# Assessment of performance and decision curve analysis

Ewout Steyerberg, Andrew Vickers

Dept of Public Health, Erasmus MC, Rotterdam, the Netherlands

Dept of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, USA

Freiburg, October 2008

Erasmus MC

## Work from Freiburg

BIOINFORMATICS

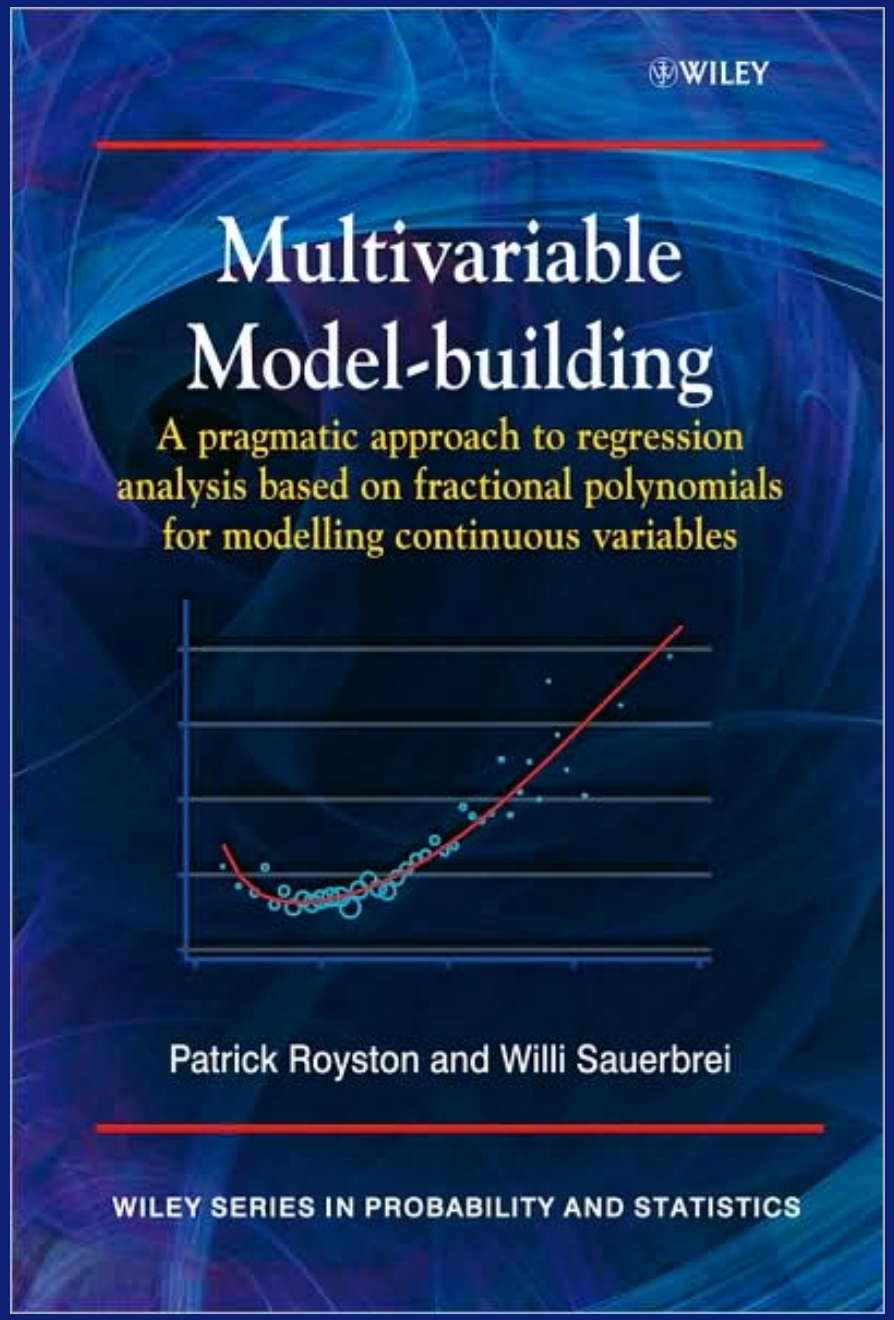
# Assessment of survival prediction models based on microarray data

Martin Schumacher [a] ,* Harald Binder [b], Thomas Gerds [b]

[a]Department of Medical Biometry and Statistics, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

[b] Freiburg Center of Data Analysis and Model Building, University Freiburg, Germany

Erasmus MC


WILEY
Multivariable Model-building
A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables
Patrick Royston and Willi Sauerbrei
WILEY SERIES IN PROBABILITY AND STATISTICS

Erasmus MC

# Erasmus MC – University Medical Center Rotterdam

Erasmus MC

## Some issues in performance assessment

- Usefulness / Clinical utility: what do we mean exactly?
  - Evaluation of predictions
  - Evaluation of decisions

- Usefulness of a marker
  - Challenges in design and analysis
  - Measurement worth the increase in complexity (physician burden) and worth the costs (patient burden, financial costs)?
  - Additional value to a model with free / easy to obtain predictors
  - Validity of the model w/o marker

