

# Will Net Benefit trump Net Reclassification Index as a measure for incremental value of markers in prediction models? A historical perspective

ISCB, Basel  
Aug 28, 2025

Ewout W. Steyerberg, PhD

*Professor of Clinical Biostatistics and  
Medical Decision Making*

Dept of Biomedical Data Sciences

Leiden University Medical Center

Chair of Julius Center, University Medical Center Utrecht

Ben Van Calster, PhD for STRATOS TG6

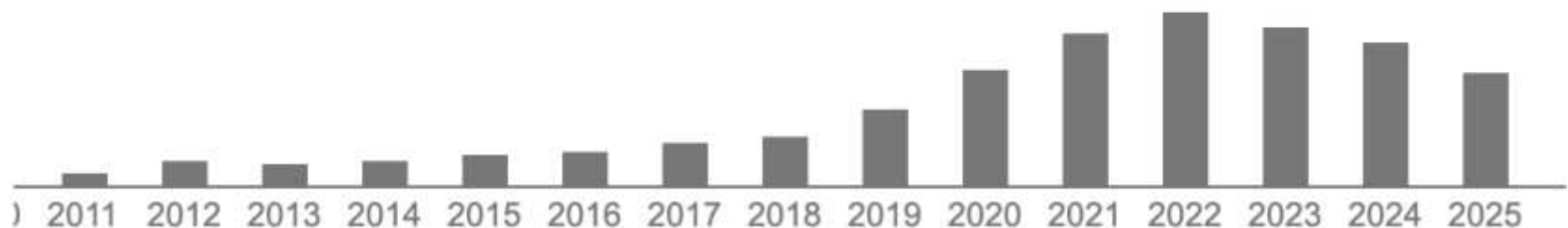


## Decision Curve Analysis: A Novel Method for Evaluating Prediction Models

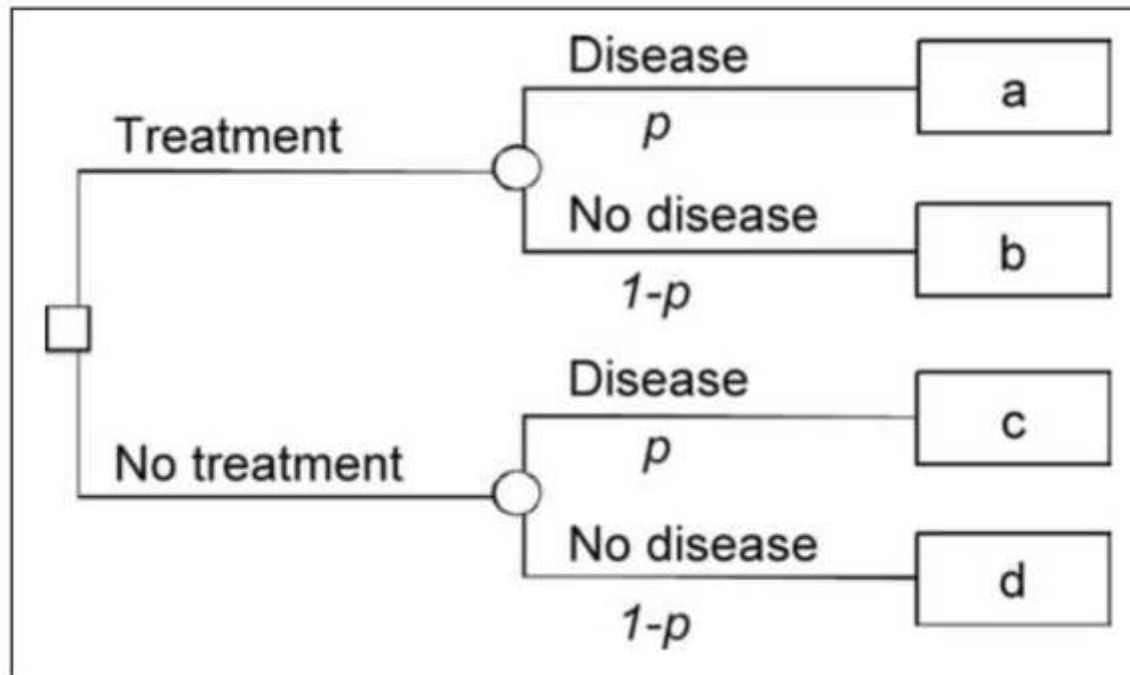
*Andrew J. Vickers, PhD, Elena B. Elkin, PhD*

