

Performance evaluation of predictive AI models to support medical decisions: overview and guidance

Ben Van Calster

KU Leuven | UMC Utrecht

KU LEUVEN

Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance

Ben Van Calster, Gary S Collins, Andrew J Vickers, Laure Wynants, Kathleen F Kerr, Lasai Barreñada, Gael Varoquaux, Karandeep Singh, Karel GM Moons, Tina Hernandez-Boussard, Dirk Timmerman, David J McLernon, Maarten van Smeden, Ewout W Steyerberg, on behalf of Topic Group 6 of the STRATOS initiative

Case study: ovarian tumor diagnosis

Aim

Externally validate the ADNEX model to estimate risk of malignancy of ovarian tumors
Support decision whether specialized surgery is needed (threshold 0.1)

External validation dataset

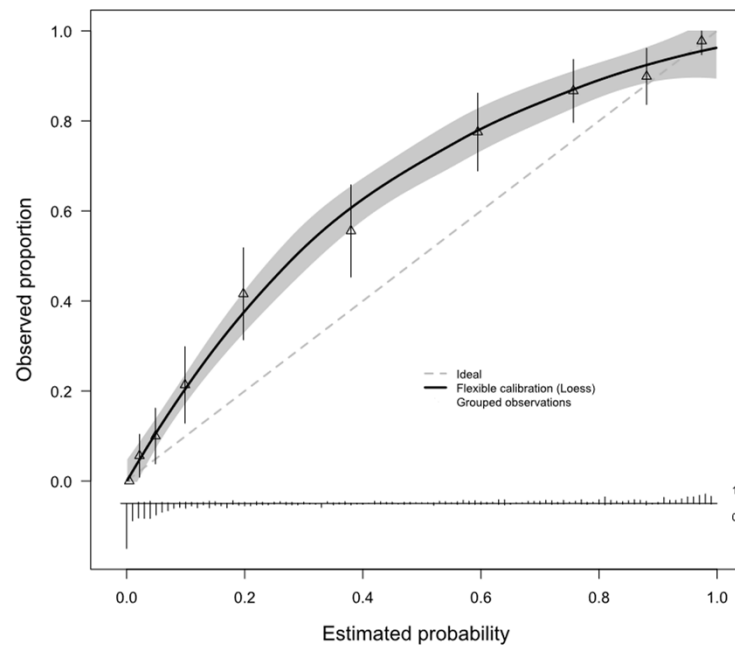
n=894, 434 malignancies (49%)

Validation

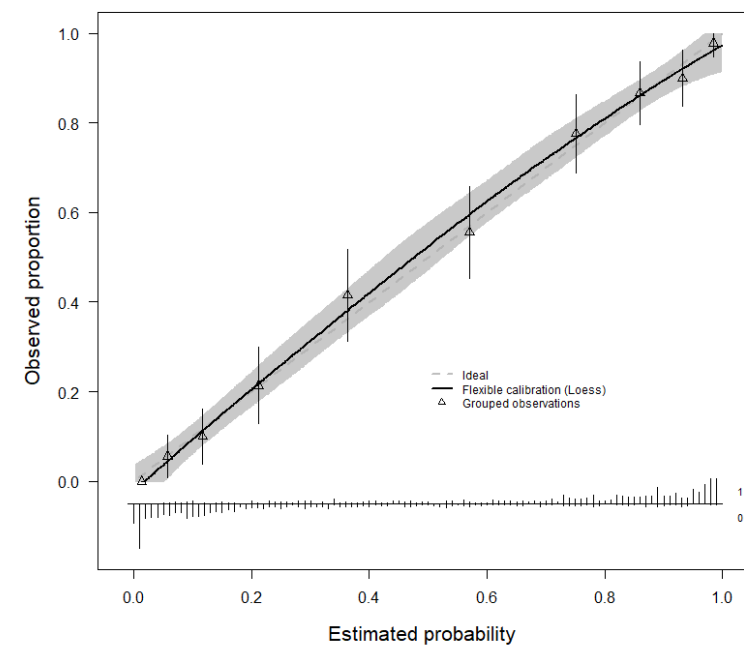
- The original ADNEX model
- An updated version to recalibrate the model; updating done on validation dataset

Calibration plots

Original



Updated



How to assess performance?

People often use measures based on classification of patients using a risk threshold t

Simplest measure: **classification accuracy**

How many patients are correctly classified (malignant and $p \geq t$, or benign and $p < t$)?

Original ADNEX: 79%

Updated ADNEX: 69%

How to assess performance?

Hot measure: **F1**

Harmonic mean of sensitivity and positive predictive value, aka precision and recall.

$$2 * \frac{precision * recall}{precision + recall}$$

Original ADNEX: 0.82

Updated ADNEX: **0.76**

How to assess performance?

How can these measures be worse for a superior model?

