[Part 3] – Model Evaluation using Novel Performance Measures

- Aasthaa Bansal, PhD
- The Comparative Health Outcomes, Policy and Economics (CHOICE) Institute
- University of Washington

TABLE 1. Characteristics of Some Traditional and Novel Performance Measures

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2, Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects
Discrimination	c statistic	ROC curve	Rank order statistic; interpretation for a pair of subjects with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean (y) versus mean ($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation; related to "shrinkage" of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification statistic		Compare observed outcomes to predicted risks within cross-classified categories
	Net reclassification index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction
	Integrated discrimination index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net benefit (NB) Decision curve analysis (DCA)	Cross-table Decision curve	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)

Traditional and Novel Performance Measures

- Overall performance
- Discrimination
- Calibration
- Reclassification
- Clinical usefulness

Traditional and Novel Performance Measures

So far (Part 1):

- Overall performance - **R^2** for survival outcomes
- Discrimination - **Time-dependent TP, FP, ROC**
- Calibration - **Hosmer-Lemeshow test** for survival outcomes

Now:

- Reclassification
- Clinical usefulness

Net Reclassification Index

- Pencina et al. (2008)
- Consider two marker/models that generate predictions
- Consider whether M_2 "moves" cases and controls relative to M_1
  - $D(t)$ disease status at time t
  - $p_1(t)$ and $p_2(t)$ based on M_1, M_2

	Case $D(t)=1$	Control $D(t)=0$
$p_2(t)-p_1(t)>0$	$P(+,1)$	$P(+,0)$
$p_2(t)-p_1(t)\leq 0$	$P(-,1)$	$P(-,0)$

- **NRI**:

$$NRI(t) = \underbrace{[P(+,1)-P(-,1)]}_{\texttt{case:up/down}} + \underbrace{[P(-,0)-P(+,0)]}_{\texttt{control:down/up}}$$

- French et al. (2016) extend to survival outcomes

Net Reclassification Index

- Criticism: Does not use decision analytic weights, leads to misleading estimates and false inferences (Pepe et al. 2013; Kerr et al. 2014; Vickers & Pepe 2014)
- Decision analytically weighted version proposed (Van Calster et al. 2013)

Discrimination Slope / IDI

- Pencina et al. (2008)
- **Integrated Discrimination Improvement**
- Idea:
  - Let $p(M_i) = P[D \mid M_i]$
  - Contrast mean risk in cases and mean risk in controls

$$\Delta P = \overline{p}_{case} - \overline{p}_{control}$$

  - IDI $= \Delta P_{new} - \Delta P_{old}$
  - Note: model-based
- Uno et al. (2012) extend to $D = 1(T \leq t)$ (cumulative cases)
- Liang & Heagerty (2016) extend to $D = 1(T = t)$ (incident cases)

Discrimination Slope / IDI

Software:

- Uno et al. (2012): R package `survIDINRI`
- Liang & Heagerty (2016): R package `hds` (https://github.com/liangcj/hds)

Traditional and Novel Performance Measures

- Overall performance - **R^2** for survival outcomes
- Discrimination - **Time-dependent TP, FP, ROC**
- Calibration - **Hosmer-Lemeshow test** for survival outcomes
- Reclassification - **IDI** for survival outcomes
- Clinical usefulness

Prediction, action, cost, and consequence

- So far, we have focused on prediction **accuracy**
- Consider the context where predictions guide clinical actions, and patients have **outcomes** (e.g. QALY) that depend on their disease status and the action taken (i.e. the treatment they receive)
- Need decision theoretic approaches to evaluate clinical usefulness of risk model

Example: Lung Transplantation

Lung Allocation Score (`http://www.unos.org`)

- "The lung allocation score (LAS) is used to prioritize waiting list candidates based on a combination of **waitlist urgency** and **post-transplant survival**."
- In this context, waitlist urgency is defined as **what is expected to happen** to a candidate, given his or her characteristics, in the next year if he or **she does not receive a transplant.**
- Post-transplant survival is defined as **what is expected to happen** to a candidate, given his or her characteristics, in the first year after a transplant if he or **she does receive the transplant.**

