Erasmus MC
Universitair Medisch Centrum Rotterdam

Evaluation of incremental value of a marker: a historic perspective on the Net Reclassification Improvement

Ewout Steyerberg

Petra Macaskill

Andrew Vickers

For TG 6 (Evaluation of diagnostic tests and prediction models)


WOMEN
MEN
Non-smoker
Smoker
Age
65
60
55
50
40
Systolic blood pressure (mmHg)
Total cholesterol (mmol/L)

Performance:
what is the quality of this prediction model?

- Statistical aspects

- Clinical perception

Overview

- NRI
  - Definition
  - Examples
- Methodological considerations
  - Positive and negative commentaries
- Clinical applications
  - Review of 67 papers
- Alternatives
  - Decision-analytic = utility respecting

How to quantify improvements in risk predictions?

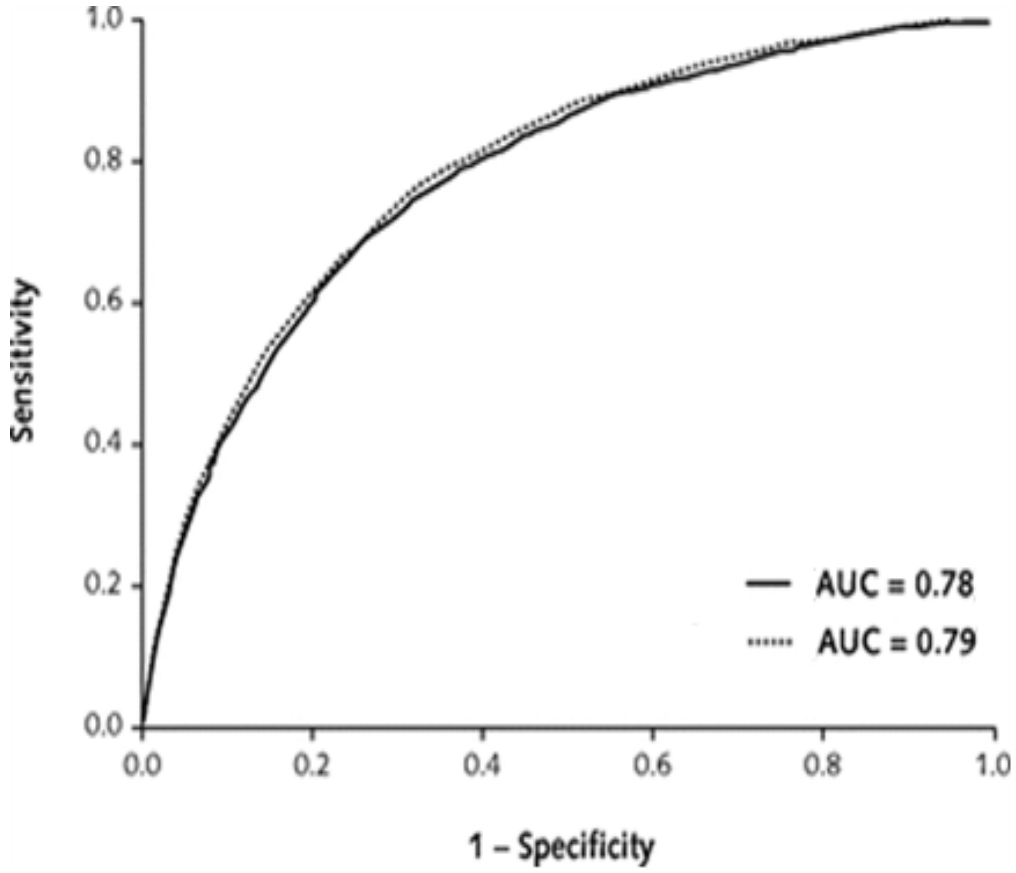
- Association: odds ratio, hazard ratio
- Discrimination: Δperformance measure
  ΔC-statistic; ΔR^2; ΔBrier
- Risk reclassification: net reclassification improvement (NRI)

Leening et al., Ann Intern Med 2014;160:122-31

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

*MJ **Pencina**, RB D'Agostino, RS Vasan - Statistics in medicine, 2008*

Geciteerd door 2170



Adding a marker to a model

- Typically small improvement in discriminative ability according to ΔC
- *c* stat blamed for being insensitive:
  "..too conservative .. as it hardly moves after a few good risk factors are already included in the model" (Pencina, Stat Med 2011).