Context:

Validation of risk prediction models meant to be used in clinical practice
to support decisions for patients

Models with binary outcomes (but principles hold more generally)

Not:

Model development

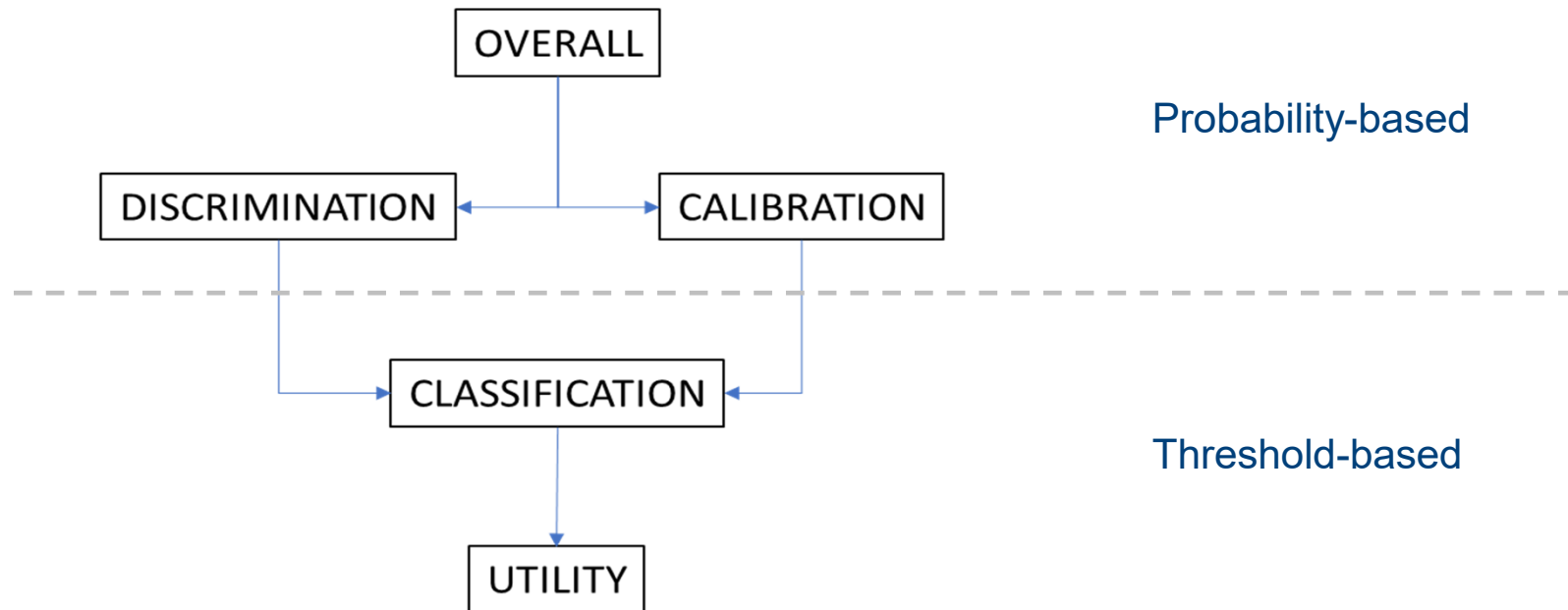
Model comparison (although mostly a natural extension)

Counterfactual prediction

5 performance domains

Domain	Target question and focus	Focus
<u>Probability-based</u>		
Discrimination	Does the model estimate higher probabilities in individuals with an event compared to individuals without an event?	Relative
Calibration	Do estimated probabilities correspond to observed event proportions?	Absolute
Overall	How close are estimated probabilities from the model (between 0 and 1) to actual outcomes (0 or 1)?	General
<u>Threshold-dependent</u>		
Classification	Are individuals classified correctly corresponding to their observed outcome?	Binary
Clinical utility	Do classifications lead to better clinical decisions?	Clinical

5 performance domains



Measures and plots

Domain	#	Measures	Plots
Discrimination	3	AUROC / c statistic AUPRC Partial AUROC	ROC curve PR curve
Calibration	6	O:E ratio Calibration intercept Calibration slope Estimated calibration index (ECI) Integrated calibration index (ICI) Expected calibration error (ECE)	Calibration plot
Overall	9	Loglikelihood Logloss (a.k.a. cross-entropy) Brier Scaled Brier (a.k.a. Brier Skill, IPA) McFadden R2 Cox-Snell R2 Nagelkerke R2 Discrimination slope MAPE	Risk distributions

Measures and plots

Domain	#	Measures	Plots
Classification	11	<u>SUMMARY MEASURES</u> (7) Classification accuracy Balanced accuracy Youden index Diagnostic odds ratio Kappa F1 Matthew's Correlation Coefficient (MCC) <u>PARTIAL MEASURES</u> (4) Sensitivity (a.k.a. recall) Specificity PPV (a.k.a. precision) NPV	Classification plot
Clinical utility	3	Net Benefit Standardized NB Expected cost	Decision curve

Core required property: properness

Definition: The expected value of measure is optimized for correct model/risks (fool proof)

Strictly proper: optimized for correct model only

Semi proper: optimized for the correct model and some wrong models

Improper: optimized for a wrong model

13 measures are improper

Two overall measures: discrimination slope, MAPE

All 11 classification measures

Case study: 8 got worse, 5 improved after recalibration



12 measures are semi-proper

All discrimination measures

All calibration measures

All clinical utility measures

Case study: these improved or did not change after recalibration

7 measures are strictly proper

Most overall measures (loglikelihood, Brier, R2 type)*

* Although 'skill scores' (relative measures) only asymptotically so (Gneiting & Raftery 2007)

Case study: all measures improved after recalibration

Why are classification measures improper?