Clinical Usefulness

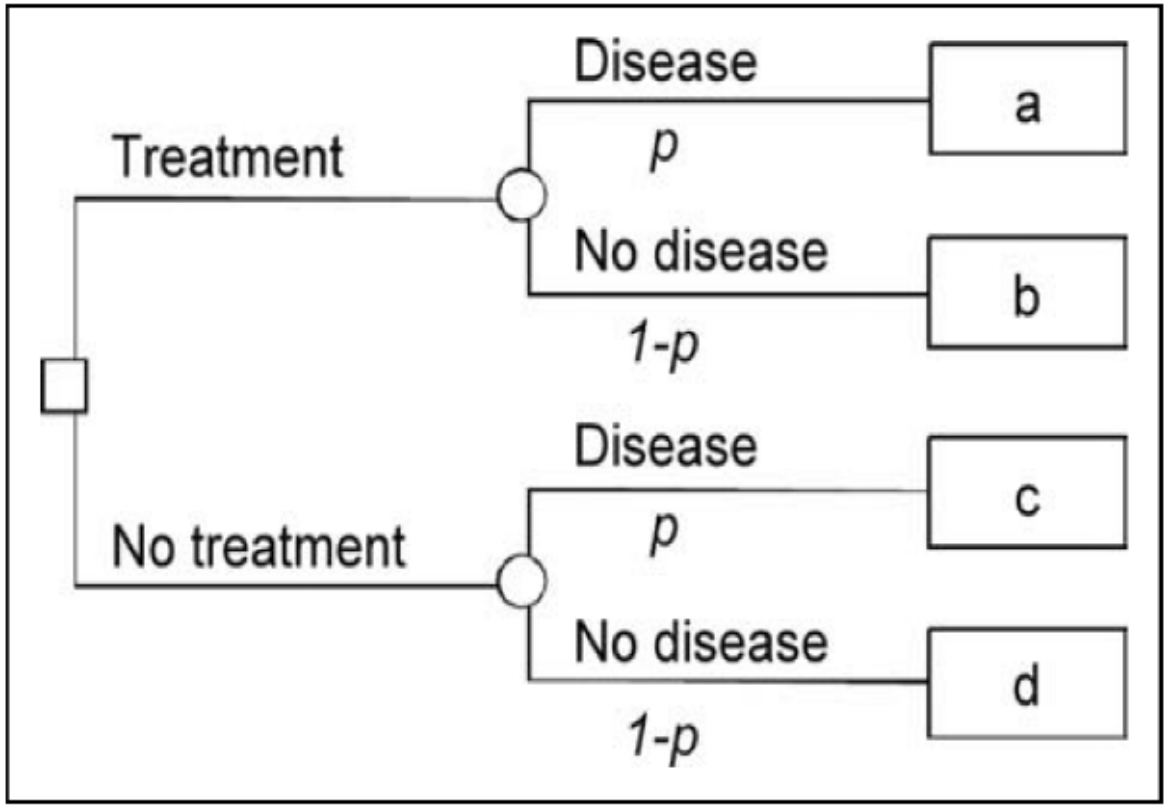

Figure 1 A decision tree for treatment. The probability of disease is given by p; a, b, c, and d give, respectively, the value of true positive, false positive, false negative, and true negative.

(Vickers & Elkin, 2006)

Decision Curve Analysis

- Vickers & Elkin (2006):
  - In a formal decision analysis, harms and benefits need to be quantified, leading to optimal decision threshold
  - Difficult to define threshold if
    (1) Insufficient data on harms and benefits
    (2) Relative weight of harms and benefits varies across patients
  - Propose decision curve analysis

Decision Curve Analysis

- Consider a risk score/marker M and a **decision function**:

$$A(M, m) \;=\; 1(M > m)$$

- Let $A(M, m) = 1$ denote that treatment is used, and let $A(M, m) = 0$ denote that no treatment is used
- **Q**: What is the population mean if **no treatment** is used?
- **Q**: What is the population mean if **universal treatment** is used?
- **Q**: What is the population mean if **selective treatment** is used?