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.

How to improve the impression of usefulness?

a) Use Δc and accept that is shows a small number

b) Multiply Δc by 2 and give it another name

c) Think about truly different measures
(and accept resulting small numbers)

Evaluation principles

- Consistency; e.g. use the same cut-off for model w/out marker
- “Proper scoring rule”
- Simple to interpret, not misleading

- Separate prediction from classification / decision making

Moving beyond the “insensitive” delta c

Net reclassification improvement (Pencina 2008)

- Sum of net percentages of correctly reclassified persons with and without the event of interest.

- =[Pr(up|event)–Pr(down|event)] + [Pr(down|nonevent)–Pr(up|nonevent)]

- Unit-less statistic due to implicit weighing for the event rate (p)
  - 1/p [= costs false negatives]
  - 1/(1-p) [= costs false positives]

- Theoretical range: -2 to +2

Leening and Steyerberg, JAMA 2013;309:2547-8

Leening et al., Ann Intern Med 2014;160:122-31

Event NRI and Non-event NRI

- The net percentage of persons with(out) the event of interest correctly reclassified

- Event NRI = Pr(up|event) – Pr(down|event)
- Non-event NRI = Pr(down|nonevent) – Pr(up|nonevent)

- Interpretable as a net percentage

- Theoretical range: -100% to +100%

Leening et al., Ann Intern Med 2014;160:122-31

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
Participants who experience a CHD Event				
<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
Participants who do not experience a CHD Event				
<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%

7

29

1/3081=.03%

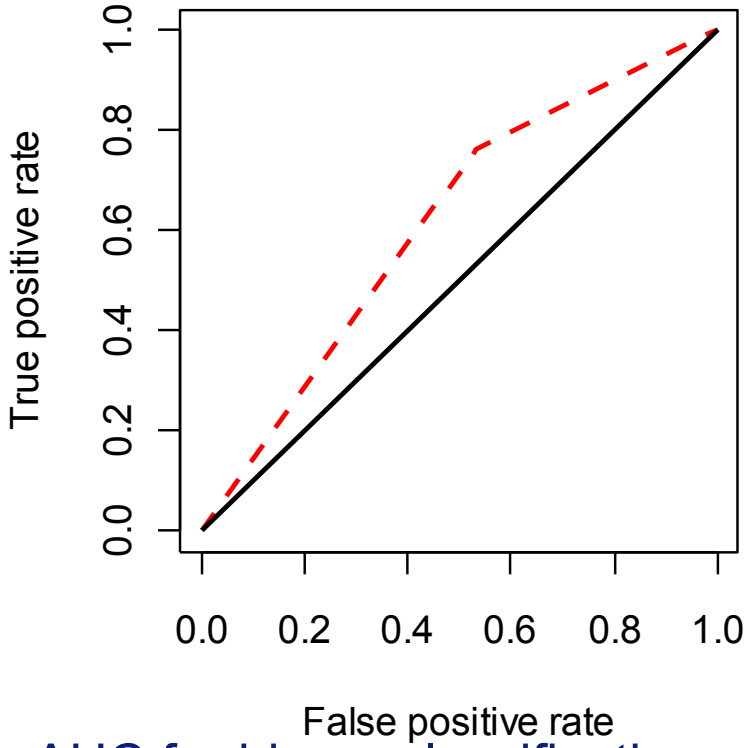
174

173

HDL cholesterol is routinely used in CHD prediction models [3–5]. Both IDI and NRI suggest that including it in the prediction model results in significant improvement in performance. That conclusion could not have been drawn relying solely on the increase in AUC. The increase in IDI, albeit significant, was of small magnitude—0.009 on the absolute scale or 7 per cent relative increase. It can be interpreted as equivalent to the increase in average sensitivity given no changes in specificity. Based on the NRI and its components, we conclude that addition of HDL improved classification for a net of 12 per cent of individuals with events, with no net loss for non-events. Even though the NRI results look convincing, caution needs to be given to their interpretation, as it is dependent on the somewhat arbitrary choice of categories.

