrap samples for calculating the net benefit for the 2 competing models in order to produce accurate and shorter confidence intervals [13]. Hence, the difference in the net benefit curves for the 2 models as a function of p_t is the statistic used for model comparison. Two models are said to be equivalent if the confidence interval for the difference in the net benefit curves includes zero, we refer to this standard approach as C-NBC. This decision can be made at a particular threshold probability or in a range of threshold probabilities of interest. However, in some situations (i.e., if model 1 is better for some values of p_t and model 2 is better for other values of p_t in the range of interest), it is difficult to identify the best model. Therefore, we initially propose the area under the net benefit curve (A-NBC) as a summary statistic for the performance of the model in the range of threshold probabilities of interest. When comparing 2 models using A-NBC, the statistic of interest is the difference in the area under the net benefit curves for the 2 competing models. The area under the net benefit curve in a range of p_t is computed using trapezoidal numerical integration. Two models are said to be equivalent if the difference in the A-NBC includes zero. The confidence intervals for the A-NBC statistic were obtained from the bootstrap distribution of the statistic.

Drawback to using the area under the net benefit curve

The drawback to using the A-NBC as a summary statistic is that the integral used to calculate the area assumes that the threshold probabilities are uniformly distributed in the range of interest [18]. In most of the clinical decision making scenarios, this is not a reasonable assumption. For example, most individuals would have lower values of p_t for highly malignant diseases (e.g., cancer) and higher values of p_t for comparatively harmless diseases (e.g., appendicitis). Hence, the distribution of p_t will depend on the disease, cost and benefits of the treatment and patient characteristics such as age, sex, etc. Because the A-NBC statistic does not utilize this information, a clinical decision using the A-NBC may not be practically optimal. By attributing the same weight to model performance at all threshold probabilities, the A-NBC statistic over weights the performance of the model when the clinical significance is comparatively lower and underweights the performance of the model at threshold probabilities where the clinical significance is higher. To overcome this obstacle, we proposed a novel method to estimate the distribution of p_t without any additional data, and calculated the weighted area under the net benefit curve based on the distribution of p_t to obtain improved estimates of model performance.

Estimating the distribution of p_t

The individual threshold probabilities are generally not available in existing datasets. However, one can estimate the distribution of p_t using the clinical decision of Z (i.e., removing or not removing the seminal vesicles during prostate surgery), and the predicted probability of SVI (p_d). The detailed derivation of the cumulative probability distribution of p_t is provided in the Appendix. Briefly, the cumulative probability distribution of p_t can be expressed as

$$P(p_t \le k) = P(Z = 1, p_t \le k) + P(Z = 0, p_t \le k).$$

The cumulative probability distribution for individuals who choose to have their seminal vesicles removed can be expressed as

$$P(Z = 1, p_t \le k) = P(Z = 1, p_d \le k) + \frac{P(pt \le k)}{P(pt \le pd)} P(Z = 1, p_d > k).$$

Similarly, the cumulative probability distribution for individuals who choose not to have their seminal vesicles removed can be expressed as

$$\begin{aligned} P(Z = 0, p_t \le k) = P(Z = 0) - P(Z = 0, p_d > k) \\ - \frac{1 - P(p_t \le k)}{1 - P(p_t \le p_d)} P(Z = 0, p_d \le k). \end{aligned}$$

Using the above equations, the final cumulative distribution of the threshold probability p_t can be expressed as

$$\begin{aligned} P(p_t \le k) &= P(Z = 1, p_d \le k) \\ &+ \frac{P(p_t \le k)}{P(p_t \le p_d)} P(Z = 1, p_d > k) \\ &+ P(Z = 0) - P(0, p_d > k) \\ &- \frac{1 - P(p_t \le k)}{1 - P(p_t \le p_d)} P(Z = 0, p_d \le k) \end{aligned} \qquad (2)$$

It is complicated to compute the solution to this equation by using traditional methods. Hence, we propose to use an iterative approach to infer the distribution. In our implementation, we used a uniform distribution as the initial estimate of the distribution of p_t. After initializing the starting distribution, the distribution of p_t is updated using equation (2). In the next iteration, the estimated distribution of p_t is used as the input value to equation (2). The process is repeated until the distribution converges (see Fig. 1).