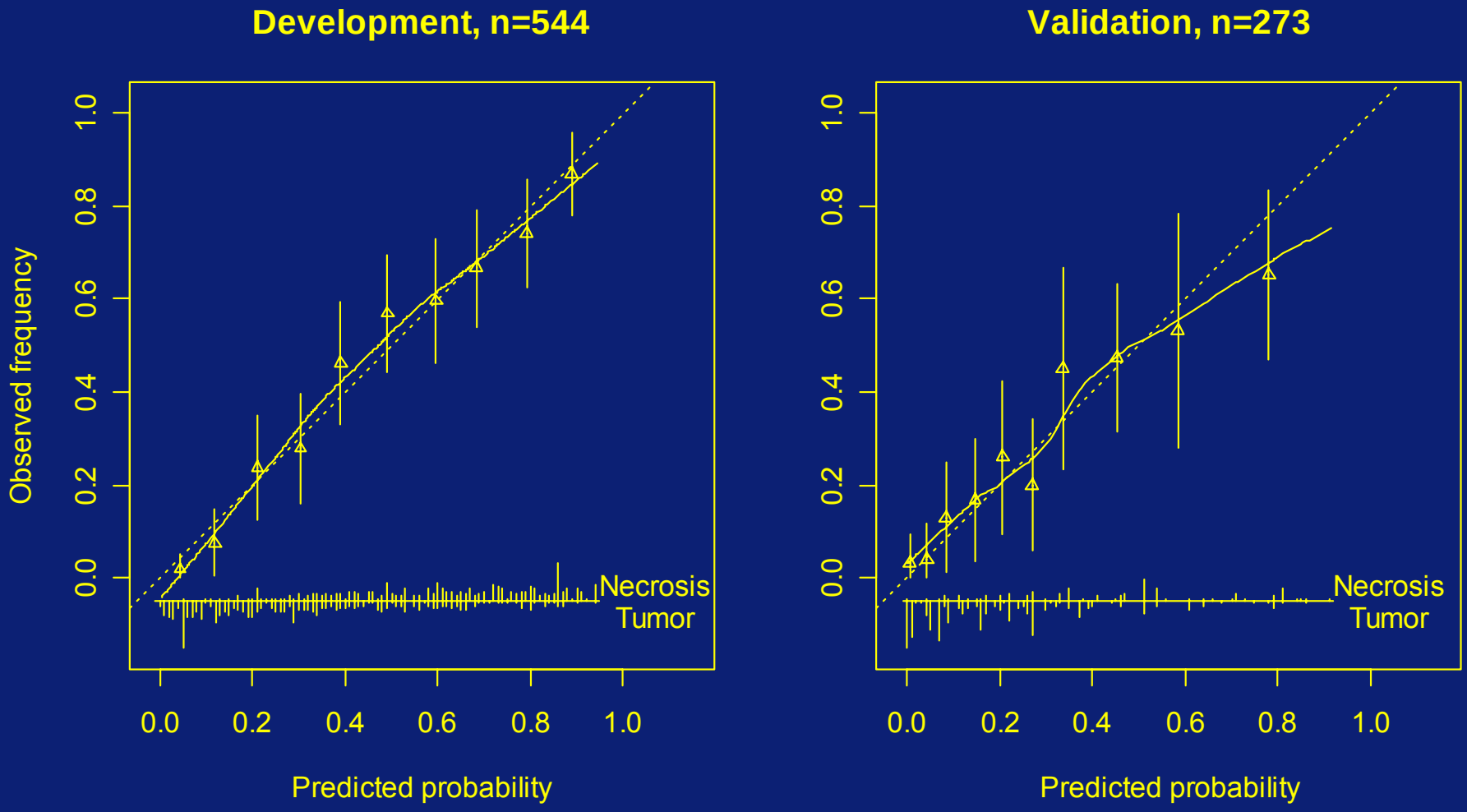
Erasmus MC

# Traditional performance evaluation of predictions

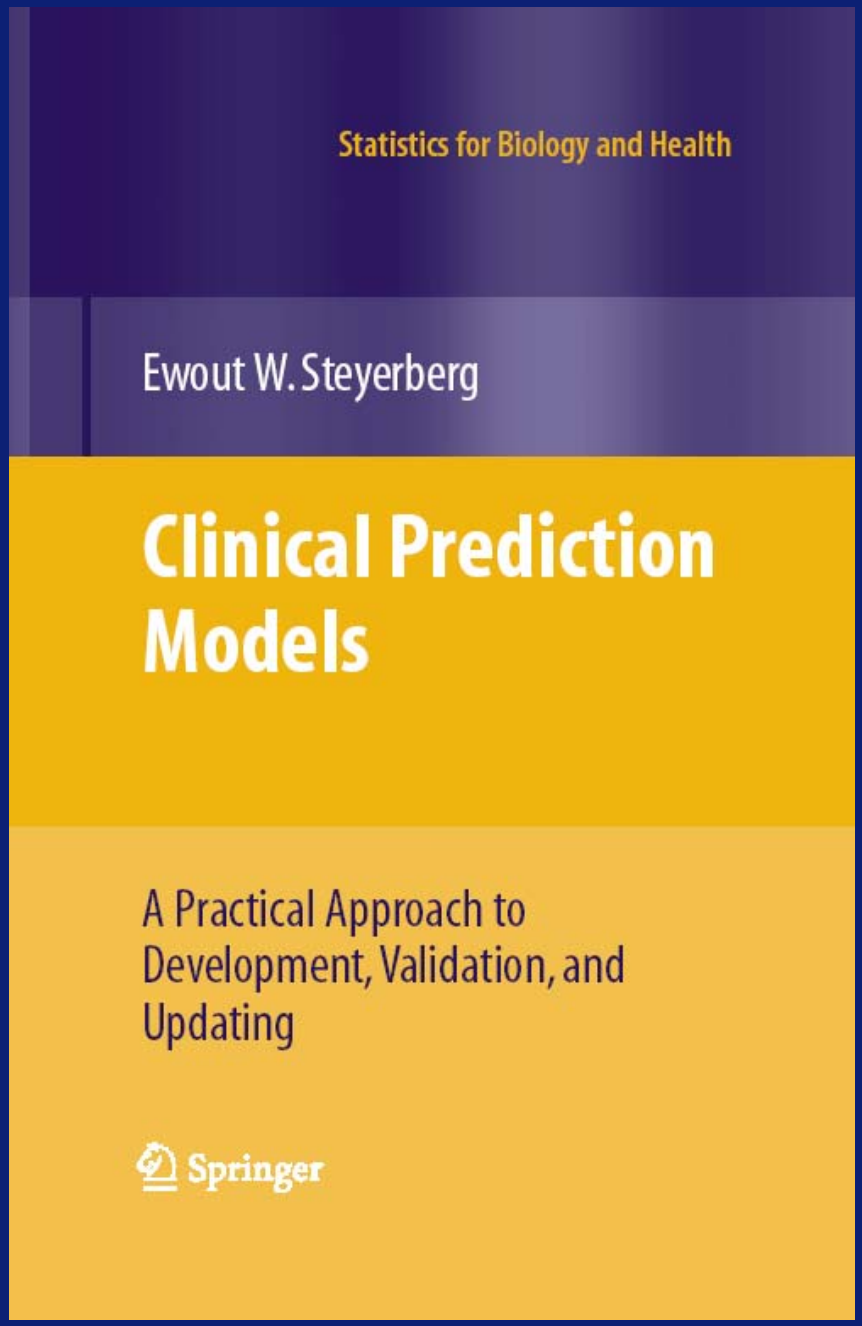
- Predictions close to observed outcomes?
  - Overall; consider residuals y – ŷ, or y – *p*
    - Brier score
    - $R^2$ (e.g. on log likelihood scale)
  - Discrimination: separate low risk from high risk
    - Area under ROC (or *c* statistic)
  - Calibration: e.g. 70% predicted = 70% observed
    - Calibration-in-the-large
    - Calibration slope

Erasmus MC

# Validation graph to visualize both calibration, discrimination, and usefulness




Erasmus MC


Statistics for Biology and Health
Ewout W. Steyerberg
Clinical Prediction Models
A Practical Approach to Development, Validation, and Updating
Springer

Erasmus MC

# Quantification of performance .. many developments

Erasmus MC

# Brier score for model performance

BIOINFORMATICS

## Assessment of survival prediction models based on microarray data

Martin Schumacher [a] ,* Harald Binder [b], Thomas Gerds [b]

[a]Department of Medical Biometry and Statistics, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

[b] Freiburg Center of Data Analysis and Model Building, University Freiburg, Germany

Erasmus MC

# Addition of a marker to a model

- Typically small improvement in discriminative ability according to *c* statistic
- *c* stat blamed for being insensitive

**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.

Erasmus MC

# Editorial JNCI, July 16, 2008 on paper by Gail

## Gauging the Performance of SNPs, Biomarkers, and Clinical Factors for Predicting Risk of Breast Cancer

Margaret S. Pepe, Holly E. Janes

Predicting risk of cancer for individuals has long been a goal of medical research. If an individual's risk could be predicted, then prevention and screening modalities could be targeted toward those at meaningfully high risk. This approach is not only more cost efficient than targeting the whole population but also more ethical, at least when interventions are burdensome to the individual. The quest for risk predictors has been revitalized with the emergence of technologies that measure genetic information and other molecular and physiological attributes of the individual. In this issue of the Journal, Gail (1) asks to what extent newly discovered associations between seven single-nucleotide polymorphisms

0.05. Although an extreme example, it illustrates the point. These two criticisms of AUC apply generally, not solely to risk prediction. The AUC really is a poor metric for evaluating markers for disease diagnosis, screening, or prognosis. The third criticism, which is specific to risk prediction, is that the AUC, and indeed the ROC curve itself, hides the values of risk calculated for subjects in the population. Indeed, the risk values are not visible from the ROC curve or the related curves in figure 2 of Gail (1). Moreover, the same ROC curve results if risk values are transformed monotonically, say, multiplied by a factor of 10, yet the clinical implications of these risk values would be very different.

Erasmus MC

## Alternatives to ROC analysis

Without harm – benefit: *Stat Med* 2008: 27:157–172;

see S. Greenland commentary

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

Michael J. Pencina[1,*,†], Ralph B. D'Agostino Sr[1], Ralph B. D'Agostino Jr[2] and Ramachandran S. Vasan[3]

*Am J Epidemiol* 2008;167:362–368

**Practice of Epidemiology**

**Integrating the Predictiveness of a Marker with Its Performance as a Classifier**

**Margaret S. Pepe[1,2], Ziding Feng[1], Ying Huang[2], Gary Longton[1], Ross Prentice[1], Ian M. Thompson[3], and Yingye Zheng[1]**

## Alternatives to ROC analysis

With harm – benefit: *Biostatistics* (2008), in press

# Estimating the capacity for improvement in risk prediction with a marker

WEN GU, MARGARET SULLIVAN PEPE*

### 5.4 *Risk thresholds and decisions to ascertain Y*

Implicit in the above discussion is the existence of threshold values for risk that are used to make decisions, in this case for or against the renal arteriography procedure. Risk thresholds vary with the clinical context and may additionally vary among individuals. How to choose a risk threshold? The classic decision theoretic solution to choosing a risk threshold is fairly simple. Let $C_0$ (and $B_0$) denote the cost

### ACKNOWLEDGMENTS

We thank Dr A. Cecile J. W. Janssens for allowing us to use the renal artery stenosis data to illustrate our methodology and Dr Patrick Bossuyt for very helpful suggestions.

## Contents

1. Developments in performance evaluation
   predictions / decisions

2. Evaluation of clinical usefulness
   A. Binary marker / test
   B. Additional value of a marker

3. Further developments

Erasmus MC

# Example: Binary markers / tests

- 2 uncorrelated binary markers with equal costs
  - 50% and 10% prevalence, outcome incidence 50%
  - Odds Ratio 4 and 16
  - Evaluate as single test

| | Test 1 | Test 2 |
|---|---|---|
| C stat | 0.67 | 0.59 |
| Brier | 0.22 | 0.23 |
| $R^2$ | 15% | 13% |

- Any role for test 2?