Criticism on NRI (Commentaries Stat Med 2008)

- Nothing new; related to other measures
- Binary NRI = delta sens + delta spec



- AUC for binary classification = (sens + spec) / 2
- NRI = 2 x delta AUC

Interpretability

- NRI cited in >2000 papers
- Typical NEJM abstract:

  “The further addition to this model of information on CRP or fibrinogen increased the C-index by 0.0039 and 0.0027, respectively (P<0.001), and yielded a net reclassification improvement of 1.52% and 0.83%, respectively, for the predicted 10-year risk categories of "low" (<10%), "intermediate" (10% to <20%), and "high" (≥20%) (P<0.02 for both comparisons).”

N Engl J Med. 2012 Oct 4;367(14):1310-20. doi: 10.1056/NEJMoa1107477.

C-reactive protein, fibrinogen, and cardiovascular disease prediction.

Emerging Risk Factors Collaboration, Kaptoge S, Di Angelantonio E, Pennells L, Wood AM, White IR, Gao P, Walker M, Thompson A, Sarwar N, Caslake M,Butterworth AS, Amouyel P, Assmann G, Bakker SJ, Barr EL, Barrett-Connor E, Benjamin EJ, Björkelund C, Brenner H, Brunner E, Clarke R, Cooper JA,Cremer P, Cushman M, Dagenais GR, D'Agostino RB Sr, Dankner R, Davey-Smith G, Deeg D, Dekker JM, Engström G, Folsom AR, Fowkes FG, Gallacher J,Gaziano JM, Giampaoli S, Gillum RF, Hofman A, Howard BV, Ingelsson E, Iso H, Jørgensen T, Kiechl S, Kitamura A, Kiyohara Y, Koenig W, Kromhout D, Kuller LH, Lawlor DA, Meade TW, Nissinen A, Nordestgaard BG, Onat A, Panagiotakos DB, Psaty BM, Rodriguez B, Rosengren A, Salomaa V, Kauhanen J, Salonen JT, Shaffer JA, Shea S, Ford I, Stehouwer CD, Strandberg TE, Tipping RW, Tosetto A, Wassertheil-Smoller S, Wennberg P, Westendorp RG, Whincup PH,Wilhelmsen L, Woodward M, Lowe GD, Wareham NJ, Khaw KT, Sattar N, Packard CJ, Gudnason V, Ridker PM, Pepys MB, Thompson SG, Danesh J.