Weighted area under the net benefit curves

After estimating the distribution of p_t, we proposed the weighted area under the net benefit curve (WA-NBC), as an improved summary statistic for comparing risk prediction models. The WA-NBC statistic is calculated as

$$\text{WA–NBC} = \int_{p_t} NBC(p_t) dp_t,$$

where NBC (p_t) is the net benefit curve for the corresponding model, f (p_t) is the density of p_t and the integration is over the range of interest of threshold probabilities. When 2 competing models are compared using the WA-NBC method, the statistic of interest is the difference in the WA-NBC statistics that correspond to the 2 models. The confidence intervals for the statistic were obtained using the standard bootstrap approach.

Results

We used 3 different approaches, C-NBC; A-NBC; and WA-NBC to compare 2 competing risk prediction models in the range of interest of threshold probabilities.

Simulation 1: Type 1 error and power

The type 1 error rates were based on the comparison of models M1 and M2. M1 included predictor variables PSA, GS, and X_1 as in equation (1) and 3 non-causal

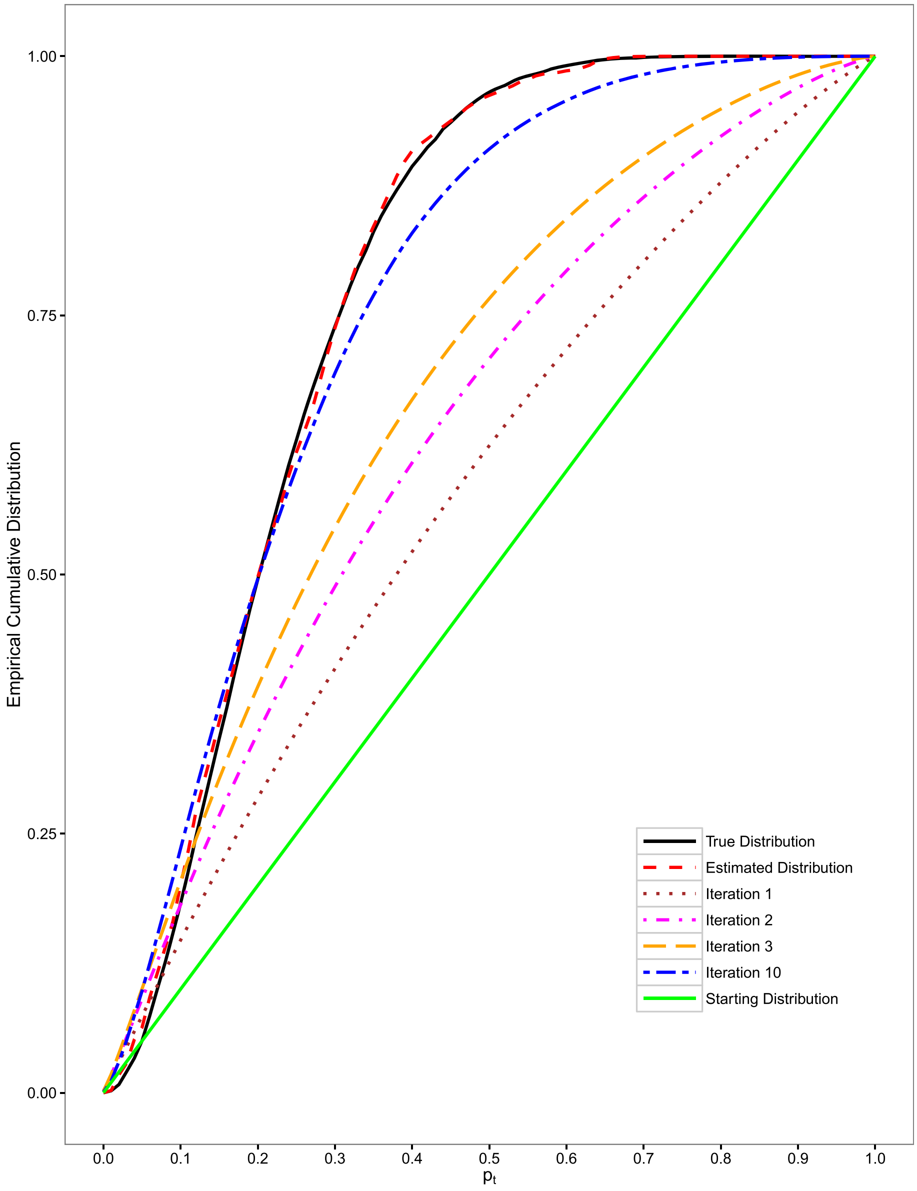

Fig. 1 Iterative steps involved in estimating the distribution of threshold probability p_t simulated using a beta distribution. The starting distribution is uniform; the intermediate distributions are shown for iterations 1, 2, 3 and 10; and the final estimated distribution computed after 100 iterations is equivalent to the true distribution