Erasmus MC

# Decision threshold and relative costs

Treatment: Risk >= cutoff
No treatment: Risk < cutoff

| | Event | No event |
|---|---|---|
| Treatment: Risk >= cutoff | cTP | cFP |
| No treatment: Risk < cutoff | cFN | cTN |

cTP and cFP: costs of True and False Positive classifications;
cFN and cTN: costs of False and True Negative classifications respectively.

- Optimal cutoff:
- Odds(cutoff) = (cFP – cTN) / (cFN – cTP)
  = harm / benefit

Erasmus MC

# Simple usefulness measures given 1 cutoff

- Naïve: Unweighted

  Sensitivity = TP / (TP+FN); Specificity = TN / (FP+TN)

  Accuracy: (TN+TP)/N; Error rate: (FN+FP)/N

# Example

- 2 uncorrelated binary markers with equal costs
  - 50% and 10% prevalence, 50% outcome incidence
  - Odds Ratio 4 and 16
  - Evaluate as single test

| | Test 1 | Test 2 |
|---|---|---|
| C stat | 0.67 | 0.59 |
| Brier | 0.22 | 0.23 |
| $R^2$ | 15% | 13% |

- Any role for test 2 alone?

| | | |
|---|---|---|
| **Sens** | **67%** | **18.8%** |
| **Spec** | **67%** | **98.7%** |

Erasmus MC

# Simple usefulness measures given 1 cutoff

- Naïve: Unweighted

  Sensitivity = TP / (TP+FN); Specificity = TN / (FP+TN)

  Accuracy: (TN+TP)/N; Error rate: (FN+FP)/N

- Weighted variants

  Weighted accuracy: $(TP + w\,TN) / (N_{Event} + w\,N_{No\ event})$ (Vergouwe 2002)

  Net Benefit: $(TP - w\,FP) / N$,

  with $w$ = harm / benefit (Pierce 1884, Vickers 2006)

Erasmus MC

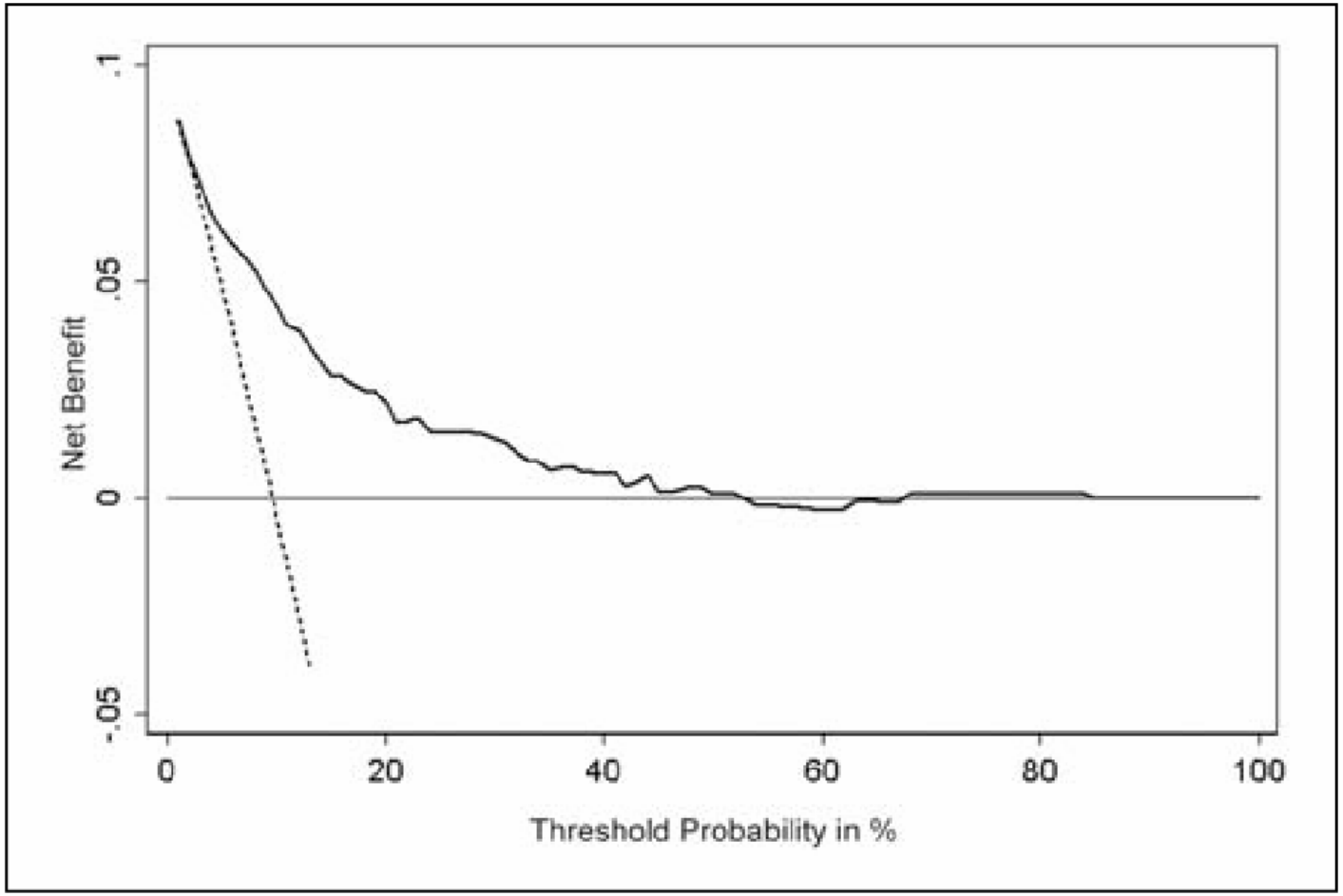
# From 1 cutoff to consecutive cutoffs

- Sensitivity and specificity → ROC curve
- Net benefit → decision curve

Erasmus MC

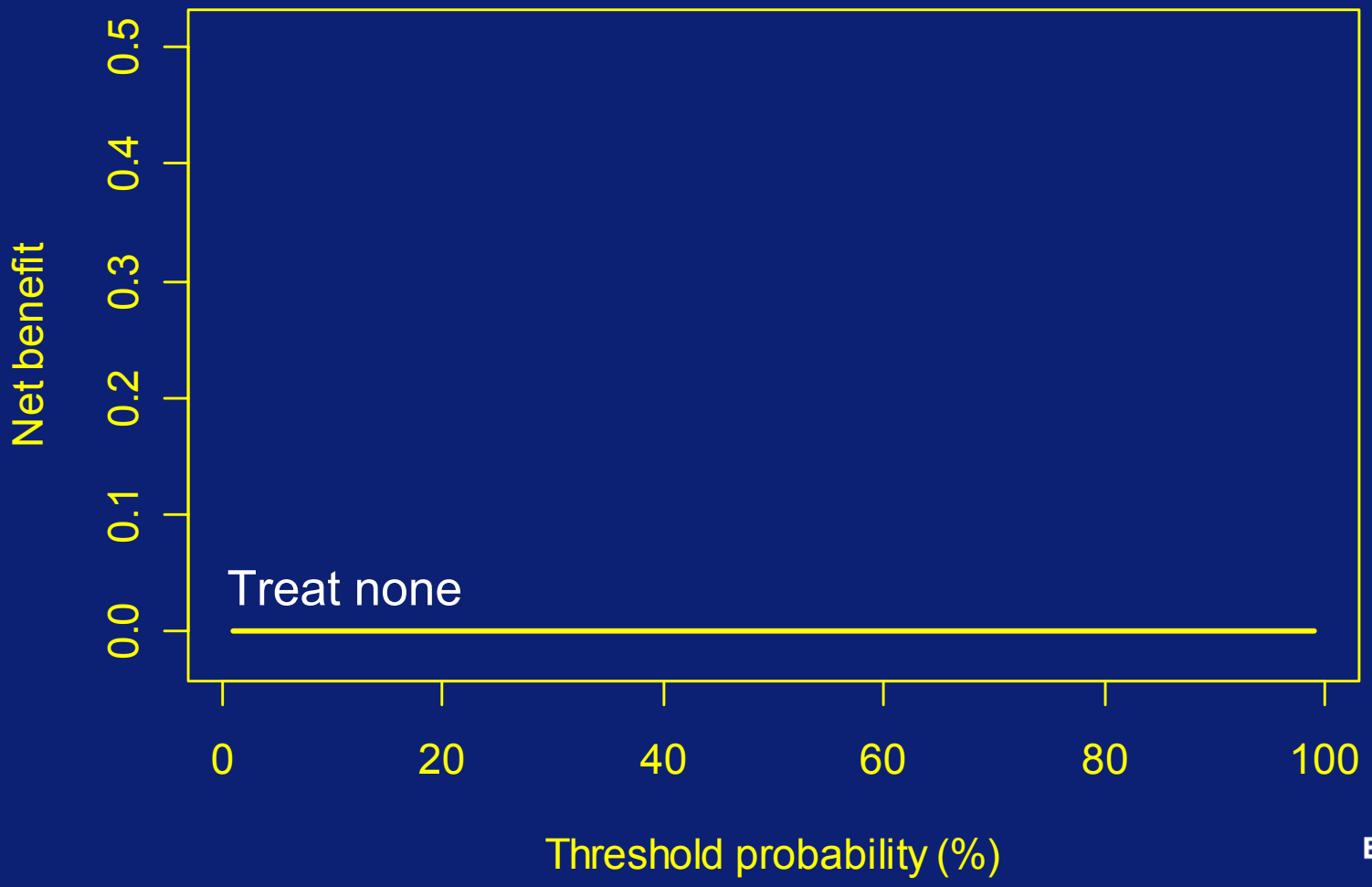
# Decision Curve Analysis: A Novel Method for Evaluating Prediction Models