They are based on t , and that means something! See Pauker & Kassirer, NEJM 1975

They do not take the misclassification costs associated with t into account

Basically, classification measures make no sense to evaluate clinical AI models

Why are discrimination, calibration and utility measures semi-proper?

1. Discrimination: rank invariant
2. Calibration: cf hierarchy in mean/weak/moderate/strong (Van Calster et al 2016)
3. NB/SNB: all models with the same classification at t have same result

Second requirement: focus

There is a clear difference between statistical performance and decision-analytic performance

Do not mix both!

Either focus on some aspect of statistical performance,
or properly address decision-analytic performance using misclassification costs

3 measures have a poor focus

Measures that mix statistical and decision-analytic performance:

AUPRC

pAUROC

F1

Which measure has the coolest name but is the worst?



F1: what and why?

Suggested alternative for the ‘misleading’ or ‘optimistic’ AUROC in “imbalanced data”:

- Hard to identify patients with the event, whereas it is important not to miss disease
- Hard to have high sensitivity and high PPV, AUROC ignores that

“Furthermore, previous studies have used AUC as the performance metric, rather than F1-score, which may have overestimated the respective model’s performance at the classification of adnexal masses, given the lack of adjustment for class imbalance” (Barcroft 2024)

F1 vs AUROC

F1

Ignores TN
Not interpretable
Depends on prevalence
F1 changes when switching outcome labels
F1 is improper

AUROC

Comprehensive
Interpretable
Prevalence-independent ($\neq$ overestimating)

A false comparison actually:
Classification vs discrimination measure

Class imbalance is not a problem, it's life

Class imbalance is not proportional to cost imbalance (conflation)

- Class imbalance: epidemiological feature
- Misclassification costs: clinical concept

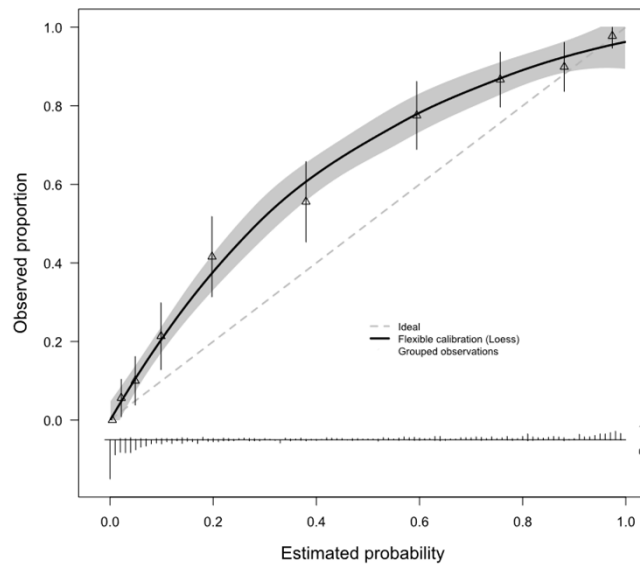
DON'T PANIC

UTILITY MEASURES

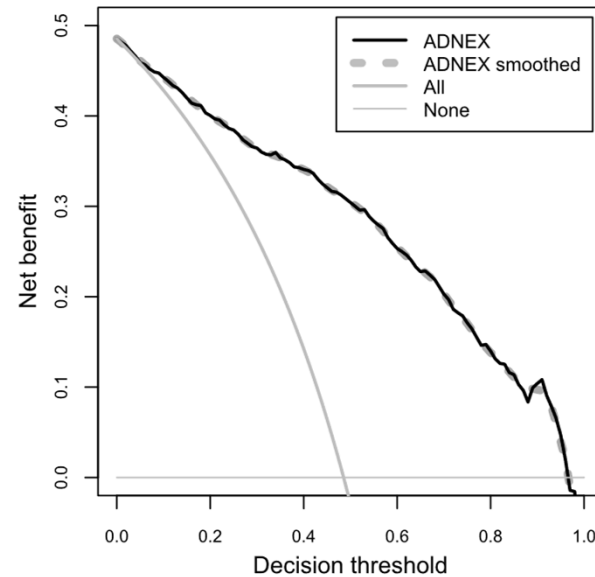
Recommended suite of measures

AUROC: 0.91 (95% CI 0.89 to 0.93)

calibration plot



NB/EC with curve



risk distributions