predictors (i.e., coefficient zero). M2 included PSA, GS, and X_1 as in equation (1) and 3 different non-causal predictors. The purpose of including different non-causal predictors in M1 and M2 was to evaluate the type 1 error rate when 2 models were equivalent but not identical. Therefore, we considered 2 models, M1 and M2, which had identical causal variables but different non-causal variables. In this simulation scenario, the models are equivalent, $M1 \equiv M2$, because the non-causal predictors have an effect size of zero. All the non-causal predictors were simulated using the standard normal distribution. The type 1 errors and powers were calculated using the bootstrap confidence interval approach. The type 1 errors for the 3 different approaches, C-NBC, A-NBC, and WA-NBC, were well controlled.

To assess power, the probability of SVI was simulated using the model

$$logit(P(Y = 1)) = -10 + 0.1PSA + 0.2X_1 + 0.5GS + \gamma n_1 + \epsilon.$$

The power for the methods was based on comparing the 2 models, M1 and M2, where M1 included all risk factors, PSA, X_1, GS, and n_1 and 2 non-causal predictors, n_2,n_3. M2 included PSA, X_1, and GS, but not n_1, and 3 additional independent, non-causal predictors. The predictors $n_1,n_2...,n_6$ were simulated using a standard normal distribution. M1 was superior to M2 because it included all the causal predictors. To compare the competing models, we simulated data using 3 values of γ (0.3,0.35 and 0.4) to illustrate the change in power while varying the effect of the causal predictor. For this simulation scenario, the range of interest for p_t was considered to be between 30 % to 50 % (i.e., we compared the performance of the methods in this range of p_t). The power for the C-NBC method was computed for threshold probabilities within the range of interest and averaged over the range (Table 1).

At γ=0.3, the statistical power achieved when using the C-NBC method was 0.31, which was lower than that achieved by the proposed WA-NBC method (0.53) at the 0.05 level of significance. For γ = (0.35,0.4) the power achieved by using the C-NBC method (0.46, 0.61) was lower than the power for the WA-NBC method (0.68, 0.79).

Simulation 2: Convergence of the iterative process of estimating the distribution of p_t

The patient threshold probabilities were modeled using a beta distribution *Beta* (2,7). The simulated data were used to estimate the distribution of the threshold probabilities using the recursive method (detailed in Methods). At each iteration, the estimated distribution of the threshold probabilities p_t approaches the true distribution of p_t. The final estimated distribution of p_t computed after 100 iterations converged to the original simulated distribution of threshold probabilities, as shown in Fig. 1. The convergence to the original simulated distribution of threshold probabilities when the patient threshold probabilities were modeled using a truncated exponential distribution (rate parameter 10 and truncated to the right at 1) is shown in Additional file 1: Figure S2.

Simulation 3: Impact of weighting the net benefit using the distribution of p_t

The following simulation shows the importance of using the estimated distribution of p_t that we proposed to use in calculating the WA-NBC statistic compared to simply using a uniform distribution for p_t that is used to calculate the A-NBC. The statistic of interest is the total net benefit for all the patients in the cohort, which summarizes the performance of a model for the cohort. We compared the total net benefit in 2 scenarios: 1) when the true threshold probabilities (p_t) were uniformly distributed; and 2) when the true threshold probabilities (p_t) were distributed as a beta distribution (Beta (2,7)). The model

$$logit(P(Y=1)) = \beta + \beta_1 PSA + \beta_2 X_1 + \beta_3 GS$$

was fitted to the training data set and the coefficients were estimated. The test data were used to calculate the net benefit curve for the model. The total net benefit for the cohort was assessed for the test data set using 3 methods: 1) Using the true values of p_t used in the simulation; 2) using the estimated distribution of p_t; and 3) using a uniform distribution. We simulated 1000 replicates to evaluate the confidence intervals for the total net benefit using the 3 methods (Table 2).