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

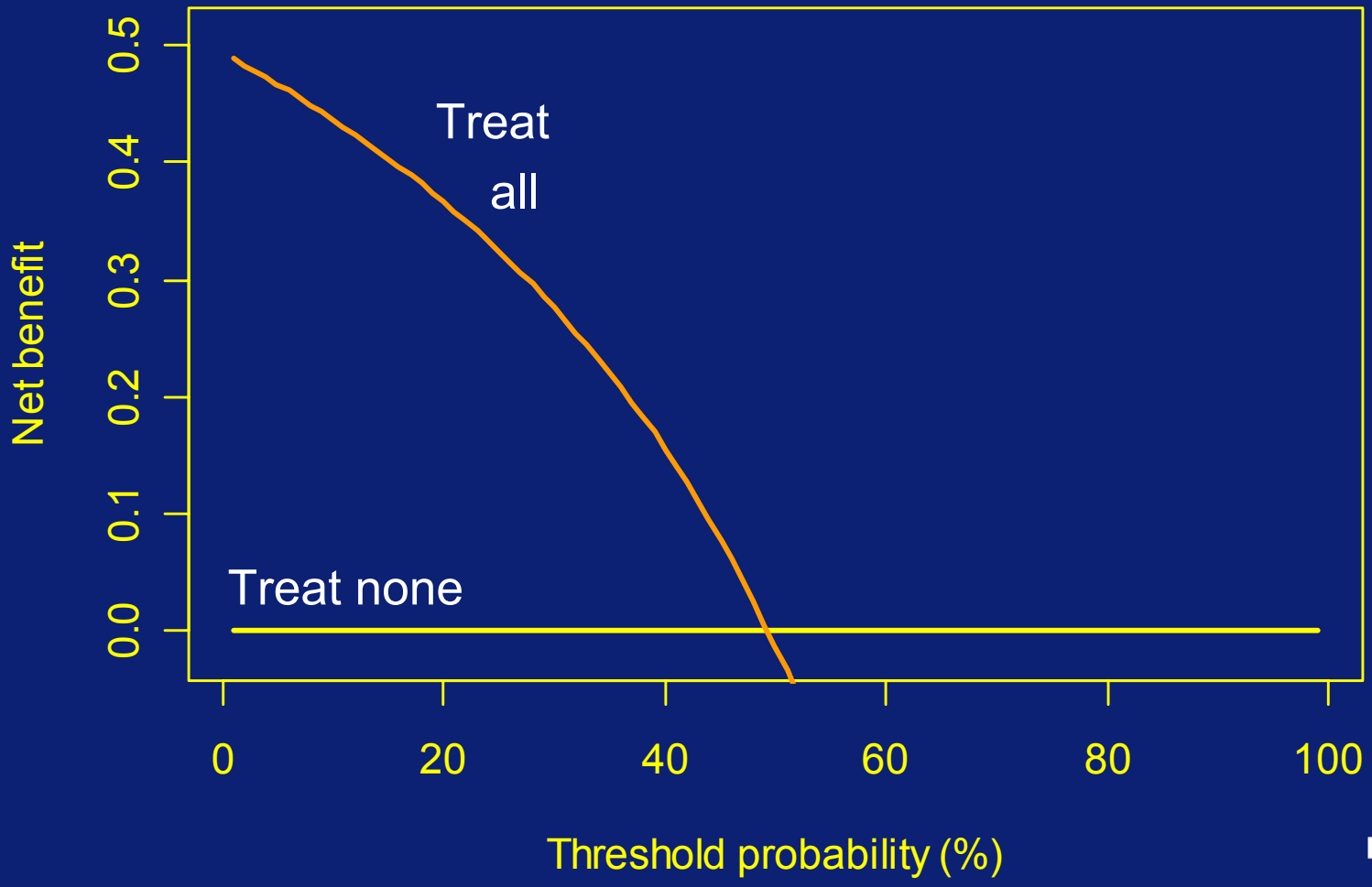



Erasmus MC

# Treat none

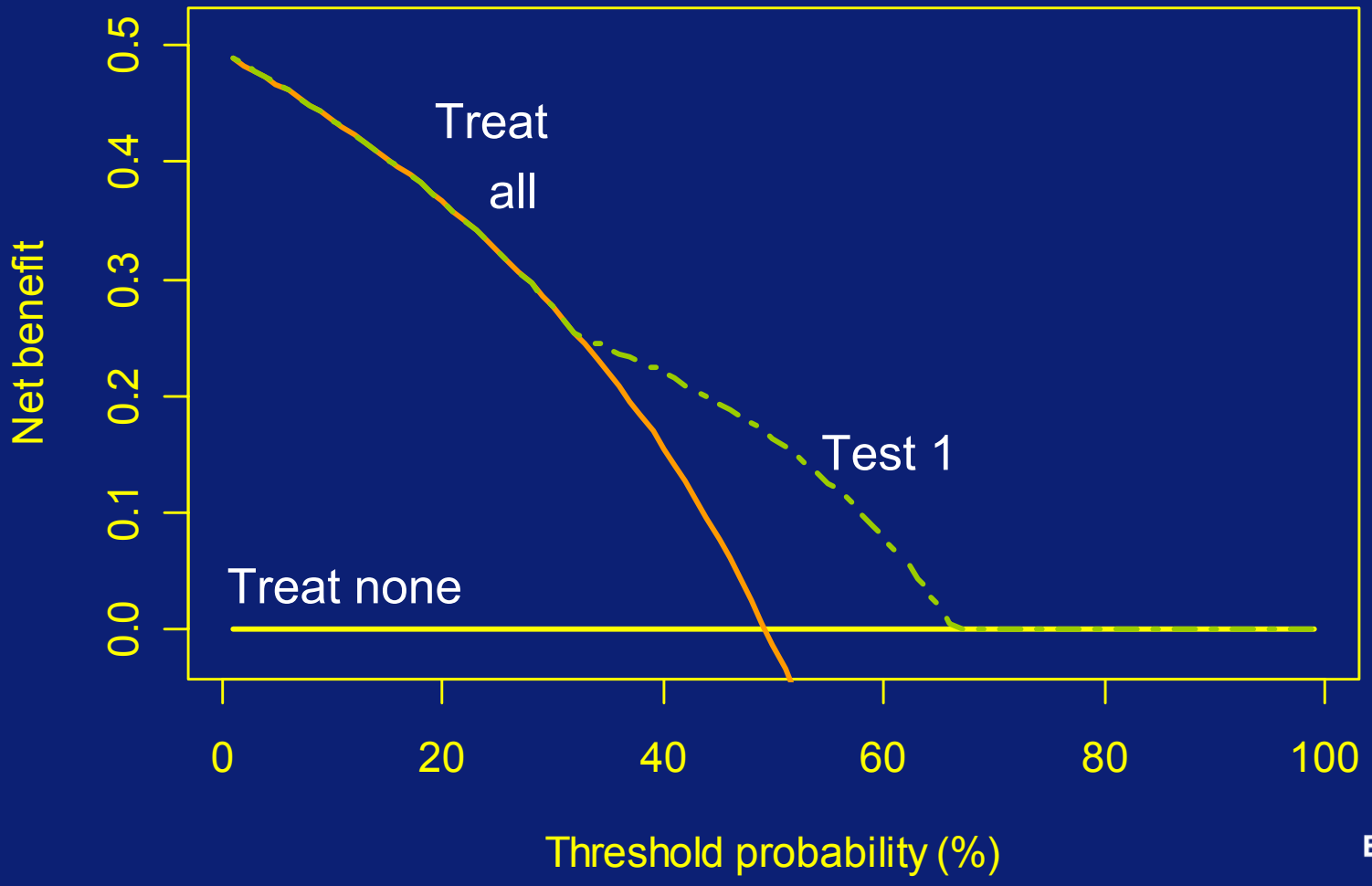



Erasmus MC


Net benefit
0.0
0.1
0.2
0.3
0.4
0.5
Treat
all
Treat none
0
20
40
60
80
100
Threshold probability (%)