NRI as a %

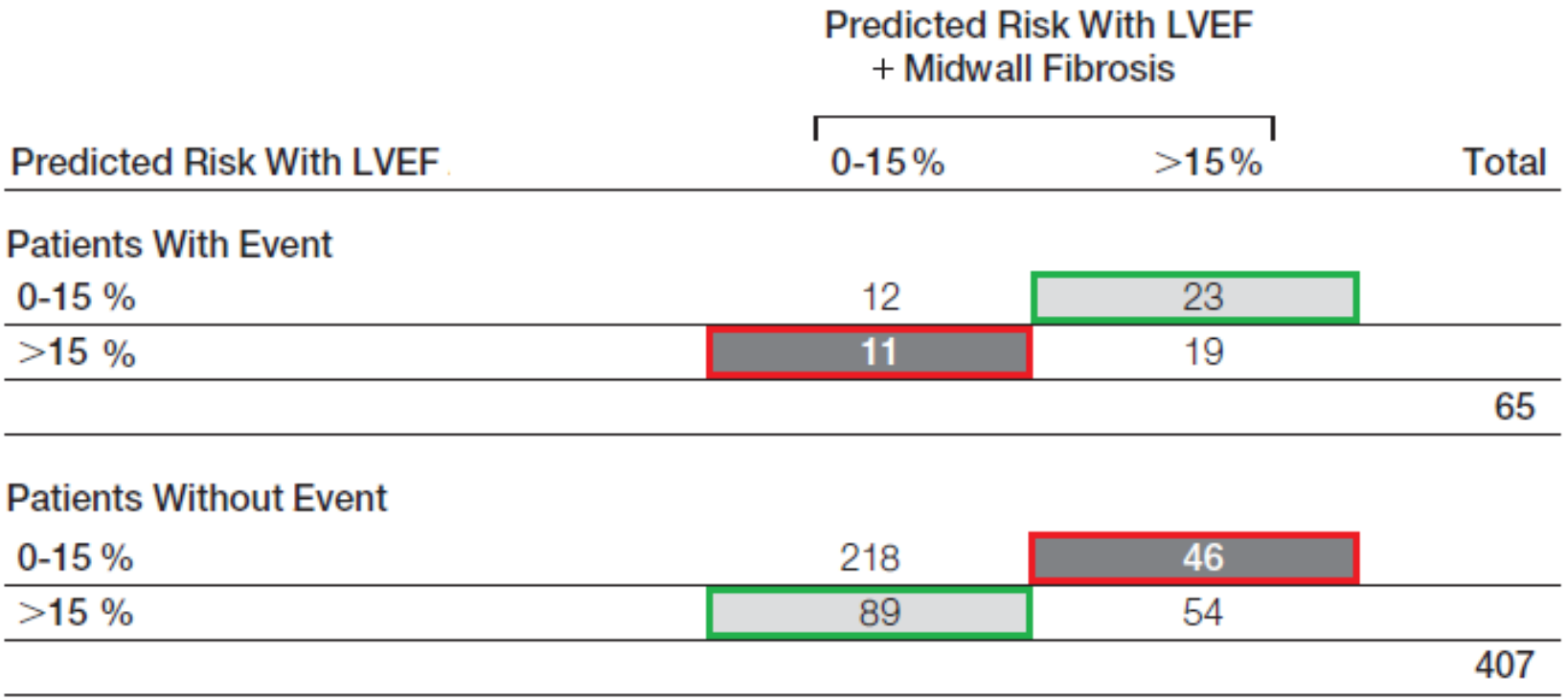
- Sum of 2 conditional probabilities

  P(move up | event – move down | event) +

  P(move down | non-event – move up | non-event)

- delta AUC smaller with more cut-offs;
  NRI larger with more cut-offs

Example JAMA 2013



Predicted Risk With LVEF	Predicted Risk With LVEF + Midwall Fibrosis: 0-15 %	Predicted Risk With LVEF + Midwall Fibrosis: >15 %	Total
Patients With Event			
0-15 %	12	23	
>15 %	11	19	
			65
Patients Without Event			
0-15 %	218	46	
>15 %	89	54	
			407

$$\text{NRI} = ([23 - 11] / 65) + ([89 - 46] / 407)$$
$$= 18\% + 11\%$$
$$= 0.29$$

Gulati et al., JAMA 2013;309:896-908

Leening and Steyerberg, JAMA 2013;309:2547-8

LTTE JAMA 2013

We believe that the authors misinterpreted the NRI results and consequently overestimated the contribution of myocardial fibrosis in risk reclassification for mortality and arrhythmias.

The authors erroneously simplified the interpretation of the NRI of 0.29 for the arrhythmic composite outcome by stating "Overall, 29% of patients were correctly reclassified after adding midwall fibrosis status to the risk model...