When the threshold probabilities were simulated using a uniform distribution, the net benefit obtained by using the estimated distribution compared to the uniform distribution was 1685.7 (1682.0–1689.3) compared to 1689.5 (1686.4–1692.5), respectively, which, as expected, was equivalent to the true net benefit of 1689.4 (1686.5–1692.3) obtained using the true values of p_t in the simulation. However, when the threshold probabilities were simulated using a beta distribution, the net benefit obtained by using the estimated distribution compared to the uniform distribution was 3013.9 (3010.5–3017.2) compared to 1692.1 (1689.2–1695.0), respectively. In this scenario, the true net benefit for the cohort, obtained using the true values of p_t used in the simulation, was 3013.7 (3010.3–3017.2), which is closer to the net benefit computed using the estimated distribution of p_t. The net benefit computed using the uniform distribution underestimated the total net benefit of the method and therefore underestimated model performance.

Table 1 Power comparison results using the net benefit curves (C-NBC) method and weighted area under the net benefit curves (WA-NBC) method to compare two models. The table shows the variation in power as the simulated coefficient of the causal predictor included in the superior model varied from 0.3 to 0.4

Method	Coefficient		
	0.3	0.35	0.4
C-NBC	0.31	0.46	0.61
WA-NBC	0.53	0.68	0.79

Table 2 Total net benefit comparison results using the estimated distribution of p_t and the uniform distribution. The two columns indicate the model for simulating p_t

	Simulated p_t	
Method	Uniform (0,1)	Beta (2,10)
True net benefit	1689.4 (1686.5–1692.3)	3013.7 (3010.3–3017.2),
Estimated net benefit	1685.7 (1682.0–1689.3)	3013.9 (3010.5–3017.2)
Uniform net benefit	1689.5 (1686.4–1692.5)	1692.1 (1689.2–1695.0)