**MEDICAL DECISION MAKING/NOV-DEC 2006**

Geciteerd door 4513

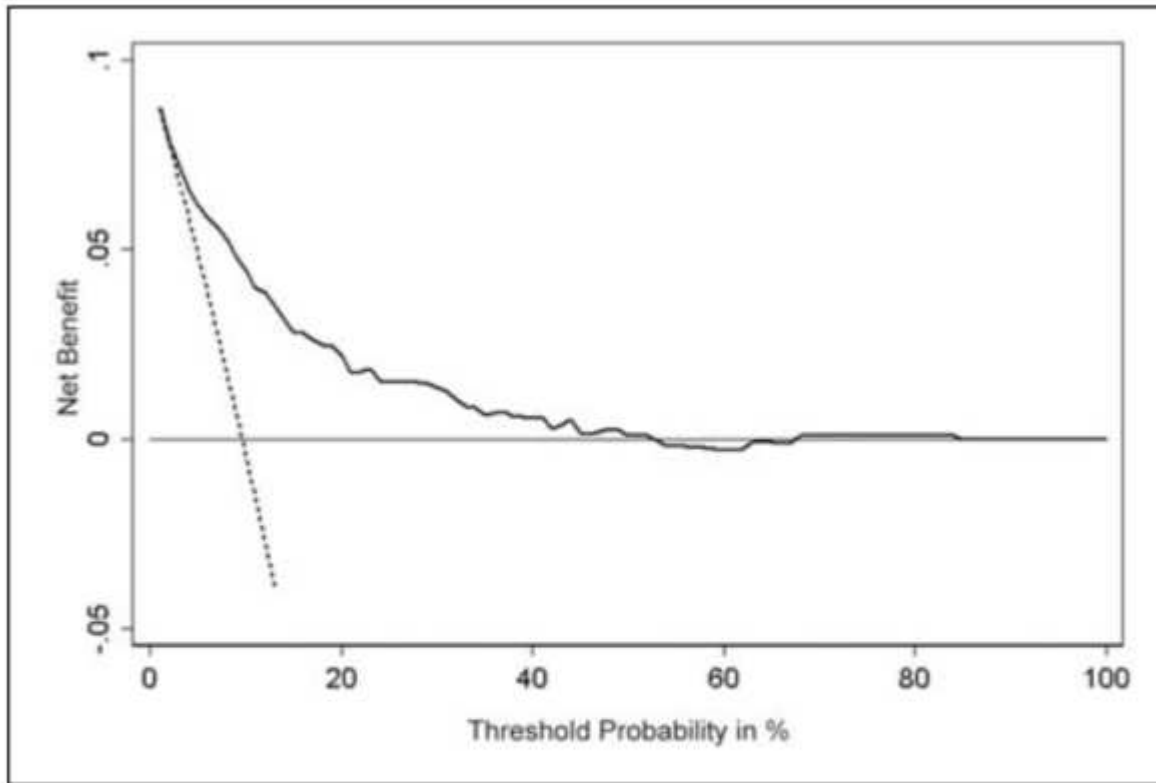


# Decision making and Net Benefit



$$\text{Net benefit} = \frac{\text{true-positive count}}{n} - \frac{\text{false-positive count}}{n} \left( \frac{p_t}{1-p_t} \right).$$

# Decision curve



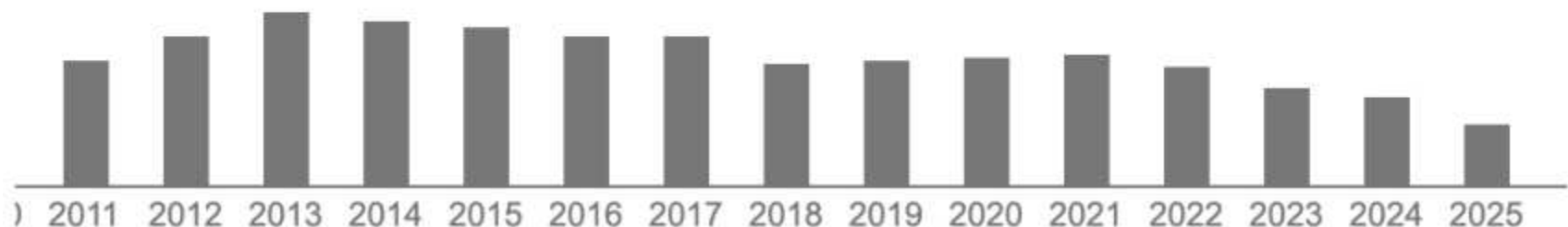
*Figure 2 Decision curve for a model to predict seminal vesicle invasion (SVI) in patients with prostate cancer. Solid line: prediction model. Dotted line: assume all patients have SVI. Thin line: assume no patients have SVI. The graph gives the expected net benefit per patient relative to no seminal 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.*

## Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond

Michael J. Pencina<sup>1,\*†</sup>, Ralph B. D'Agostino Sr<sup>1</sup>, Ralph B. D'Agostino Jr<sup>2</sup>  
and Ramachandran S. Vasan<sup>3</sup>

<sup>1</sup>*Department of Mathematics and Statistics, Framingham Heart Study, Boston University, 111 Cummington St., Boston, MA 02215, U.S.A.*

Geciteerd door 6819





## NRI definition

Net Reclassification Index:

$$\begin{aligned} & (\text{move up} \mid \text{event} - \text{move down} \mid \text{event}) + \\ & (\text{move down} \mid \text{non-event} - \text{move up} \mid \text{non-event} ) \end{aligned}$$

If dichotomy:

improvement in sensitivity + improvement in specificity

# Calculation of NRI

Table II. Reclassification among people who experience a CHD event and those who do not experience a CHD event on follow-up.

Model without HDL	Model with HDL			
Frequency (Row per cent)	<6 per cent	6–20 per cent	>20 per cent	Total
<i>Participants who experience a CHD Event</i>				
<6 per cent	39 (72.22)	15 (27.78)	0 (0.00)	54
6–20 per cent	4 (3.81)	87 (82.86)	14 (13.33)	105
>20 per cent	0 (0.00)	3 (12.50)	21 (87.50)	24
Total	43	105	35	183
<i>Participants who do not experience a CHD Event</i>				
<6 per cent	1959 (93.24)	142 (6.76)	0 (0.00)	2101
6–20 per cent	148 (16.78)	703 (79.71)	31 (3.51)	882
>20 per cent	1 (1.02)	25 (25.51)	72 (73.47)	98
Total	2108	870	103	3081

$$22/183=12\%$$

$$1/3081=.03\%$$

# Historical perspective

Birth of NB, DCA & NRI

Antenatal works

Peirce 1884

Cook 2007

Perinatal works

A happy youth?

Death / eternal life?