Results: interpretation

NRI ≠ “net percentage of persons correctly reclassified”

Leening & Steyerberg, JAMA 2013:accepted

Results: interpretation

Predicted Risk With LVEF	Predicted Risk With LVEF + Midwall Fibrosis: 0-15%	Predicted Risk With LVEF + Midwall Fibrosis: >15%	Total
Patients With Event			
0-15 %	12	23	
>15 %	11	19	
			65
Patients Without Event			
0-15 %	218	46	
>15 %	89	54	
			407

NRI = 18% + 11% = 0.29

Net percentage correctly reclassified

= [23 – 11] + [89 – 46] / [65 + 407]

= 12%

Gulati et al., JAMA 2013;209(9):896-908
Leening & Steyerberg, JAMA 2013:accepted

Literature review

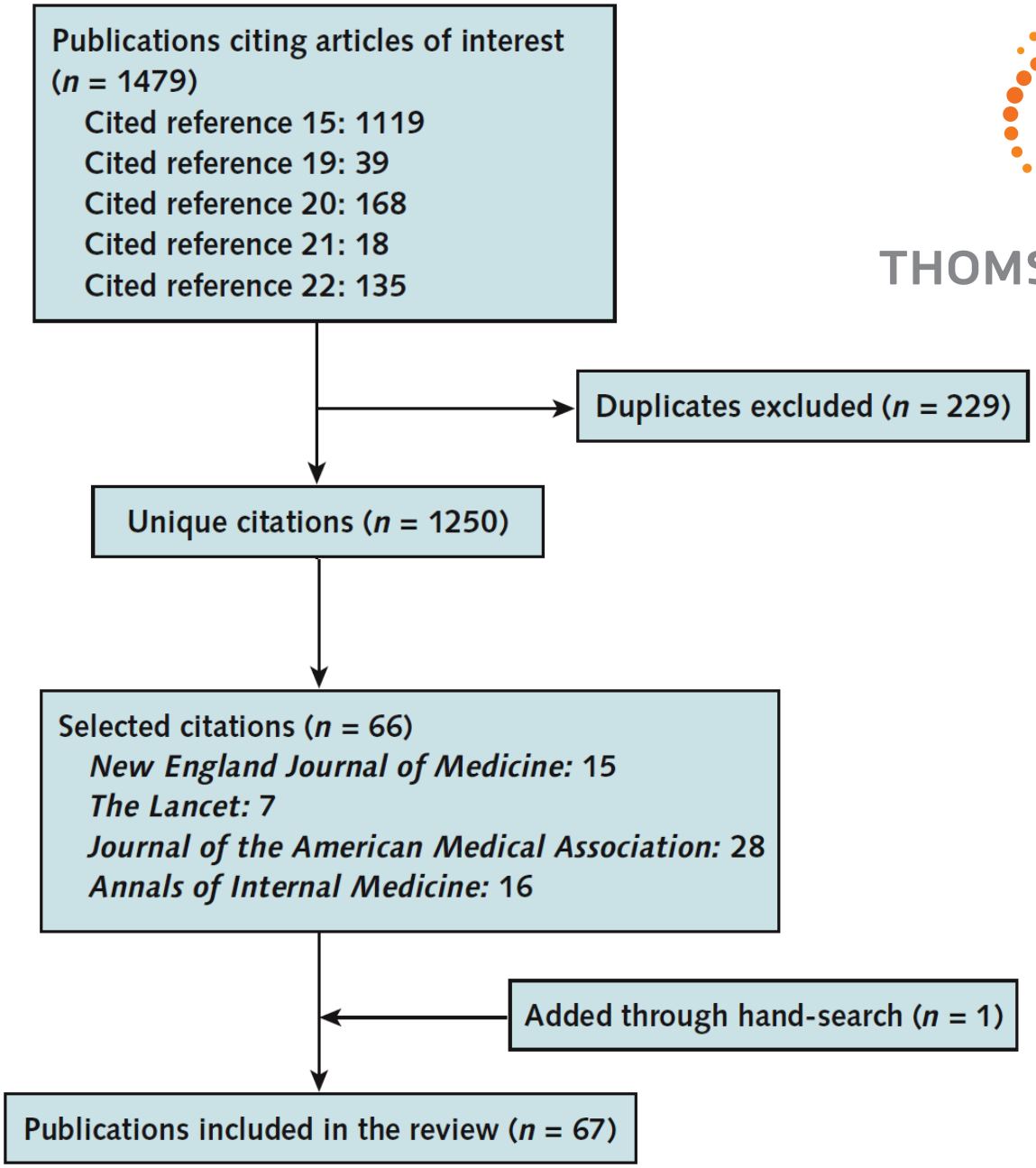
Erasmus MC
Universitair Medisch Centrum Rotterdam