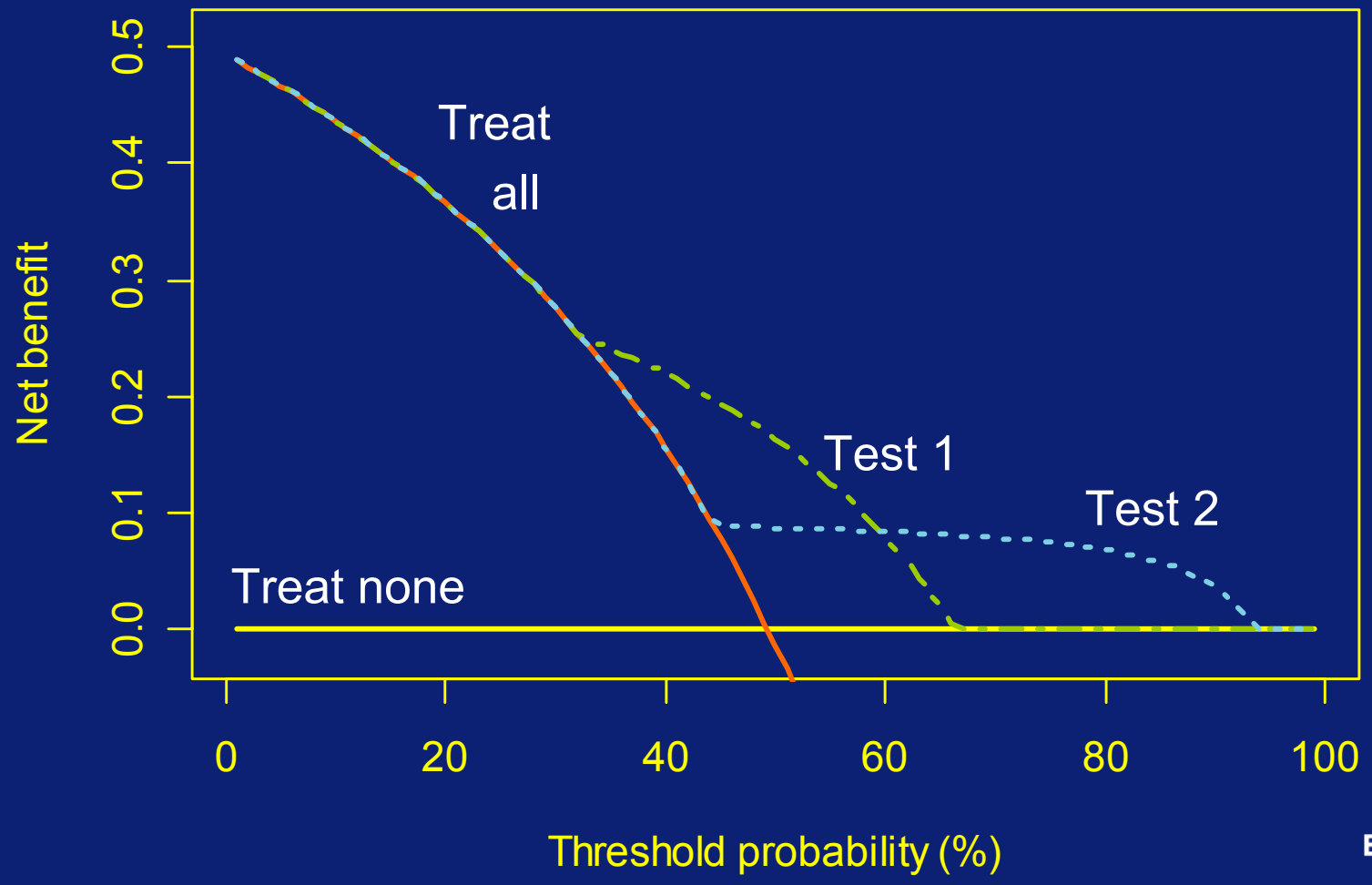
Erasmus MC

# Decision curve for example: Test 1 alone




Erasmus MC

# Decision curve for example: Test 1 and test 2 each alone




Erasmus MC

# Addition of a marker to a model

Erasmus MC

## Additional value of a marker: prostate cancer

- Men with elevated PSA are referred to prostate biopsy
- Only 1 in 4 men with high PSA have prostate cancer
- Could an additional marker help predict biopsy result?
  - Free PSA (a subfraction of PSA)
  - PSA velocity (measure change in PSA)
- Assume decision threshold around 20%

Erasmus MC

## Data set

- Data from European Randomized Study of Prostate Cancer screening (ERSPC)
- 2742 previously screened men with:
  - Elevated PSA
  - No previous biopsy
- 710 cancers (26%)

Erasmus MC

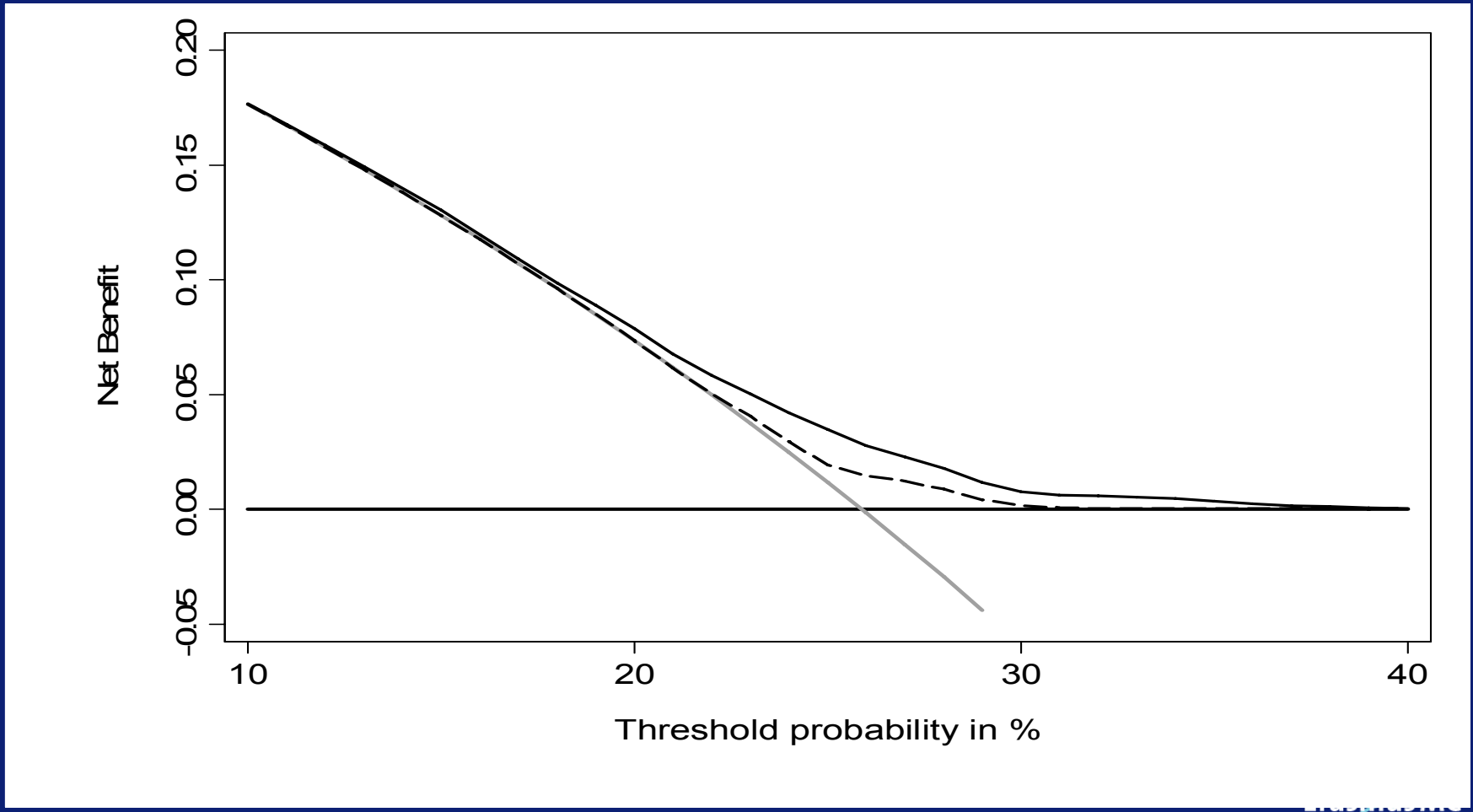
# Accuracy metrics

| Model | Sens.* | Spec.* | PPV* | NPV* | Brier | AUC | NRI |
|---|---|---|---|---|---|---|---|
| PSA only | 100 | 0 | 26 | 0 | .191 | .544 | |
| + PSA velocity | 95 | 10 | 27 | 86 | .189 | .580 | .053 |
| + Free PSA | 98 | 4 | 26 | 84 | .186 | .592 | .018 |
| + Free PSA & PSA velocity | 95 | 8 | 27 | 83 | .184 | .610 | .037 |