# Antenatal works

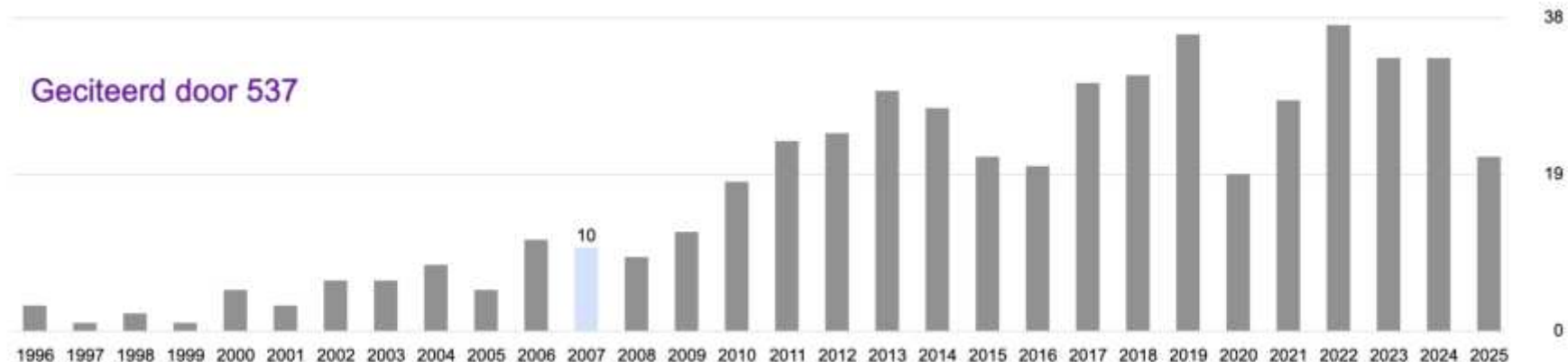
NOVEMBER 14, 1884.]

*SCIENCE.*

453

Youden index:  $\text{sens} + \text{spec} - 1$

Net Benefit:  $\text{TP} - w\text{FP} / n$



NOVEMBER 14, 1884.]

SCIENCE.

453

## The numerical measure of the success of predictions.

Suppose we have a method by which questions of a certain kind, presenting two alternatives, can in every case be answered, though not always rightly. Suppose, further, that a large number of such answers have been tabulated in comparison with the events, so that we have given the following four numbers:—

- (*aa*), the number of questions for which the answers were the first way and the events the first way;
- (*ab*), the number of questions for which the answers were the first way and the events the second way;
- (*ba*), the number of questions for which the answers were the second way and the events the first way;
- (*bb*), the number of questions for which the answers were the second way and the events the second way.

obtained by solving these equations is the measure of the science of the method. This value is,

$$i = \frac{(aa)}{(aa) + (ba)} - \frac{(ab)}{(ab) + (bb)},$$

$$= \frac{(aa)}{(aa) + (ba)} + \frac{(bb)}{(ab) + (bb)} - 1,$$

Another problem is to measure the utility of the method of prediction. For this purpose, let *p* be the profit, or saving, from predicting a tornado, and let *l* be the loss from every unfulfilled prediction of a tornado (outlay in preparing for it, etc.); then the average profit per prediction would be,

$$\frac{p \cdot (aa) - l (ab)}{(aa) + (ab) + (ba) + (bb)}.$$

C. S. PEIRCE.

# Incremental value of marker

- Classic approach:
  - Define a reference model, add marker to evaluate incremental value
  - Regression coefficient problematic (scaling); p-value assumed to be low
  - Increase in AUC / c statistic usually small (typically: +0.01)
    - something must be wrong: AUC “insensitive”; “only a rank order statistic”

## Special Report

### Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction

Nancy R. Cook, ScD

### Letter by Pepe et al Regarding Article, “Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction”

*To the Editor:*

Current statistical approaches for evaluation of risk prediction markers are unsatisfactory. We applaud Cook’s criticisms of the c-index, or area under the receiver operating characteristic curve. This index is based on the notion of pairing subjects, one with poor outcome (eg, cardiovascular event within 10 years) and one without, and determination of whether the risk for the former (ie, the case) is larger than the risk for the latter (ie, the control). This probability of correct ordering of risks is not a relevant measure of clinical value. It should not play a central role in evaluation of risk markers.

# Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women

The Reynolds Risk Score

---

Paul M Ridker, MD, MPH

---

Julie E. Buring, ScD

---

Nader Rifai, PhD

---

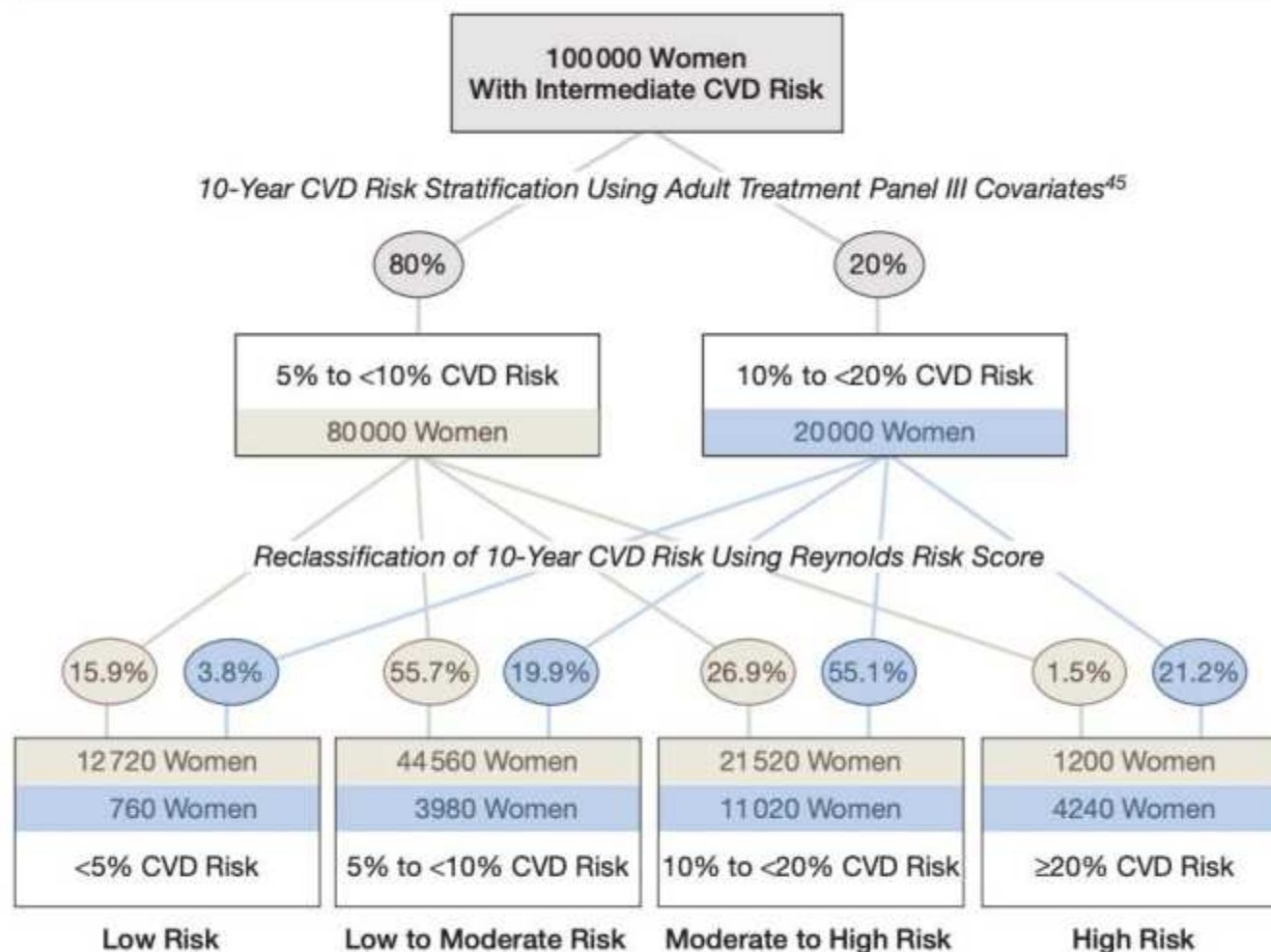
Nancy R. Cook, ScD

---

*JAMA. 2007;297:611-619*



**Figure.** Reclassification of Risk Using the Reynolds Risk Score for a Representative Population of 100 000 Intermediate-Risk US Women Without Diabetes



Percentages shown in ovals indicate the proportion of women distributed to risk categories based on Adult Treatment Panel III (top) and the Reynolds Risk Score (bottom). Reclassification using the Reynolds Risk Score is based on data shown in Table 5, Model B. CVD indicates cardiovascular disease.



## Use and misuse of the receiver operating characteristic curve in risk prediction.

**Cook NR.**

Circulation. 2007 Feb 20;115(7):928-35. doi: 10.1161/CIRCULATIONAHA.106.672402.

PMID: 17309939

Accepted risk factors such as lipids, hypertension, and smoking have only marginal impact on the c statistic individually yet lead to more accurate **reclassification** of large proportions of patients into higher-risk or lower-risk categories. ...

## Advances in measuring the effect of individual predictors of cardiovascular risk: the role of **reclassification** measures.

**Cook NR**, Ridker PM.

Ann Intern Med. 2009 Jun 2;150(11):795-802. doi: 10.7326/0003-4819-150-11-200906020-00007.

PMID: 19487714      **Free PMC article.**

Methods based on risk stratification have recently been proposed to compare predictive models. Such methods include the **reclassification** calibration statistic, the net **reclassification** improvement, and the integrated discrimination improvement. This article demonstr ...

## Performance of **reclassification** statistics in comparing risk prediction models.

**Cook NR**, Paynter NP.

Biom J. 2011 Mar;53(2):237-58. doi: 10.1002/bimj.201000078. Epub 2011 Feb 3.

PMID: 21294152      **Free PMC article.**

Concerns have been raised about the use of traditional measures of model fit in evaluating risk prediction models for clinical use, and **reclassification** tables have been suggested as an alternative means of assessing the clinical utility of a model. Several measures based ...

# Historical perspective

Birth of NB, DCA & NRI

Antenatal works

Peirce 1884

Cook 2007

**Perinatal works**

A happy youth?

Death / eternal life?

## 7 invited commentaries, Stat Med 2008

☐ 2 Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, R. S. Vasan, Statistics in Medicine (DOI: 10.1002/sim.2929).  
Greenland P.  
Stat Med. 2008 Jan 30;27(2):188-90. doi: 10.1002/sim.2976.  
PMID: 17579827 No abstract available.  
[View PDF](#)

☐ 3 Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929).  
Ware JH, Cai T.  
Stat Med. 2008 Jan 30;27(2):185-7. doi: 10.1002/sim.2985.  
PMID: 17668917 No abstract available.  
[View PDF](#)

☐ 4 Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929).  
Pepe MS, Feng Z, Gu JW.  
Stat Med. 2008 Jan 30;27(2):179-81. doi: 10.1002/sim.2991.  
PMID: 17671958 No abstract available.  
[View PDF](#)

☐ 5 Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929).  
Cook NR.  
Stat Med. 2008 Jan 30;27(2):191-5. doi: 10.1002/sim.2987.  
PMID: 17671959 No abstract available.  
[View PDF](#)

☐ 6 The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929).  
Che YY, Zhou XH.  
Stat Med. 2008 Jan 30;27(2):182-4. doi: 10.1002/sim.2986.  
PMID: 17712782 No abstract available.  
[View PDF](#)

☐ 7 The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929).  
Greenland S.  
Stat Med. 2008 Jan 30;27(2):199-204. doi: 10.1002/sim.2995.  
PMID: 17729377 No abstract available.  
[View PDF](#)

☐ 8 Comments on 'Evaluating the added predictive ability of a new marker' by M. Pencina, R. D'Agostino, R. D'Agostino Jr, R. Vasan, Statistics in Medicine (DOI: 10.1002/sim.2929).  
Kraemer HC.  
Stat Med. 2008 Jan 30;27(2):196-8. doi: 10.1002/sim.2948.