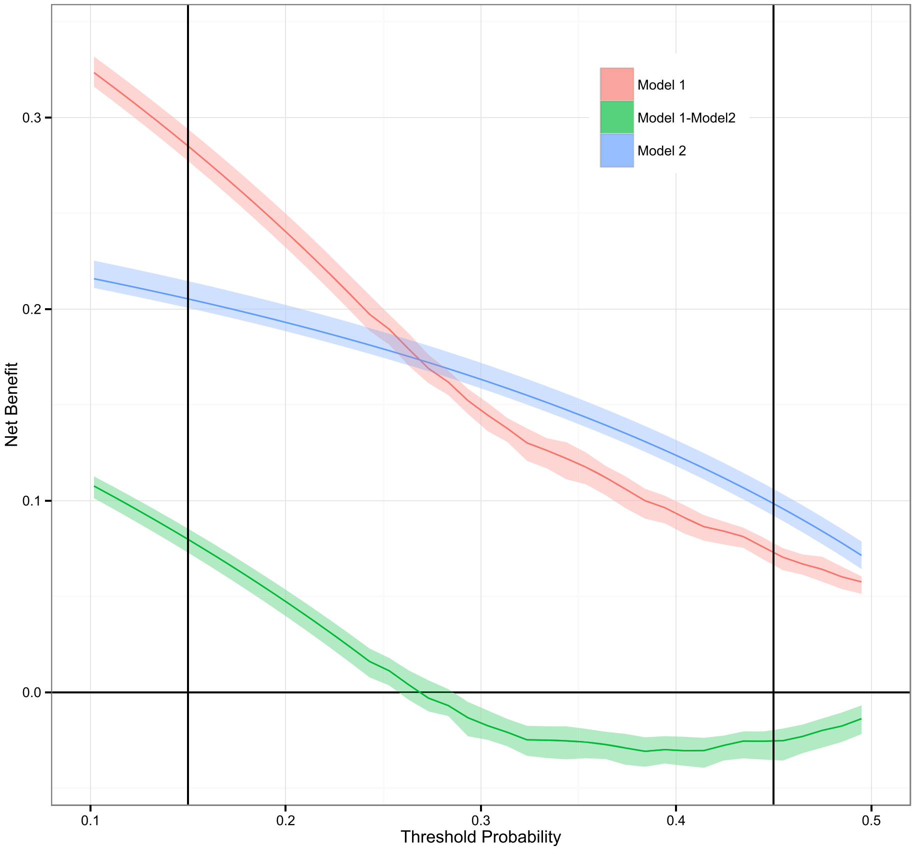

Fig. 2 Net benefit curves for models M1 and M2, along with their confidence intervals. The range of interest is from 0.15–0.45 as indicated by the *vertical lines*. This example showcases that model 1 is better than model 2 for study participants even though model 2 is better than model 1 for most of the risk thresholds (0.27–0.45) in the range of interest

Simulation 4: Example showcasing the utility of WA-NBC

The threshold probabilities were simulated from a beta distribution (Beta (2,7)). In this analysis, the performance of the 2 models M1 and M2 was compared in the p_t range of 15 % to 45 %. The models are

$$\text{Model M1} : logit(P(Y=1)) = \beta_0 + \beta_1 PSA$$
$$\text{Model M2} : logit(P(Y=1)) = I(GS{\geq}6\, and\, X_1 > 25).$$

Net benefit curves were constructed for each of the models, along with confidence intervals (Fig. 2). As the 2 models cross in the region of interest, we cannot determine whether M1 > M2, M2 > M1, or M1 ≡ M2. We used the WA-NBC method and A-NBC method to evaluate the model performances in the range of interest. Using the A-NBC method, the confidence interval for the test statistic was (−0.0012, 0.0041), which includes zero, thus implying that M1 and M2 are equivalent in the range of interest. However, using the WA-NBC method, the confidence interval for the test statistic was (0.0064, 0.0157) which is above zero, thus implying that M1 is superior to M2 in the range of interest. As the threshold probabilities were simulated form a beta distribution that is right-skewed, most of the individuals would have lower threshold probabilities. And because model M1 is superior to model M2 at lower threshold probabilities (Fig. 2), most individuals would benefit from using model M1 compared to model M2, which is reflected in our analysis using WA-NBC. Thus, our simulation demonstrates the utility of the weighted area under the curve statistic in a particular situation when one cannot determine the best model using traditional DCA.

Discussion

In this paper, we present a novel method for estimating the distribution of threshold probabilities for individuals in clinical scenarios. This work was motivated by the absence of a straightforward way to compare 2 risk prediction models when the decision curves cross in the range of threshold probabilities of interest. The key deterrent for the use of a simple summary measure such as the area under the net benefit curve is its unrealistic assumption that the threshold probabilities are uniformly distributed in the range of interest. That assumption is largely unrealistic as the threshold probabilities depend on the given disease and other patient characteristics. For malignant diseases, these probabilities are right-skewed and for non-life-threatening diseases, they are left-skewed.

It is important to note that the individual values of p_t cannot be estimated without additional data; however, the distribution of the values of p_t can be estimated using existing data by solving the recursive equation for cumulative distribution of p_t. The recursive equation does not have a closed form solution because the cumulative distribution function is present in both the numerator and denominator in different forms. Therefore, an iterative solution was adopted to solve the equation to estimate the cumulative distribution of p_t. The rate of convergence was quick for several forms of hypothesized distributions of p_t (i.e., beta, uniform and truncated exponential) and almost always converged within 100 iterations, which took less than 3 min on a computer with a single processor and a speed of 3.4 GHz. We then used the estimated distribution to propose the weighted area under the net benefit curve as a novel summary measure for model comparison in the range of threshold probabilities of interest.

We performed several simulations to assess the performance of the proposed WA-NBC statistic and compared it to A-NBC and the standard approach, C-NBC. The type 1 errors were well controlled for all the methods. The statistical power for WA-NBC was higher than the power achieved when using the C-NBC method. We also showed that the total net benefit for the cohort obtained by using the estimated distribution was closer to the true total net benefit compared to using the uniform distribution to calculate the area under the net benefit curve. Thus, using the estimated threshold probability distribution accurately quantifies the model performance.