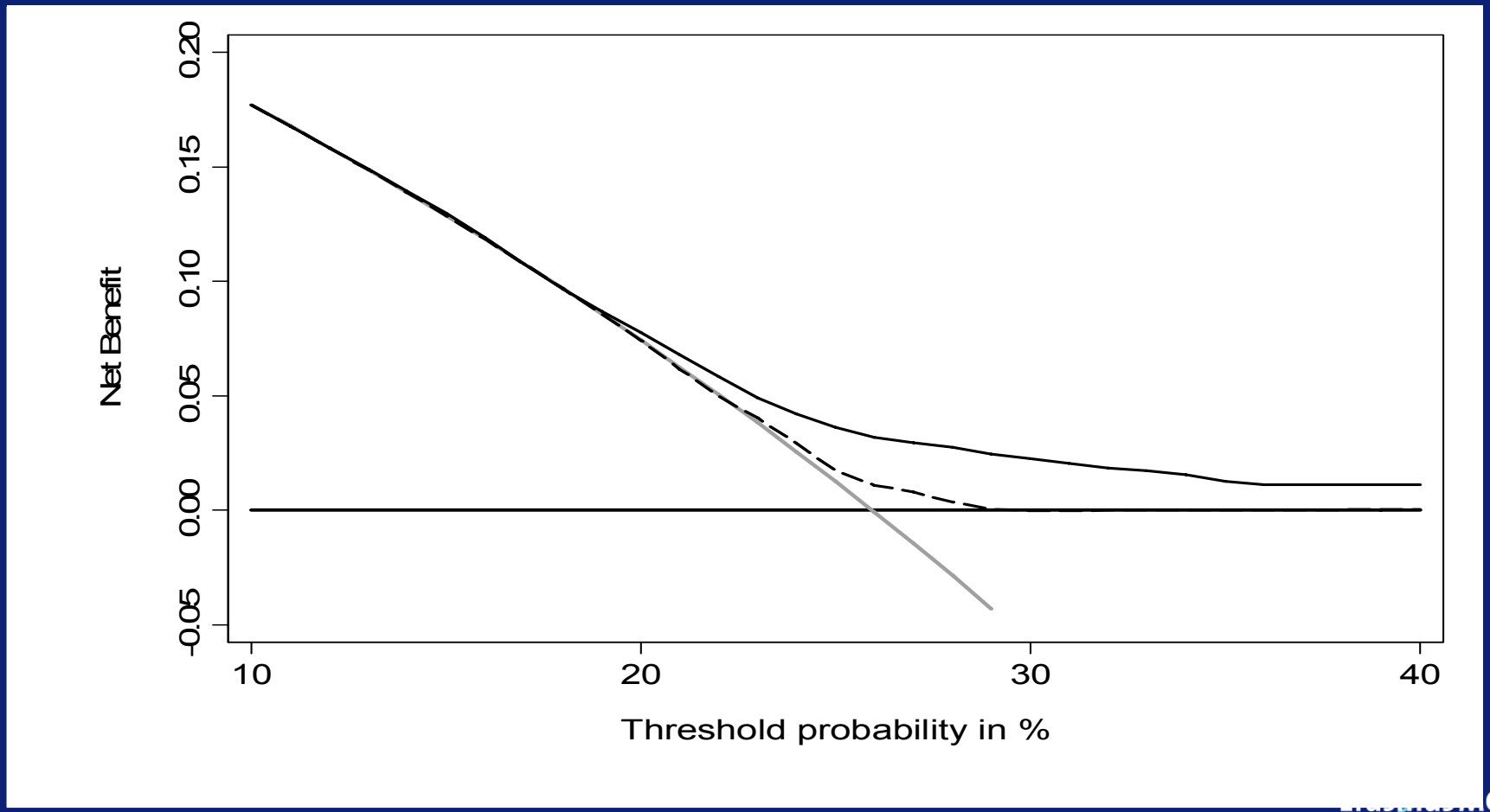
* At Risk threshold of 20%

Erasmus MC

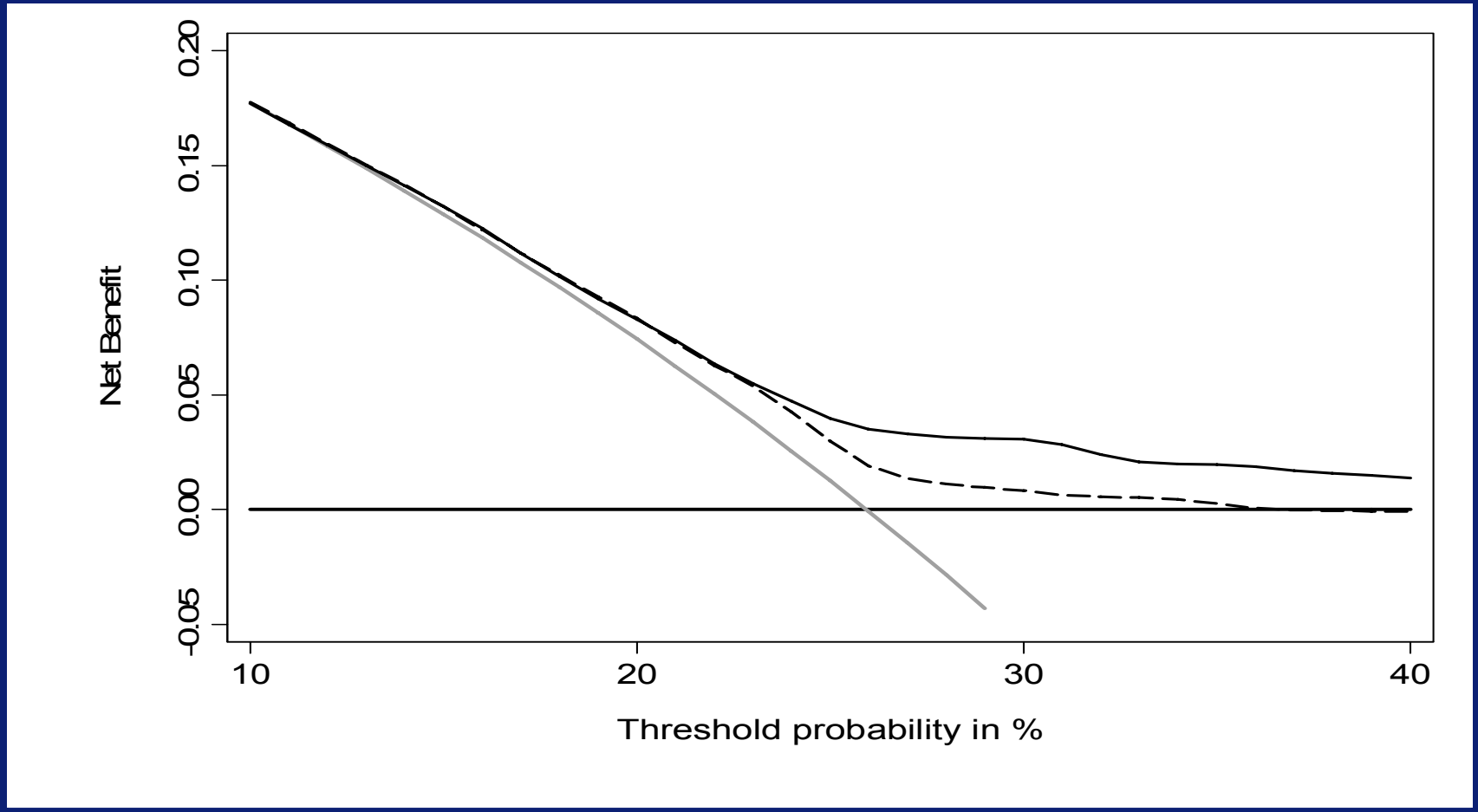
# Add PSA velocity to base model?