# NRI has 'absurd' weighting?

STATISTICS IN MEDICINE

*Statist. Med.* 2008; **27**:199–206

Published online 30 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2995

## COMMENTARY

The need for reorientation toward cost-effective prediction:  
Comments on 'Evaluating the added predictive ability of a new  
marker: From area under the ROC curve to reclassification and  
beyond' by M. J. Pencina *et al.*, *Statistics in Medicine*  
(DOI: 10.1002/sim.2929)

Sander Greenland<sup>\*,†</sup>

*Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, U.S.A.*

Any decision rule entails an implicit loss function, and the loss functions implicit in rules that appear to neglect loss functions **are usually clinically absurd**. One property of the loss function

The test criterion  $\Delta$  involves cost parameters that can be far beyond the scope of statistical expertise, involving matters of valuation and quality of life. It is then natural and may often suffice to focus statistical efforts on maximizing the accuracy of the risk score with and without  $X$ , to provide an accurate basis for further evaluations. Nonetheless, by including costs as free parameters in a loss function, a statistician can (with the aid of contextual experts) perform a sensitivity analysis over a range of reasonable values, rather than **rely on potentially absurd** implicit defaults. Occasionally, it may even be deemed worthwhile to statistically estimate costs as well as risks from available data, to provide a complete health-service evaluation.

# Relative utility (Stuart Baker)

## Putting Risk Prediction in Perspective: Relative Utility Curves

Stuart G. Baker

J Natl Cancer Inst 2009;101:1538–1542

Relative utility curves evaluate risk prediction models by comparing their net benefit at different risk thresholds to that of perfect prediction providing a normalized score (0 to 1) of clinical usefulness

Journal of the  
Royal Statistical Society

SERIES A  
Statistics  
in Society



Geciteerd door 86

*J. R. Statist. Soc. A* (2009)  
172, Part 4, pp. 729–748

### Using relative utility curves to evaluate risk prediction

Stuart G. Baker,

*National Cancer Institute, Bethesda, USA*

Nancy R. Cook,

*Brigham and Women's Hospital, Boston, USA*

Andrew Vickers

*Memorial Sloan–Kettering Cancer Center, New York, USA*

and Barnett S. Kramer

*National Institutes of Health, Bethesda, USA*

Geciteerd door 159



## Evaluation of Markers and Risk Prediction Models: Overview of Relationships between NRI and Measures

**Table 2** Overview of Relationships between Measures That Compare Classifications of 2 Competing Models at Risk Threshold  $T$

Condition Regarding $T$ and $P$	Relationships between Measures
If $T < P$	$\Delta RU_T = \frac{1}{w(1-P)} \Delta NB_T$
If $T \geq P$	$\Delta RU_T = \frac{1}{P} \Delta NB_T$ $wNRI_T = \frac{P}{T} \Delta RU_T$
If $T = P$	$NRI_T = \frac{1}{P} \Delta NB_T = \Delta RU_T = wNRI_T$
Irrespective of $T$ and $P$	$NRI_T = \Delta \text{Youden}$ $NRI_T = 2 * \Delta AUC_T$ (i.e., twice the difference in the areas under the prediction rules' single-point receiver operating characteristic curves) $wNRI_T = \frac{1}{T} \Delta NB_T$

Ben Van Calster

Stuart G. Baker,

[Med Decis Making 2013;3

cina, PhD,

rberg, PhD

**Table 2** Overview of Relationships between Measures That Compare Classifications of 2 Competing Models at Risk Threshold  $T$

Condition Regarding $T$ and $P$	Relationships between Measures
If $T < P$	$\Delta RU_T = \frac{1}{w(1-P)} \Delta NB_T$
If $T \geq P$	$\Delta RU_T = \frac{1}{P} \Delta NB_T$ $wNRI_T = \frac{P}{T} \Delta RU_T$
If $T = P$	$NRI_T = \frac{1}{P} \Delta NB_T = \Delta RU_T = wNRI_T$
Irrespective of $T$ and $P$	$NRI_T = \Delta \text{Youden}$ $NRI_T = 2 * \Delta AUC_T$ (i.e., twice the difference in the areas under the prediction rules' single-point receiver operating characteristic curves) $wNRI_T = \frac{1}{T} \Delta NB_T$

# Historical perspective

Birth of NB, DCA & NRI

Antenatal works

Peirce 1884

Cook 2007

Perinatal works

**A happy youth?**

**Extensions / reflections in methodological literature**

**Tremendous acceptance in medical literature**

Death / eternal life?

# Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers

Michael J. Pencina,<sup>a,b,\*†</sup> Ralph B. D'Agostino Sr<sup>c</sup> and Ewout W. Steyerberg<sup>d</sup>

Appropriate quantification of added usefulness offered by new markers included in risk prediction algorithms is a problem of active research and debate. Standard methods, including statistical significance and c statistic are useful but not sufficient. Net reclassification improvement (NRI) offers a simple intuitive way of quantifying improvement offered by new markers and has been gaining popularity among researchers. However, several aspects of the NRI have not been studied in sufficient detail.

In this paper we propose a prospective formulation for the NRI which offers immediate application to survival and competing risk data as well as allows for easy weighting with observed or perceived costs. We address the issue of the number and choice of categories and their impact on NRI. We contrast category-based NRI with one which is category-free and conclude that NRIs cannot be compared across studies unless they are defined in the same manner. We discuss the impact of differing event rates when models are applied to different samples or definitions of events and durations of follow-up vary between studies. We also show how NRI can be applied to case-control data. The concepts presented in the paper are illustrated in a Framingham Heart Study example.

In conclusion, NRI can be readily calculated for survival, competing risk, and case-control data, is more objective and comparable across studies using the category-free version, and can include relative costs for classifications. We recommend that researchers clearly define and justify the choices they make when choosing NRI for their application. Copyright © 2010 John Wiley & Sons, Ltd.