RESEARCH AND REPORTING METHODS | Annals of Internal Medicine

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



THOMSON REUTERS

Erasmus MC

Leening et al., Ann Intern Med 2014;160:122-31

Reporting of NRI Feature	Studies, %
Risk categorization	
Categorization for computing NRI justified in text	27
Reference given for NRI categorization	38
Categorization for computing NRI corresponded to diagnostic or therapeutic implications in clinical guidelines	11
Unit	
Reported as a percentage	67
Interpreted as a percentage or proportion	22

Leening et al., Ann Intern Med 2014;160:122-31

Recommendations

- Methods
  - incomplete follow-up
  - meaningful categories
- Results
  - focus on components
- Discussion
  - interpretation

Methods	
Type of NRI	Specify the type of NRI computed in the methods section of the manuscript (category-based and/or continuous NRI).
Follow-up	Specify the horizon of risk prediction if the NRI was computed for prognostic evaluations (e.g., 10-y risk). Describe how censored observations (e.g., persons lost to follow-up before the specified horizon) were handled. Use the event status at the predicted time horizon and ignore events occurring beyond the predicted time horizon (e.g., when predicting 10-y risk for CHD, consider participants with a myocardial infarction occurring after 10 y of follow-up as nonevents).
Cutoffs	For category-based NRI, the categorization should ideally have clear consequences in clinical practice. When possible, give references to formal clinical guidelines used to define the risk categories for the computation of the NRI. If alternative cutoffs were used, clearly motivate them.
Results	
Components	Report the NRIs for events and nonevents separately. Reclassification tables stratified for persons with and without the event of interest are informative beyond the NRI
Unit	The event and nonevent NRIs can be presented as percentages. However, the overall NRI has no units and should therefore not be presented as a percentage.
Calibration	Provide information on the calibration of the models being compared.
Discussion	
Interpretation	The components of the overall NRI can be interpreted as a net percentage of the number of persons with or without events. However, the overall NRI should not be interpreted as a net percentage of the study population correctly reclassified.
Comparisons	Do not draw strong comparative conclusions based on direct comparisons of NRIs obtained in different populations or using different outcomes or cutoffs.

Leening et al., Ann Intern Med 2014;160:122-31

Net reclassification risk graph (*Biom J 2014*)

Steyerberg et al., 2013: in preparation

Further criticism on NRI

- Statistical properties (Gerds, Hilden, Pepe)
  - Misleading with miscalibrated predictions
  - Too high values and low p-values under H0

- Better alternatives (Vickers, Baker, van Calster, ..)
  - Decision-analytic = utility respecting

Net Reclassification Indices for Evaluating Risk Prediction Instruments

A Critical Review

Erasmus MC

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?

Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective

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

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*,†

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 Δ 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.