To demonstrate the utility of the weighted area under the curve, we simulated a scenario in which 2 competing models, M1 and M2, cross in the region of interest. In this scenario, it was difficult to make a decision using the C-NBC method. Using the A-NBC method led to the false conclusion that M1 was equivalent to M2, and using the proposed WA-NBC method provided the correct conclusion that M1 was superior to M2. The proposed weighted area under the net benefit curve is a superior measure for comparing risk prediction models when there is a crossover of net benefit curves in the region of interest. Importantly, we recommend using this measure even when the curves do not cross because this measure weights the net benefit curve with respect to the distribution of p_t, leading to a more practical estimation of the net benefit for future study participants.

Traditional measures of prediction such as AUC and NRI have limited value for risk prediction in clinical scenarios as they do not account for the cost of the treatment and the associated side effects. The proposed methodology based on the net benefit curves can be easily extended to include the cost and side effects by using a novel risk prediction model with additional predictors. The net harm corresponding to the additional predictors or existing predictors can be incorporated into the net benefit curve estimation using the following definition of net benefit [27]:

$$Net\,Benefit = \frac{True\,Positives}{n} - \frac{False\,Positives}{n}\left(\frac{p_t}{1-p_t}\right) - Net\,Harm$$

The net harm can also include cost effectiveness of the risk prediction models and other measures of utility. Accounting for the net harm will provide a more practical measure to identify the model that is most relevant to a particular clinical scenario.

The proposed methodology is generally used to compare 2 predefined risk prediction models. However, we can improve the efficiency of decision curve analysis using relative utility curves [28]. Relative utility curves are based on the contribution of the risk prediction model to clinical utility compared to a hypothesized perfect prediction. The test trade off, which is analogous to the net harm (discussed above) for net benefit curves, can then be used to evaluate the practical utility of the models when using relative utility curves. The main contribution of the proposed methodology is to recursively estimate the distribution of p_t, and subsequently use a weighted summary measure. This framework can be used to improve the performance of relative utility curves in the same manner as net benefit curves.

Conclusions

We have proposed a novel way to estimate the distribution of threshold probabilities without any additional data. Using the estimated distribution of threshold probabilities, we proposed the weighted area under the net benefit curve as a novel summary statistic for model comparison. We performed several simulation studies to demonstrate the improvement in model comparison using the weighted area under the net benefit curve statistic compared to the standard net benefit curves in various scenarios.

Additional file

Additional file 1: Supplementary Material.docx. Includes the Appendix and two supplementary figures referred in the manuscript. (DOCX 327 kb)

Abbreviations

A-NBC, Area based net benefit curve method; AUC, Area under the curve; C-NBC, Confidence interval based net benefit curve method; DCA, Decision curve analysis; PSA, Prostate specific antigen; GS, Gleason score; SVI, Seminal vesicular invasion; WA-NBC: Weighted area based net benefit curve method

Acknowledgements

We wish to thank 2 anonymous reviewers and associate editor for constructive comments on earlier version of the manuscript. We also thank Lee Ann Chastain for editing the manuscript.

Funding

This work was supported in part by the National Institutes of Health [grants R01CA131324, R01DE022891, and R25DA026120 to S. Shete]; a cancer prevention fellowship for R. Talluri supported by a grant from the National Institute of Drug Abuse [NIH R25DA026120]; the Barnhart Family Distinguished Professorship in Targeted Therapy (to S. Shete); and the Cancer Prevention Research Institute of Texas grant RP130123 (to S. Shete) and the National Cancer Institute grant CA016672 (to S. Shete).

Availability of data and materials

The data supporting the conclusions of this article are included within the article in Tables 1 and 2.

Authors' contributions

RT and SS conceived and designed the methodology. RT implemented the method. RT and SS wrote the manuscript. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing financial interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 22 January 2016 Accepted: 9 July 2016
Published online: 18 July 2016