Geciteerd door 2565



Technical advance

Open Access

## **Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers**

Andrew J Vickers\*, Angel M Cronin, Elena B Elkin and Mithat Gonen

Address: Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, New York, NY 10065, USA

Email: Andrew J Vickers\* - [vickersa@mskcc.org](mailto:vickersa@mskcc.org); Angel M Cronin - [serioa@mskcc.org](mailto:serioa@mskcc.org); Elena B Elkin - [elkine@mskcc.org](mailto:elkine@mskcc.org); Mithat Gonen - [gonenm@mskcc.org](mailto:gonenm@mskcc.org)

\* Corresponding author

Published: 26 November 2008

Received: 3 June 2008

*BMC Medical Informatics and Decision Making* 2008, **8**:53 doi:10.1186/1472-6947-8-53

Accepted: 26 November 2008

Geciteerd door 1256





## Practice of Epidemiology

### Interpreting Incremental Value of Markers Added to Risk Prediction Models

The discrimination of a risk prediction model measures that model's ability to distinguish between subjects with and without events. The area under the receiver operating characteristic curve (AUC) is a popular measure of discrimination. However, the AUC has recently been criticized for its insensitivity in model comparisons in which the baseline model has performed well. Thus, 2 other measures have been proposed to capture improvement in discrimination for nested models: the integrated discrimination improvement and the continuous net reclassification improvement. In the present study, the authors use mathematical relations and numerical simulations to quantify the improvement in discrimination offered by candidate markers of different strengths as measured by their effect sizes. They demonstrate that the increase in the AUC depends on the strength of the baseline model, which is true to a lesser degree for the integrated discrimination improvement. On the other hand, the continuous net reclassification improvement depends only on the effect size of the candidate variable and its correlation with other predictors. These measures are illustrated using the Framingham model for incident atrial fibrillation. The authors conclude that the increase in the AUC, integrated discrimination improvement, and net reclassification improvement offer complementary information and thus recommend reporting all 3 alongside measures characterizing the performance of the final model.

# Net Reclassification Improvement: Computation, Interpretation, and Controversies

## A Literature Review and Clinician's Guide

Maarten J.G. Leening, MD, MSc; Moniek M. Vedder, MSc; Jacqueline C.M. Witteman, PhD; Michael J. Pencina, PhD; and Ewout W. Steyerberg, PhD

*Ann Intern Med.* 2014;160:122-131.

- Use clinically meaningful risk cutoffs for the category-based NRI
- Report both NRI components
- Address issues of calibration
- **Do not interpret the overall NRI as a % of the study population**
- Promising NRI findings need to be followed with decision analytic or formal cost-effectiveness evaluations

# Many applications of NRI, many in top journals

ORIGINAL ARTICLE

# Carotid-Wall Intima–Media Thickness and Cardiovascular Events

Joseph F. Polak, M.D., M.P.H., Michael J. Pencina, Ph.D.,  
Karol M. Pencina, Ph.D., Christopher J. O'Donnell, M.D., M.P.H.,  
Philip A. Wolf, M.D., and Ralph B. D'Agostino, Sr., Ph.D.

N Engl J Med 2011;365:213-21.

## RESULTS

A total of 296 participants had a cardiovascular event. The risk factors of the Framingham risk score predicted these events, with a C statistic of 0.748 (95% confidence interval [CI], 0.719 to 0.776). The adjusted hazard ratio for cardiovascular disease with a 1-SD increase in the mean intima–media thickness of the common carotid artery was 1.13 (95% CI, 1.02 to 1.24), with a nonsignificant change in the C statistic of 0.003 (95% CI, 0.000 to 0.007); the corresponding hazard ratio for the maximum intima–media thickness of the internal carotid artery was 1.21 (95% CI, 1.13 to 1.29), with a modest increase in the C statistic of 0.009 (95% CI, 0.003 to 0.016). The net reclassification index increased significantly after addition of intima–media thickness of the internal carotid artery (7.6%,  $P<0.001$ ) but not intima–media thickness of the common carotid artery (0.0%,  $P=0.99$ ). With the presence of plaque, defined as intima–media thickness of the internal carotid artery of more than 1.5 mm, the net reclassification index was 7.3% ( $P=0.01$ ), with an increase in the C statistic of 0.014 (95% CI, 0.003 to 0.025).

Download

dAUC 0.009; NRI 7.6%

dAUC 0.014; NRI 7.3%



# Historical perspective

Birth of NB, DCA & NRI

Antenatal works

Peirce 1884

Cook 2007

Perinatal works

A happy youth? **Mixed**

**Death / eternal life?**

**Kerr, Janes, Pepe, .. 2014**

**Hilden, Gerds, .. 2014**



Cite



Share



Favorites



Permissions

## METHODS

# Net Reclassification Indices for Evaluating Risk Prediction Instruments A Critical Review

Kerr, Kathleen F.<sup>a</sup>; Wang, Zheyu<sup>a</sup>; Janes, Holly<sup>b</sup>; McClelland, Robyn L.<sup>a</sup>; Psaty, Bruce M.<sup>c</sup>; Pepe, Margaret S.<sup>b</sup>

[Author Information](#) ☺


*Epidemiology* 25(1):p 114-121, January 2014. | DOI: 10.1097/EDE.0000000000000018

[Geciteerd door 460](#)


The paper argues that the NRI can be a misleading and unreliable measure due to its statistical properties and clinical interpretation, recommending alternative methods like

# Key Arguments from the Critical Review


## Misleading Results:

The study demonstrates that the NRI statistic can yield positive results even when the new marker has no predictive information, leading to incorrect conclusions about a marker's value. 


## Statistical Properties:

The authors highlight the lack of in-depth exploration of the NRI's statistical properties and show that its calculation can be influenced by poorly fitting risk models. 

## Lack of Clinical Relevance:

NRIs, especially overall NRIs, are criticized for lacking clear interpretation and ignoring the differential harms of different types of misclassification errors, which is crucial for clinical decision-making. 

