Assessing performance of clinical risk prediction models in the era of machine learning

BEST OUTRUNNERS PER GOAL

	SCORE	NAME	ROUTE	RECORD
A.	5000000	YU.		
B.	4800000	BIN		
C.	4600000	KAG		
D.	4400000	MIY		
E.	4200000	MAT		

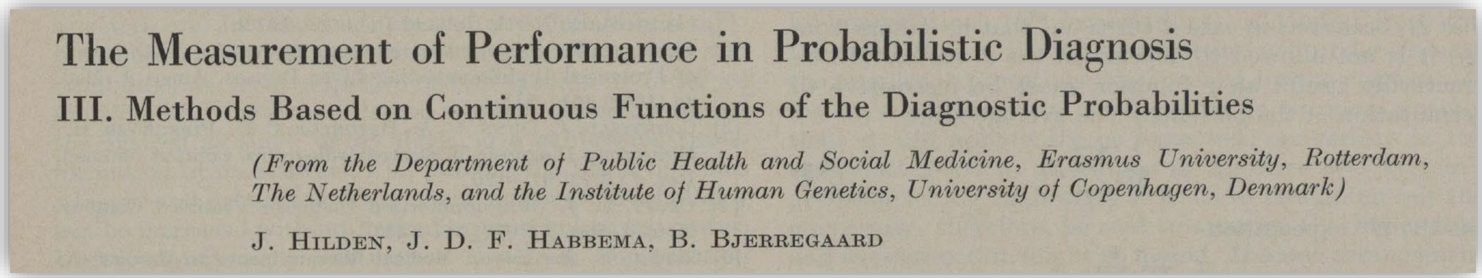
Ben Van Calster

KU Leuven (B); LUMC Leiden (NL)

KU LEUVEN

ISCB, July 2024, Thessaloniki

Binary outcomes: a lot of measures



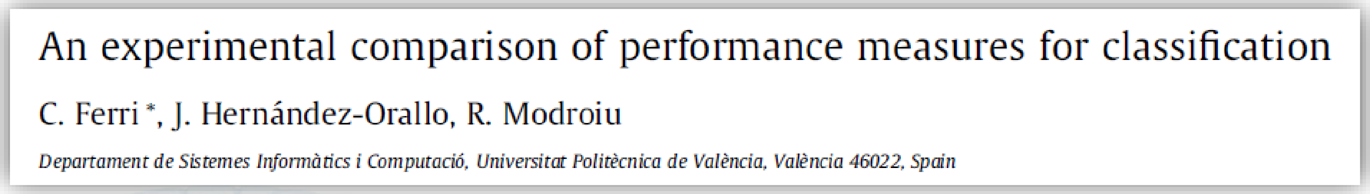
The Measurement of Performance in Probabilistic Diagnosis

III. Methods Based on Continuous Functions of the Diagnostic Probabilities

(From the Department of Public Health and Social Medicine, Erasmus University, Rotterdam, The Netherlands, and the Institute of Human Genetics, University of Copenhagen, Denmark)