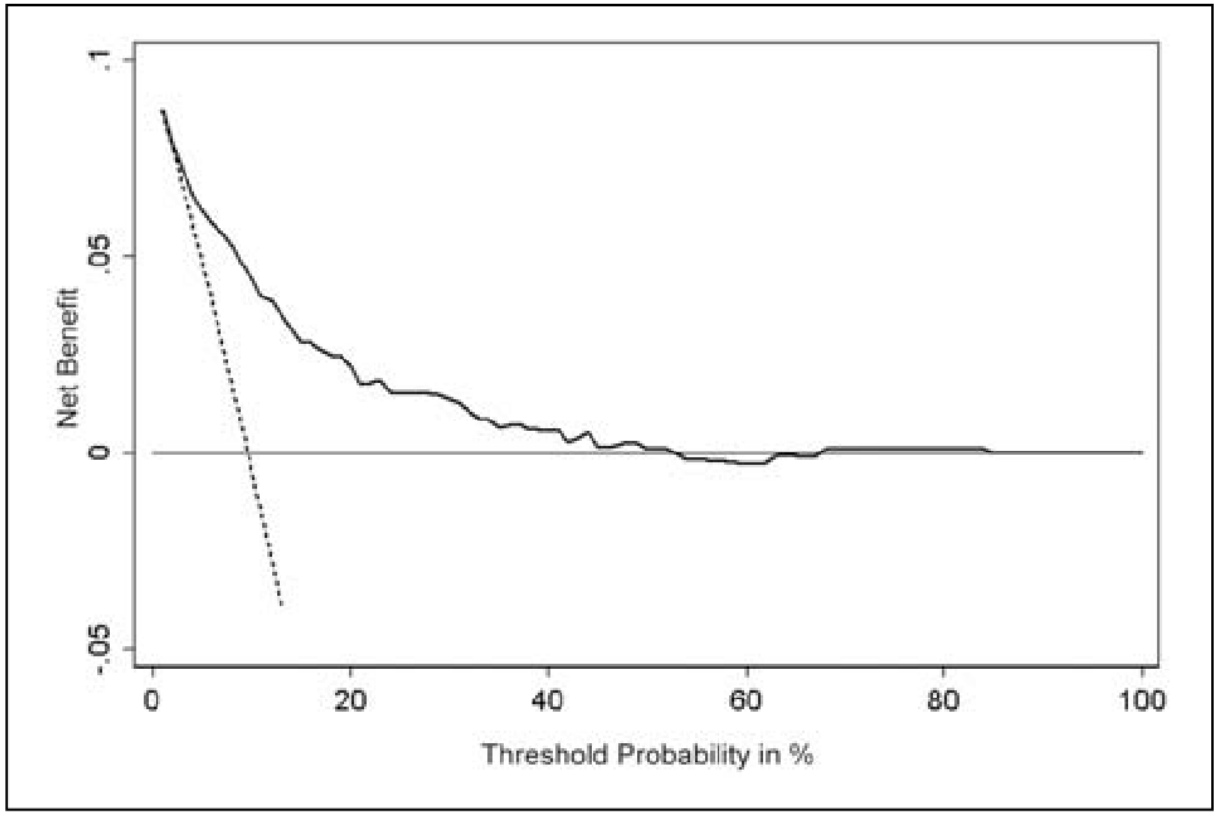
Decision-analytic

- Decision analytic, e.g. Net Benefit in Decision Curve
  - Net Benefit = (TP – w FP) / N
    w = threshold probability / (1-threshold probability)
    e.g.: threshold 50%: w = .5/.5=1; threshold 20%: w=.2/.8=1/4
  - Number of true-positive classifications,
    penalized for false-positive classifications
- Choosing a cut-off on the probability scale implies a relative weight of TP vs FP, or harm vs benefit ; and vice versa
  (Peirce Science 1884; Pauker NEJM 1975;
  Localio AR, Goodman S: Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med 2012)

Decision curve: theory (MDM 2006)

Decision Curve Analysis: A Novel Method for Evaluating Prediction Models

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



Conclusions

- Evaluation of incremental value of a marker by the NRI:
  - Methodological limitations
    - → Only for calibrated models, at model development
  - Reporting limitations
    - → Clinically meaningful risk categories
    - → Time horizon and handling of incomplete follow-up
    - → NOT be interpreted as a percentage
  - Rightly points at net reclassification
    - → Report NRI for events and non-events separately
    - → Reclassification table → decision-analytic measures
      - NRI: easy interpretation, but wrong
      - Decision-analytic: more difficult interpretation, but proper

Limitations of utility-respecting measures

- Harm to benefit ratio may be uncertain (no evidence, opinion driven, patient preferences) → consider a range
- Definition of ‘important gain’ remains subjective;
  formal cost-effectiveness at the end of evaluation pyramid