## Incorrect Interpretations:

The paper points out that the NRI is often misinterpreted as a proportion of correctly reclassified patients, which is a fundamental error in its use. 



## Recommendations from the Critical Review

### Prefer Alternative Metrics:

The review suggests using more appropriate measures for evaluating prediction performance improvement, including:

- **Area Under the ROC Curve (C-index)**: Measures overall discrimination.
- **Net Benefit**: A decision-analytic metric that accounts for the costs of misclassification and provides a single-number summary of prediction increment.
- **Brier Score**: Evaluates prediction accuracy and calibration.

# Net Reclassification Indices for Evaluating Risk Prediction Instruments

## *A Critical Review*

**A note on the evaluation of novel  
biomarkers: do not rely on integrated  
discrimination improvement and net  
reclassification index**

On NRI, IDI, and “Good-Looking” Statistics with Nothing Underneath

**Net Risk Reclassification *P* Values: Valid or Misleading?**

**Does the Net Reclassification Improvement Help Us Evaluate Models and Markers?**

# Laure's questions

# Questions for methods development

1. When and how were negative aspects of the method discovered and published?

*At birth .. 7 Commentaries; most vigorous by Sander Greenland ('absurd')*

2. How was neutrality sought for in phase 3 and phase 4?

*Mathematics to align NRI with Net Benefit as wNRI*

*Hypothetical examples to expose problems (miscalibration)*

3. When and how was the method first used in applications?

*Immediately, we don't like '0.009' improvement, we like '8%'.*



Phase	Scope	Typical Activities	Outcome	NRI
I	Introduction of a new method	Theoretical development, proofs, asymptotics, basic illustrations	Demonstrates theoretical validity	Mixed reception
II	Initial application and evaluation	Limited simulations, real-life applications that are not too complex (with „cleaned data“, ...).  Inventor usually involved	Demonstrates usefulness, but still with caution, restricted to the specific investigated settings	Marker researchers ++ Epidemiologists + / – Decision-scientists – Statisticians – –
III	Broader evaluation and comparison of still relatively new method (compared to other probably established or new methods)	Neutral comparison studies (inventor bias avoided or transparently disclosed), extensive simulations, diverse real-world examples	Comparative performance, strengths, limitations	Kerr & Hilden: <b>killing</b>
IV	Evidence synthesis and increased understanding about a method that has been in use for some time	Reviews, complex applications, wide simulations in new, previously unconsidered settings, identification of possible pitfalls,	Clarifies when the method is preferable over others, or comparable to others, or when it	

# The Swiss city where even fun is serious



# Historical perspective

Birth of NB, DCA & NRI

Antenatal works

Perinatal works

A happy youth?

Death / eternal life?



# Net Benefit

$$\text{Net Benefit} = (\text{TP} - w \text{ FP}) / N$$

$$w = \text{cut-off} / (1 - \text{cut-off})$$

- e.g.: cut-off 50%:  $w = .5 / .5 = 1$ ;

cut-off 20%:  $w = .2 / .8 = 1/4$

- $w = H : B$  ratio

“Number of true-positive classifications,  
penalized for false-positive classifications”

Display as curve for plausible thresholds



## Poll utility-based measures

Net Benefit analysis is here to stay; it should become a standard element of prediction model performance assessment if predictions are intended to support decision making

*Agree / disagree*

Decision Curve Analysis (DCA) is here to stay; a picture is worth a 1000 words

*Agree / disagree*

The key difficulty with Net Benefit analysis and DCA is to determine a plausible threshold

*Agree / disagree*

## Poll NRI

NRI was a historical mistake, suffering from Frankenstein's law: you are responsible for the monster you create

**Agree**

Overall NRI should not be used to quantify incremental value of a marker

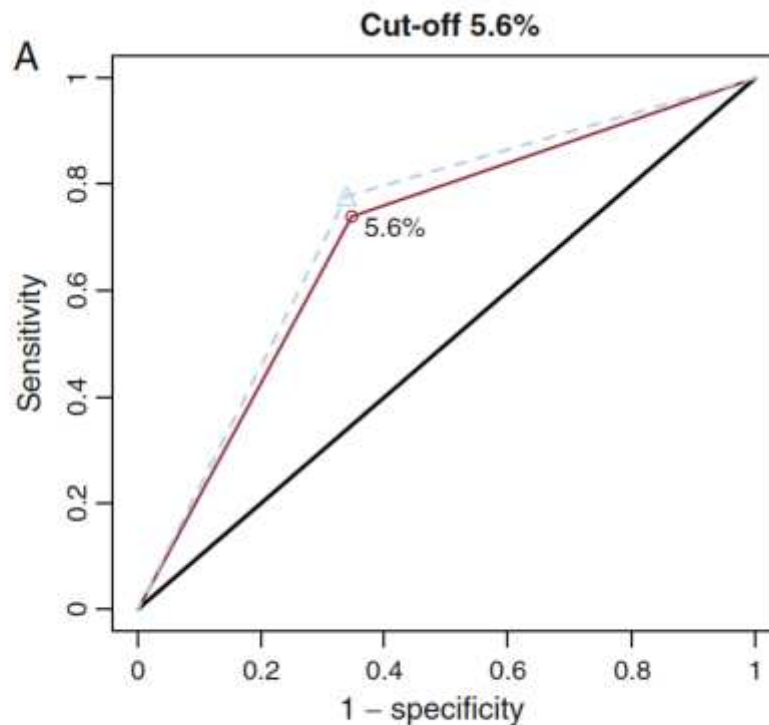
Agree / Disagree

## Delta AUC at 5.6% cut-off

AUC 0.696  $\rightarrow$  0.719, +0.023

Sens 0.738  $\rightarrow$  0.776, +0.038; spec 0.654  $\rightarrow$  0.661, +0.008