References

1. Fears TR, Guerry D, Pfeiffer RM, Sagebiel RW, Elder DE, Halpern A, Holly EA, Hartge P, Tucker MA. Identifying individuals at high risk of melanoma: a practical predictor of absolute risk. J Clin Oncol. 2006;24(22):3590–6.
2. Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, Gail MH, Pfeiffer RM. Colorectal cancer risk prediction tool for white men and women without known susceptibility. J Clin Oncol. 2009;27(5):686–93.
3. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting Individualized Probabilities of Developing Breast-Cancer for White Females Who Are Being Examined Annually. J Natl Cancer Inst. 1989;81(24): 1879–86.
4. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: The Framingham heart study. Circulation. 2008;118(4):E86.
5. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino RB, Kannel WB, Vasan RS. A risk score for predicting near-term incidence of hypertension: The Framingham Heart Study. Ann Intern Med. 2008;148(2):102–10.
6. Talluri R, Wilkinson AV, Spitz MR, Shete S. A risk prediction model for smoking experimentation in Mexican American youth. Cancer Epidemiol Biomarkers Prev. 2014;23(10):2165–74.
7. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults-The Framingham Offspring Study. Arch Intern Med. 2007;167(10):1068–74.
8. Delong ER, Delong DM, Clarkepearson DI. Comparing the Areas under 2 or More Correlated Receiver Operating Characteristic Curves-a Nonparametric Approach. Biometrics. 1988;44(3):837–45.
9. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928–35.
10. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157–72.
11. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33(19):3405–14.
12. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? J Natl Cancer Inst. 2014;106(4):dju041.
13. 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.
14. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
15. Augustin H, Sun M, Isbarn H, Pummer K, Karakiewicz P. Decision curve analysis to compare 3 versions of Partin Tables to predict final pathologic stage. Urologic Oncology-Seminars and Original Investigations. 2012;30(4): 396–401.
16. Pulleyblank R, Chuma J, Gilbody SM, Thompson C. Decision Curve Analysis for Assessing the Usefulness of Tests for Making Decisions to Treat: An Application to Tests for Prodromal Psychosis. Psychol Assess. 2013;25(3):730–7.
17. Zastrow S, Brookman-May S, Cong TAP, Jurk S, Von Bar I, Novotny V, Wirth M. Decision curve analysis and external validation of the postoperative Karakiewicz nomogram for renal cell carcinoma based on a large single-center study cohort. World J Urol. 2015;33(3):381–8.
18. Steyerberg EW, Vickers AJ. Decision curve analysis: A discussion. Med Decis Making. 2008;28(1):146–9.
19. Guzzo TJ, Vira M, Wang YL, Tomaszewski J, D'amico A, Wein AJ, Malkowicz SB. Preoperative parameters, including percent positive biopsy, in predicting seminal vesicle involvement in patients with prostate cancer. J Urol. 2006; 175(2):518–21.
20. Zlotta AR, Roumeguere T, Ravery V, Hoffmann P, Montorsi F, Turkeri L, Dobrovrits M, Scattoni V, Ekane S, Bollens R, et al. Is seminal vesicle ablation mandatory for all patients undergoing radical prostatectomy? A multivariate analysis on 1283 patients. Eur Urol. 2004;46(1):42–9.
21. John H, Hauri D. Seminal vesicle-sparing radical prostatectomy: a novel concept to restore early urinary continence. Urology. 2000;55(6):820–4.
22. Gallina A, Chun FK, Briganti A, Shariat SF, Montorsi F, Salonia A, Erbersdobler A, Rigatti P, Valiquette L, Huland H, et al. Development and split-sample validation of a nomogram predicting the probability of seminal vesicle invasion at radical prostatectomy. Eur Urol. 2007;52(1):98–105.
23. Makarov DV, Trock BJ, Humphreys EB, Mangold LA, Walsh PC, Epstein JI, Partin AW. Updated nomogram to predict pathologic stage of prostate cancer given prostate-specific antigen level, clinical stage, and biopsy Gleason score (Partin tables) based on cases from 2000 to 2005. Urology. 2007;69(6):1095–101.
24. Kettermann AE, Ferrucci L, Trock BJ, Metter EJ, Loeb S, Carter HB. Interpretation of the prostate-specific antigen history in assessing life-threatening prostate cancer. BJU Int. 2010;106(9):1284–90. discussion 1290–1282.
25. Boyce S, Fan Y, Watson RW, Murphy TB. Evaluation of prediction models for the staging of prostate cancer. BMC Med Inform Decis Mak. 2013;13:126.
26. Hoad K, Robinson S, Davies R. Automated selection of the number of replications for a discrete-event simulation. J Oper Res Soc. 2010;61(11): 1632–44.
27. Peirce CS. The numerical measure of the success of predictions. Science. 1884;4(93):453–4.
28. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc. 2009;172(4):729–48.