Decision Curve Analysis

- **No treatment**: $Tx = A(M, +\infty) \equiv 0$

treatment	disease	group size	mean
Tx	D	**0**	a
	$\overline{D}$	**0**	b
$\overline{Tx}$	D	$p(D)$	c
	$\overline{D}$	$p(\overline{D})$	d

Population mean:

$$\mu(0) = c \cdot p(D) + d \cdot p(\overline{D})$$

Decision Curve Analysis

- **Universal treatment**: $Tx = A(M, -\infty) \equiv 1$

treatment	disease	group size	mean
Tx	D	$p(D)$	a
	$\overline{D}$	$p(\overline{D})$	b
$\overline{Tx}$	D	**0**	c
	$\overline{D}$	**0**	d

Population mean:

$$\mu(1) = a \cdot p(D) + b \cdot p(\overline{D})$$

Decision Curve Analysis

- **Selective treatment**: $A(M, m)$

treatment	disease	group size	mean
Tx	D	$p[A(M,m) = 1 \mid D] \cdot p(D)$	a
	$\overline{D}$	$p[A(M,m) = 1 \mid \overline{D}] \cdot p(\overline{D})$	b
$\overline{Tx}$	D	$p[A(M,m) = 0 \mid D] \cdot p(D)$	c
	$\overline{D}$	$p[A(M,m) = 0 \mid \overline{D}] \cdot p(\overline{D})$	d

Population mean:

$$\mu(m) = \underbrace{(a-c)\ TP(m)\ p(D) + (b-d)\ FP(m)\ p(\overline{D})}_{\Delta=\text{Net benefit}} + \underbrace{c \cdot p(D) + d \cdot p(\overline{D})}_{\mu(0)}$$

Decision Curve Analysis

Consider a policy based on decision function $A(M, m)$ that recommends **selective treatment** based on risk threshold m.

- $TP(m)$ represents proportion of cases recommended treatment based on threshold m, experiencing consequence $a - c$
- $FP(m)$ represents proportion of non-cases recommended treatment based on threshold m, experiencing consequence $b - d$

Expected net benefit (NB) of policy using model/marker compared to $\mu(0)$ (no treatment) is:

$$\begin{aligned} \text{NB} &= (a-c)\ TP(m)\ p(D) \ + \ (b-d)\ FP(m)\ p(\overline{D}) \\ &= (a-c)\ TP(m)\ p(D) \ - \ (d-b)\ FP(m)\ p(\overline{D}) \end{aligned}$$

Interpretation:

- $(a - c)$: consequence of treating a **case** (vs not treating)
- $(d - b)$: consequence of not treating a **non-case** (vs treating)

Decision Curve Analysis

- **Equipoise**: If $\mu(1)$ was thought to be overall just as beneficial as $\mu(0)$ when $p(D) = p^*$ then we would have a classic result from decision theory:

$$\underbrace{a \cdot p(D) + b \cdot p(\overline{D})}_{\mu(1)} = \underbrace{c \cdot p(D) + d \cdot p(\overline{D})}_{\mu(0)}$$

$$\frac{(a-c)}{(d-b)} = \frac{1-p^*}{p^*}$$

Decision Curve Analysis

- If we standardize and set $(a - c) = 1$ then we obtain the relative benefit of not treating a non-case (i.e. $-$cost):

$$(d - b) = \frac{p^*}{(1 - p^*)}$$

- For any cost/benefit ratio, we can then compare use of a model/marker via $A(M, m)$ to $\mu(0)$ (no treatment) to obtain the standardized **net benefit**:

$$\begin{aligned} \text{NB} &= (a - c)\ TP(m)\ p(D) \ - \ (d - b)\ FP(m)\ p(\overline{D}) \\ &= TP(m)\ p(D) \ - \ \frac{p^*}{(1 - p^*)} FP(m) p(\overline{D}) \end{aligned}$$

Decision Curve Analysis

- Vickers & Elkin (2006) for personalized risk threshold:

  "At some probability between 0 and 1, the patient will be unsure whether to be treated. This threshold probability, p_t, is where the expected benefit of treatment is equal to the expected benefit of avoiding treatment."