J. HILDEN, J. D. F. HABBEMA, B. BJERREGAARD

Stats focus

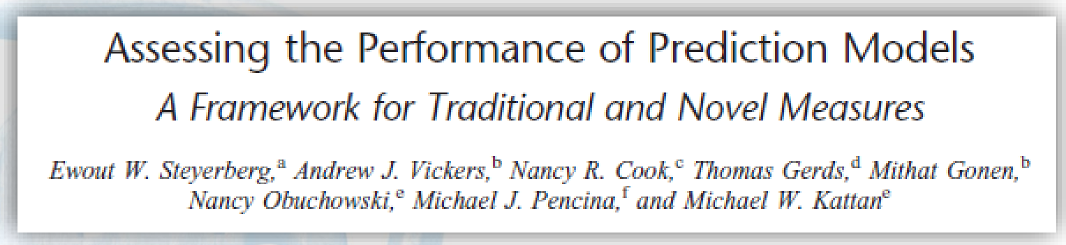


An experimental comparison of performance measures for classification

C. Ferri *, J. Hernández-Orallo, R. Modroiu

Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, València 46022, Spain

Machine learning focus

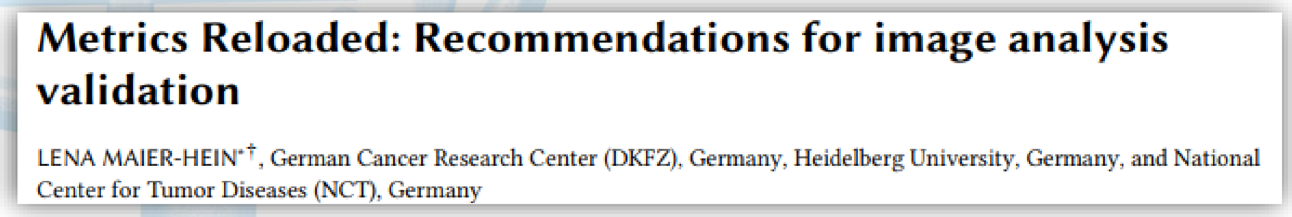


Assessing the Performance of Prediction Models

A Framework for Traditional and Novel Measures

Ewout W. Steyerberg,[a] Andrew J. Vickers,[b] Nancy R. Cook,[c] Thomas Gerds,[d] Mithat Gonen,[b] Nancy Obuchowski,[e] Michael J. Pencina,[f] and Michael W. Kattan[e]

Stats focus



Metrics Reloaded: Recommendations for image analysis validation

LENA MAIER-HEIN*†, German Cancer Research Center (DKFZ), Germany, Heidelberg University, Germany, and National Center for Tumor Diseases (NCT), Germany

Machine learning focus

Hilden et al. Meth Inform Med 1978.
Ferri et al, Pattern Recogn Lett 2009.
Steyerberg et al, Epidemiology 2010.
Maier-Hein et al, Nature Methods 2024.

STRATOS INITIATIVE

Topic Group 6: Evaluating diagnostic tests and prediction models

Performance measures for predictive AI in clinical medicine: a comprehensive overview

B VAN CALSTER, GS COLLINS, L WYNANTS, AJ VICKERS, G VAROQUAUX, KF KERR, K SINGH, M VAN SMEDEN, KGM Moons, T HERNANDEZ-BOUSSARD, D TIMMERMAN, DJ McLERNON, EW STEYERBERG

KU LEUVEN

Case study: ovarian tumor diagnosis

Aim:

Externally validate the ADNEX model to estimate risk of malignancy of a detected ovarian tumor

Support decision whether specialized surgery is needed (threshold 0.1)

External validation dataset:

n=894, 434 malignancies (49%)

Updating using logistic recalibration:

Linear transformation so rank-preserving

Landolfo et al, Br J Cancer 2024;130:934-940.

Performance domains

Domain	Focus	Target question
Probability-based evaluation		
Discrimination	Relative	Does the model estimate a higher probability in individuals with an event than individuals without an event?
Calibration	Absolute	Are probably estimates from the model reliable?
Overall	General	How close are estimated probabilities from the model (between 0 and 1) to actual outcomes (0 or 1)?
Threshold-dependent evaluation		
Classification	Binary	Are individuals classified correctly corresponding to their observed outcome?
Clinical utility	Clinical	Do classifications lead to useful decisions?

Statistical

Decision-analytic

Key desirable characteristics

Properness

Expected value of measure is optimized for correct model (fool proof)