# Add free PSA to base model?



# Does Free PSA add anything if velocity included?




Erasmus MC

# Accuracy metrics

| Model | Sens.* | Spec.* | PPV* | NPV* | Brier | AUC | NRI |
|---|---|---|---|---|---|---|---|
| PSA only | 100 | 0 | 26 | 0 | .191 | .544 | |
| + PSA velocity | 95 | 10 | 27 | 86 | .189 | .580 | .053 |
| + Free PSA | 98 | 4 | 26 | 84 | .186 | .592 | .018 |
| + Free PSA & PSA velocity | 95 | 8 | 27 | 83 | .184 | .610 | .037 |

* At Risk threshold of 20%

Erasmus MC

# Which performance measure when?

| Application area | Calibration | Discrimination | Clinical usefulness |
|---|---|---|---|
| *Public health* | | | |
| Targeting of preventive interventions | | | |
| Predict incident disease | x | X | x |
| *Clinical practice* | | | |
| Diagnostic work-up | | | |
| Test ordering | X | x | X |
| Starting treatment | X | x | X |
| Therapeutic decision making | | | |
| Surgical decision making | X | x | X |
| Intensity of treatment | X | x | X |
| Delaying treatment | X | x | X |
| *Research* | | | |
| Inclusion in a RCT | X | x | X |
| Covariate adjustment in a RCT | | X | |
| Confounder adjustment with a propensity score | | | |
| Case-mix adjustment | | | |

1. Discrimination: if poor, usefulness unlikely, but NB >= 0
2. Calibration: if poor **in new setting**, risk of NB<0;
   Prediction model may harm rather than support decision-making

Erasmus MC

# Phases of marker evaluation (Pepe, *Stat Med* 2005;24(24):3687-96)

| *Phase* | *Objective* | *Study design* |
|---|---|---|
| 1 Preclinical exploratory | Promising directions identified | Case–control (convenient samples) |
| 2 Clinical assay and validation | Determine if a clinical assay detects established disease | Case–control (population based) |
| 3 Retrospective longitudinal | Determine if the biomarker detects disease before it becomes clinical. Define a 'screen positive' rule | Nested case–control in a population cohort |
| 4 Prospective screening | Extent and characteristics of disease detected by the test; false referral rate | Cross-sectional population cohort |
| 5 Cancer control | Impact of screening on reducing the burden of disease on the population | Randomized trial |

Erasmus MC

# Phases of model development (Reilly *Ann Intern Med* 2006;144(3):201-9)

| Level of evidence | Definitions and standards of evaluation | Clinical implications |
|---|---|---|
| Level 1<br>▪ Derivation of prediction model | ▪ Identification of predictors for multivariable model; blinded assessment of outcomes. | ▪ Needs validation and further evaluation before using in actual patient care. |
| Level 2<br>▪ Narrow validation of prediction model | ▪ Assessment of predictive ability when tested prospectively in 1 setting; blinded assessment of outcomes. | ▪ Needs validation in varied settings; may use predictions cautiously in patients similar to sample studied. |
| Level 3<br>▪ Broad validation of prediction model | ▪ Assessment of predictive ability in varied settings with wide spectrum of patients and physicians. | ▪ Needs impact analysis; may use predictions with confidence in their accuracy. |
| Level 4<br>▪ Narrow impact analysis of prediction model used as decision rule | ▪ Prospective demonstration in 1 setting that use of decision rule improves physicians' decisions (quality or cost-effectiveness of patient care). | ▪ May use cautiously to inform decisions in settings similar to that studied. |
| Level 5<br>▪ Broad impact analysis of prediction model used as decision rule | ▪ Prospective demonstration in varied settings that use of decision rule improves physicians' decisions for wide spectrum of patients. | ▪ May use in varied settings with confidence that its use will benefit pateint care quality or effectiveness. |

Erasmus MC

# Conclusions

- Evaluation of p(outcome) may include overall performance, discrimination and calibration aspects
  - Confusion: overall performance and discrimination measures can be interpreted as evaluation of decision-making
- Evaluation of quality of decision-making requires utility-based loss functions, such as decision-curves

- References


- Vickers AJ, Elkin EB: Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 26:565-74, 2006
- Steyerberg EW, Vickers AJ: Decision Curve Analysis: A Discussion. *Med Decis Making* 28; 146, 2008
- Steyerberg EW, et al: Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: analysis of individual patient data from six study groups. *J Clin Oncol* 13:1177-87, 1995
- Vergouwe et al: Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: a model update and multicentre validation with more than 1000 patients. *Eur Urol* 51: 424-32, 2007