- "Personal threshold for equipoise"
- E.g.: A patient might opt for treatment if benefit of treating disease is at least 5x greater than the cost of unnecessary treatment for no disease. Then

$$5 = \frac{p^*}{(1-p^*)} \Rightarrow p^* = 0.83$$

- Cost/benefit ratio may vary across patients:
  - $1 = \frac{p^*}{(1-p^*)} \Rightarrow p^* = 0.5$
  - $9 = \frac{p^*}{(1-p^*)} \Rightarrow p^* = 0.9$

Decision Curve Analysis

- **Decision curve**: Net Benefit (NB) across a range of probability thresholds or cost/benefit ratios

$$\Delta(p^*) = TP(p^*) \cdot p(D) \; - \; FP(p^*) \cdot p(\overline{D}) \cdot \frac{p^*}{(1-p^*)}$$

  where $p^* \in (0, 1)$

- Plot: $[\, p^*, \; \Delta(p^*) \,]$

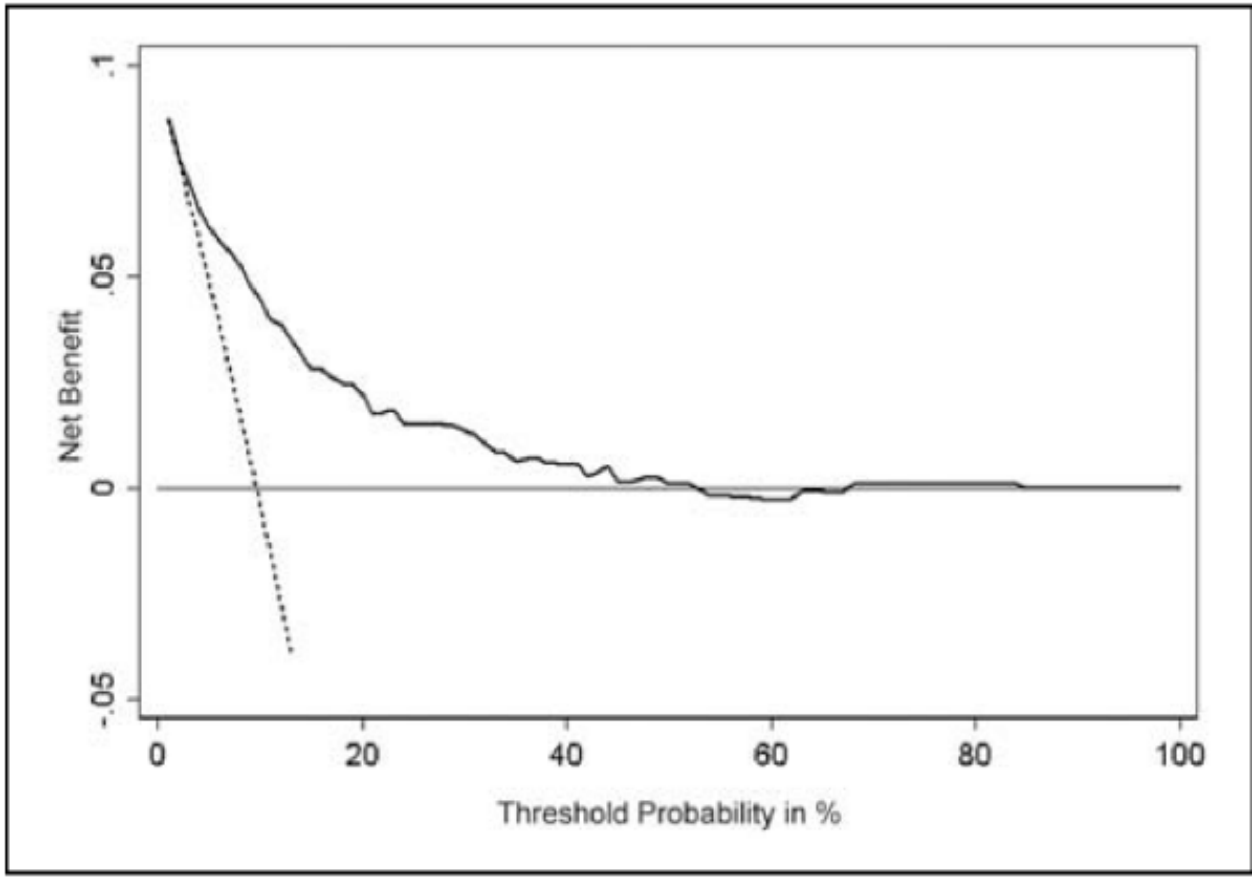

Figure 2 Decision curve for a model to predict seminal vesicle invasion (SVI) in patients with prostate cancer. Solid line: prediction model. Dotted line: assume all patients have SVI. Thin line: assume no patients have SVI. The graph gives the expected net benefit per patient relative to no seminal vesicle tip removal in any patient ("treat none"). The unit is the benefit associated with 1 SVI patient duly undergoing surgical excision of the seminal vesicle tip.