Clear performance focus

Clear separation of statistical vs decision-analytic performance

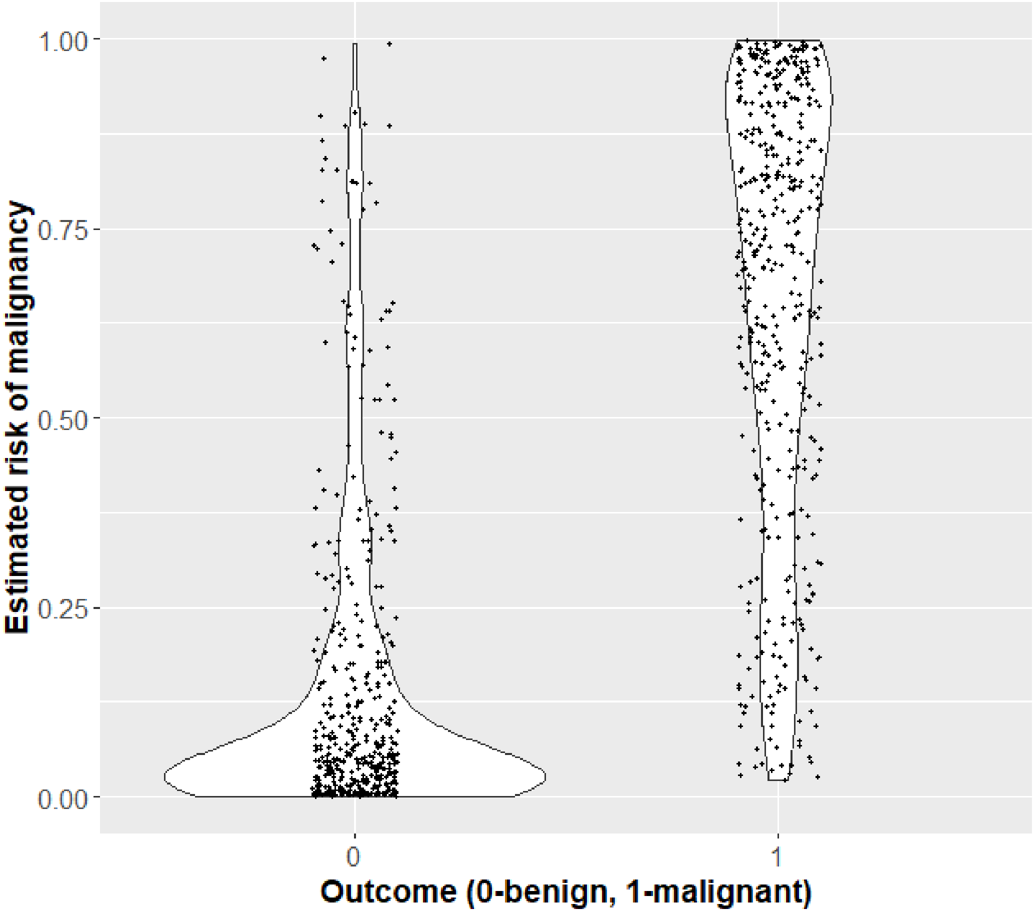
Domain	#	Measures	Plots
Discrimination	3	AUROC AUPRC (area under precision-recall curve) Partial AUROC	ROC curve Precision-recall curve
Calibration	6	O:E ratio calibration intercept Calibration slope Estimated calibration index (ECI) Integrated calibration index (ICI) Expected calibration error (ECE)	Calibration plot
Overall	9	Loglikelihood logloss (cross-entropy) Brier Scaled Brier (Brier Skill, IPA) McFadden R2 Cox-Snell R2 Nagelkerke R2 Discrimination slope (coeff. of discrimination) MAPE	Risk distributions

Domain	#	Measures	Plots
Classification	11	SUMMARY MEASURES (7) Accuracy Youden index Balanced accuracy DOR Kappa F1 Matthew’s Correlation Coefficient (MCC) PARTIAL MEASURES (4) Sensitivity (recall) Specificity PPV (precision) NPV	Classification plot
Utility	3	Net Benefit Standardized NB Expected cost	Decision curve Cost curve