NRI = 0.038 + 0.008 = 0.046



**Figure 3.** Receiver operating characteristic curves with single cut-offs of 5.6% (A) and 0.719 for the 5.6% cut-off, and 0.550 and 0.579 for the 20% cut-off, for the

Steyerberg et al, *Rev Esp Cardiol.* 2011

# NRI and delta AUC

$$\text{NRI} = \text{delta}(\text{sens}) + \text{delta}(\text{spec})$$

$$\text{AUC for binary classification} = (\text{sens} + \text{spec}) / 2$$

$$\text{Delta AUC} = (\text{delta}(\text{sens}) + \text{delta}(\text{spec})) / 2$$

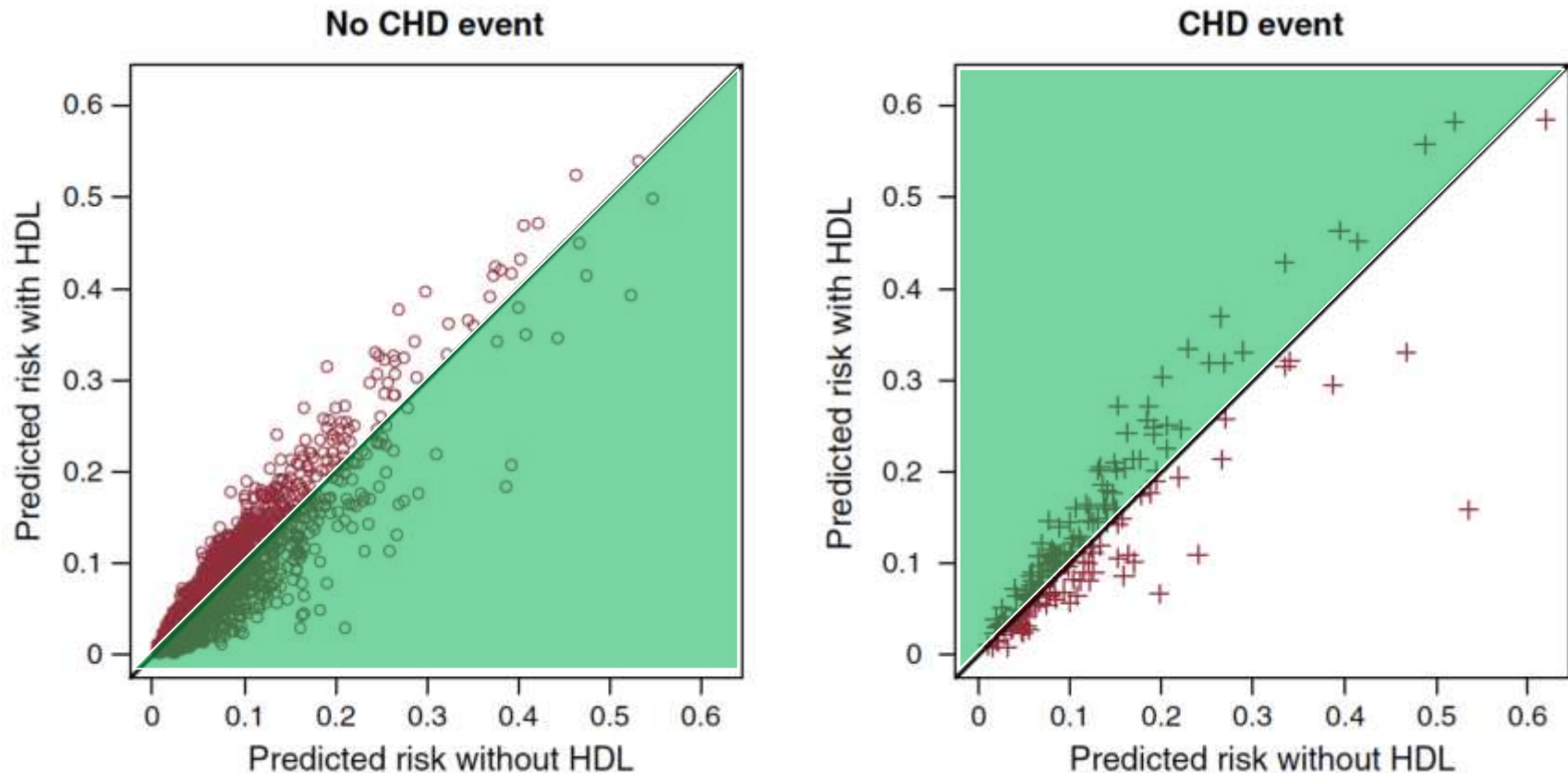
$$\text{NRI} = 2 \times \text{delta}(\text{AUC})$$

$$\text{Delta(Youden)} = \text{delta}(\text{sens}) + \text{delta}(\text{spec})$$

$$\text{NRI} = \text{delta}(\text{Youden})$$



# Reclassification plot



**Figure 5.** Reclassification plot. CHD, coronary heart disease; HDL, high-density lipoprotein.

Summary measure: Integrated Discrimination Index (IDI)

Similar to  $\Delta \text{Discrimination slope} / \Delta R^2$

# Decision-analytic variants

## Weighted NRI

Extensions of net **reclassification** improvement calculations to measure usefulness of new biomarkers.

**Pencina MJ**, D'Agostino RB Sr, Steyerberg EW.

Stat Med. 2011 Jan 15;30(1):11-21. doi: 10.1002/sim.4085. Epub 2010 Nov 5.

## Delta NB (Vickers)

## Delta Relative Utility (Baker) / standardized NB (Pepe / Janes)

# Historical perspective

Peirce, Science, 1884

Vergouwe, 2003

Vickers, MDM, 2006

Cook, Circulation, 2007

Pencina, Stat Med, 2008

Vickers, Extension in BMC, 2008

Baker, JNCI, 2009

Pencina, Extensions in Stat Med, 2011

Pencina, Interpretation in AmJEpi, 2012