Extension to Censored Survival Outcomes

- Vickers et al (2008)
- Use time-dependent TP, FP shown earlier:

$$\begin{aligned} TP_t^C(c) &= P[M > c \mid T \leq t] \\ FP_t^D(c) &= P[M > c \mid T > t] \end{aligned}$$

- so that

$$\Delta_t(p^*) = TP_t(p^*) \cdot p(D) \; - \; FP_t(p^*) \cdot p(\overline{D}) \cdot \frac{p^*}{(1-p^*)}$$

Decision Curve Analysis

Comments/Discussion:

- Vickers & Elkin (2006) focus on choosing a risk threshold and calculating corresponding cost-benefit ratio.
    - This approach works if clinically defined threshold exists
    - If clinically defined threshold does not exist, might make more sense to first choose cost-benefit ratio, then calculate corresponding risk threshold
- Kerr et al (2016) highlight some **challenges** in using and interpreting decision curves appropriately:
    - Inability of decision curves to identify optimal risk threshold for recommending treatment
    - Impact of miscalibration
    - Assumption that every patient has the same expected benefit and cost of treatment

Decision Curve Analysis

Software:

- Vickers & Elkin (2006): Stata, R, SAS code and tutorial at `www.decisoncurveanalysis.org`
- Kerr et al (2016):
  - R package `DecisionCurve` (https://cran.r-project.org/web/packages/DecisionCurve/)
  - Include several additional features, e.g. allowing for x-axis to be labeled in terms of both risk threshold and cost:benefit ratio

Value of Information (VOI)

- Weinstein & Fineberg (1980); Ades, Lu & Claxton (2004)
- Consider two risk models that generate predictions:

  M1: Clinical variables
  M2: Clinical variables + biomarker measurement

- **Typical comparison**: Measure biomarker, compare outcomes under decisions made using M1 versus M2
- **VOI**: **Before** measuring biomarker, evaluate the value of additional information, while accounting for the uncertainty in measurements that we do not currently have
- Biomarker information may be highly accurate, but its clinical value depends on
  - the probability that the decision would change given biomarker information, and
  - the consequences/outcomes of decisions
- So far, methods have focused on diagnostic setting. Ongoing work by us to extend to longitudinal setting.

Traditional and Novel Performance Measures

- Overall performance - **R^2** for survival outcomes
- Discrimination - **Time-dependent TP, FP, ROC**
- Calibration - **Hosmer-Lemeshow test** for survival outcomes
- Reclassification - **IDI** for survival outcomes
- Clinical usefulness - **Decision curve analysis**, **VOI** - under development for dynamic decision-making

Decision theoretic measures versus ROC measures

- Vickers and Elkin (2006):

  "[The AUC metric] cannot tell us whether the model is worth using at all or which of 2 or more models is preferable. This is because metrics that concern accuracy do not incorporate information on consequences."

- High accuracy may not translate to high clinical utility.
- But need accuracy for clinical utility (accuracy is necessary, but not sufficient). Discrimination and calibration form the basis for measures of clinical usefulness, such as the net benefit (Vickers & Elkin, 2006).
- Steyerberg et al (2010):

  "we suggest that reporting discrimination and calibration will always be important for a prediction model. Decision-analytic measures should be reported if the model is to be used for making clinical decisions."

Conclusions

- We provided an array of methods for risk prediction and evaluation of predictions in longitudinal cohort studies.
- Evaluation measures are defined based on risk, thus allow for incorporation of multiple longitudinal markers and other covariates.
- Our estimation procedures are robust and perform well in a wide range of scenarios.

Thanks!