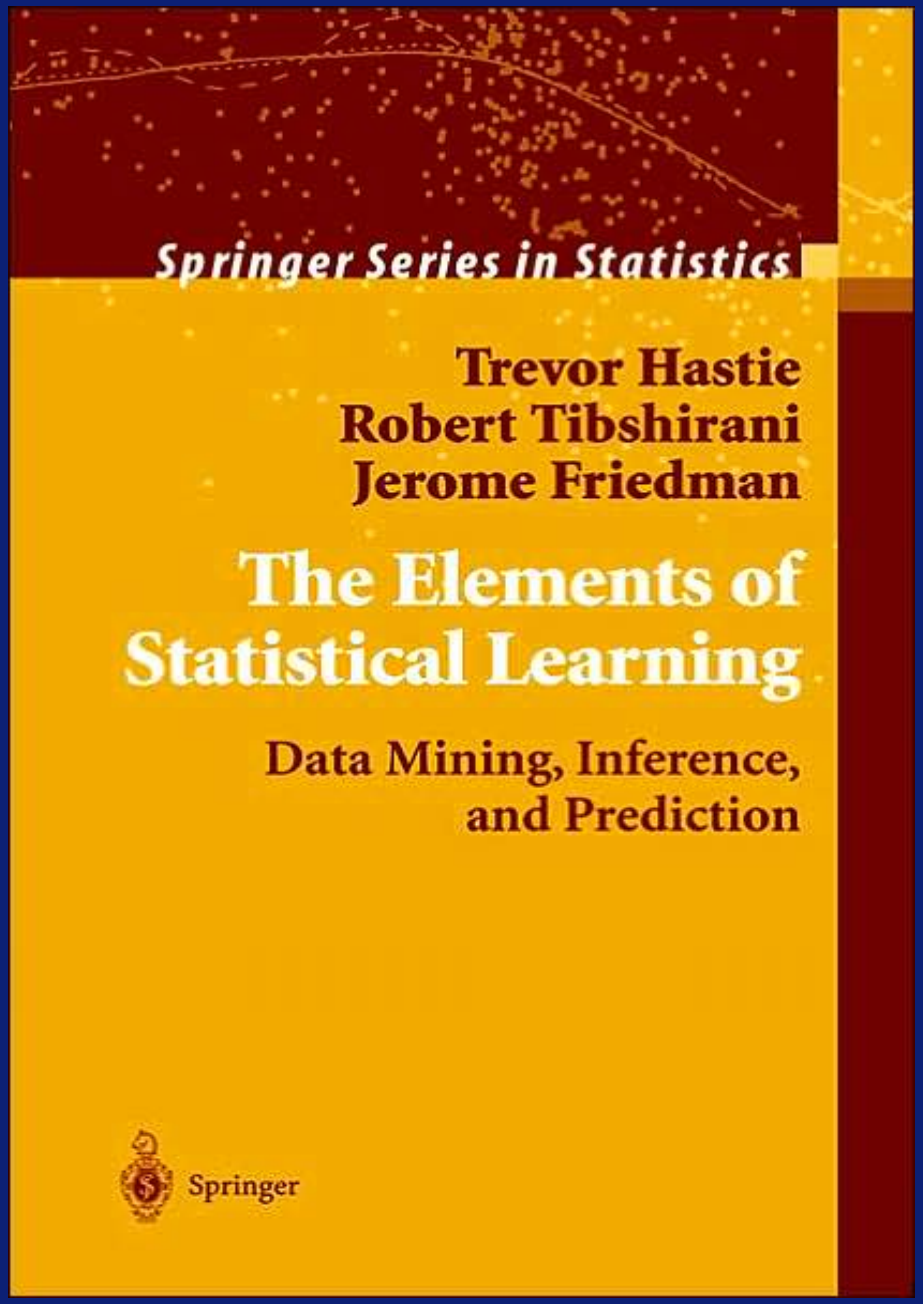

Erasmus MC

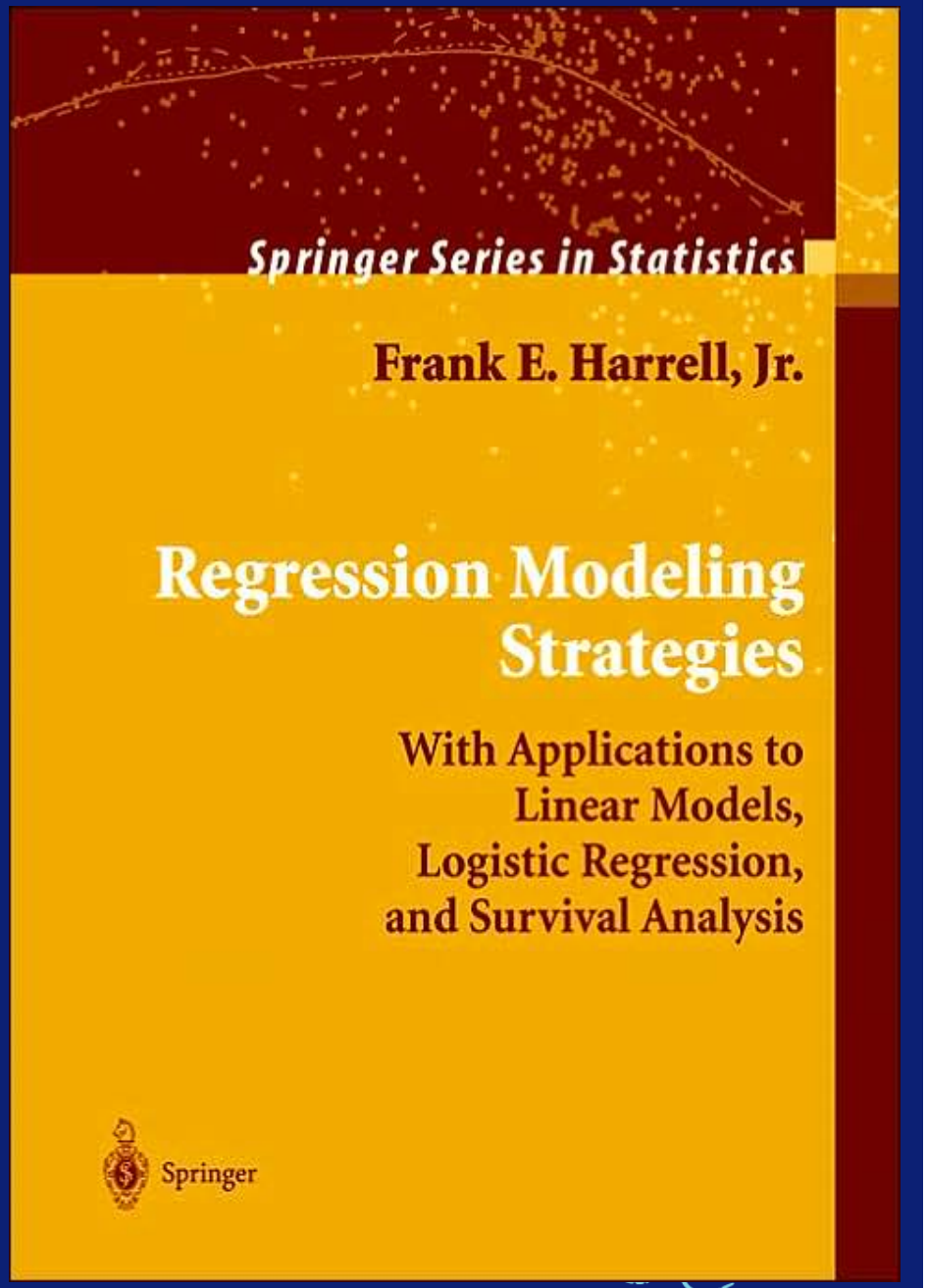
# Read more …

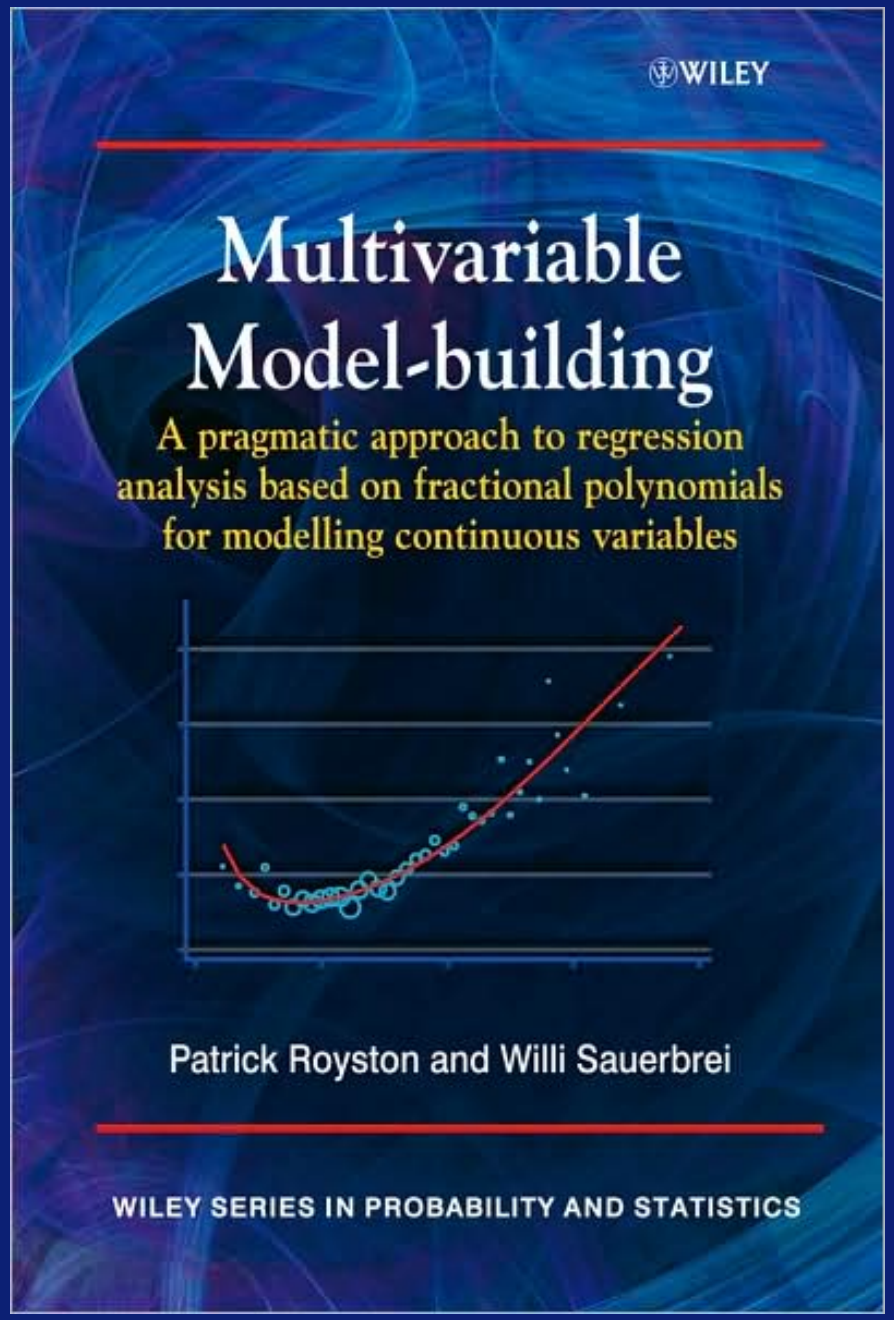
Books on prediction models

- Cost-effectiveness
  - Costs/page?
  - Costs/formula?
  - Costs/New information
  - Accessibility/Mathematical correctness
- 2 classics + 2 new

Erasmus MC


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Springer


Springer Series in Statistics
Frank E. Harrell, Jr.
Regression Modeling Strategies
With Applications to Linear Models, Logistic Regression, and Survival Analysis
Springer


WILEY
Multivariable Model-building
A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables
Patrick Royston and Willi Sauerbrei
WILEY SERIES IN PROBABILITY AND STATISTICS

Erasmus MC

E.Steyerberg@ErasmusMC.nl

Statistics for Biology and Health

Ewout W. Steyerberg

# Clinical Prediction Models

## A Practical Approach to Development, Validation, and Updating

Springer

# Thank you for your attention

Erasmus MC

# Comparison of performance measures

| Aspect | Measure | Development * | Validation |
|---|---|---|---|
| Overall performance | $R^2$ | 38% | 27% |
| | $Brier_{scaled}$ | 28% | 20% |
| Discrimination | C statistic | 0.81 | 0.79 |
| Calibration | Calibration-in-the-large | - | –0.03 |
| | Calibration slope | 0.97 | 0.74 |
| | Test for miscalibration | p=1 | p=0.13 |
| Clinical usefulness cutoff 30% | Accuracy | 69% | 75% |
| | Net Benefit – resection in all | 0.39 – 0.36 = 0.03 | 0.60 – 0.60 = 0 |

Erasmus MC