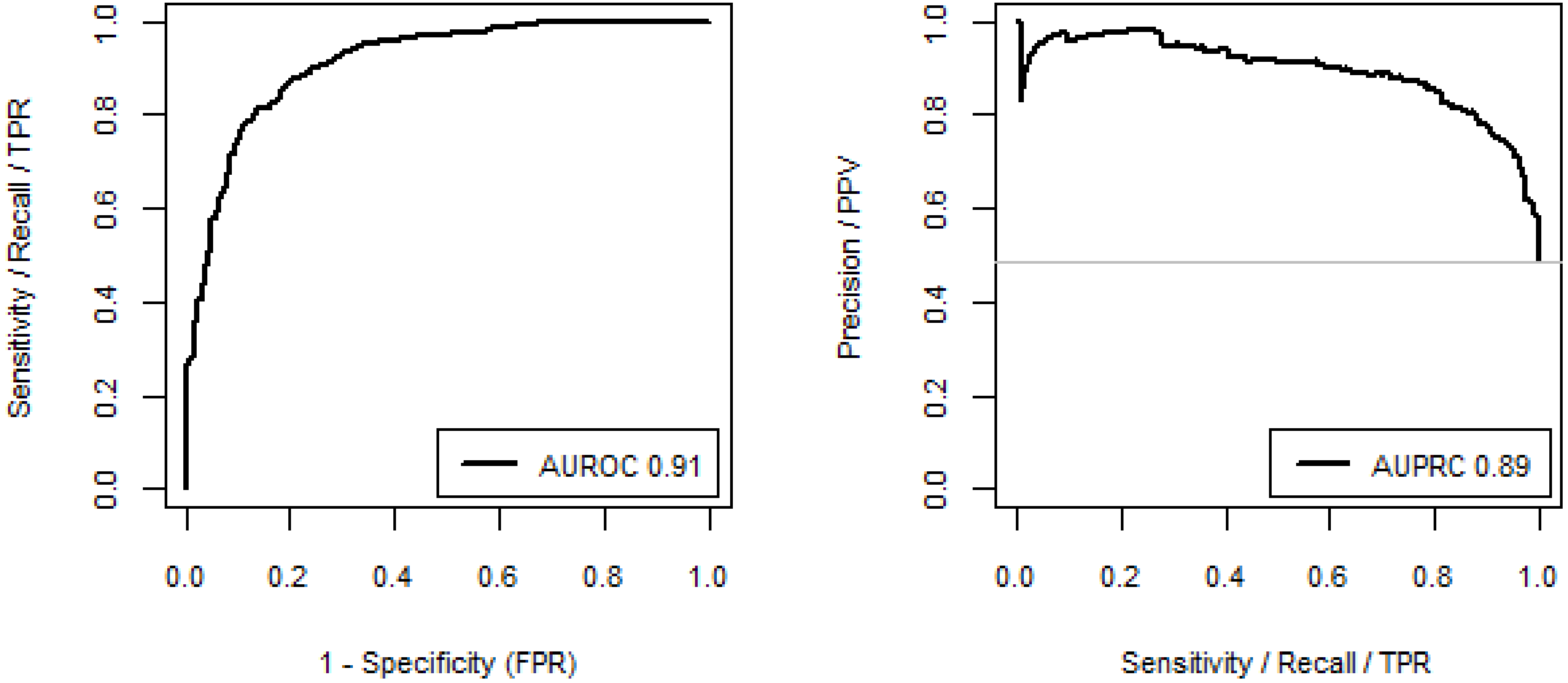
Domain	Measure	Properness	Stat vs DA focus
Discrimination	AUROC / concordance (c) statistic	Semi	OK
	AUPRC	Semi	Mixed
	Partial AUROC	Semi	Mixed
Calibration	O:E ratio	Semi	OK
	Calibration intercept	Semi	OK
	Calibration slope	Semi	OK
	Estimated calibration index (ECI)	**Strict**	OK
	Integrated calibration index (ICI)	**Strict**	OK
	Expected calibration error (ECE)	**Strict**	OK
Overall performance	Loglikelihood	**Strict**	OK
	Logloss/cross-entropy	**Strict**	OK
	Brier score	**Strict**	OK
	Scaled Brier / Brier Skill Score	**Strict**	OK
	McFadden R-squared	**Strict**	OK
	Cox-Snell R-squared	**Strict**	OK
	Nagelkerke R-squared	**Strict**	OK
	Discrimination slope	Improper	OK
	MAPE	Improper	OK

Domain	Measure	Properness	Stat vs DA focus
Classification	Classification accuracy at t	Improper	OK
	Balanced accuracy at t	Improper	OK
	Youden index at t	Improper	OK
	Diagnostic odds ratio at t	Improper	OK
	Kappa at t	Improper	OK
	F1 at t	Improper	Mixed
	Matthew’s Correlation Coefficient (MCC) at t	Improper	OK
	Sensitivity at t	Improper	OK
	Specificity at t	Improper	OK
	Positive predictive value (PPV) at t	Improper	OK
	Negative predictive value (NPV) at t	Improper	OK
Clinical utility	Net benefit	Semi	OK
	Standardized net benefit	Semi	OK
	Expected cost	Semi	OK

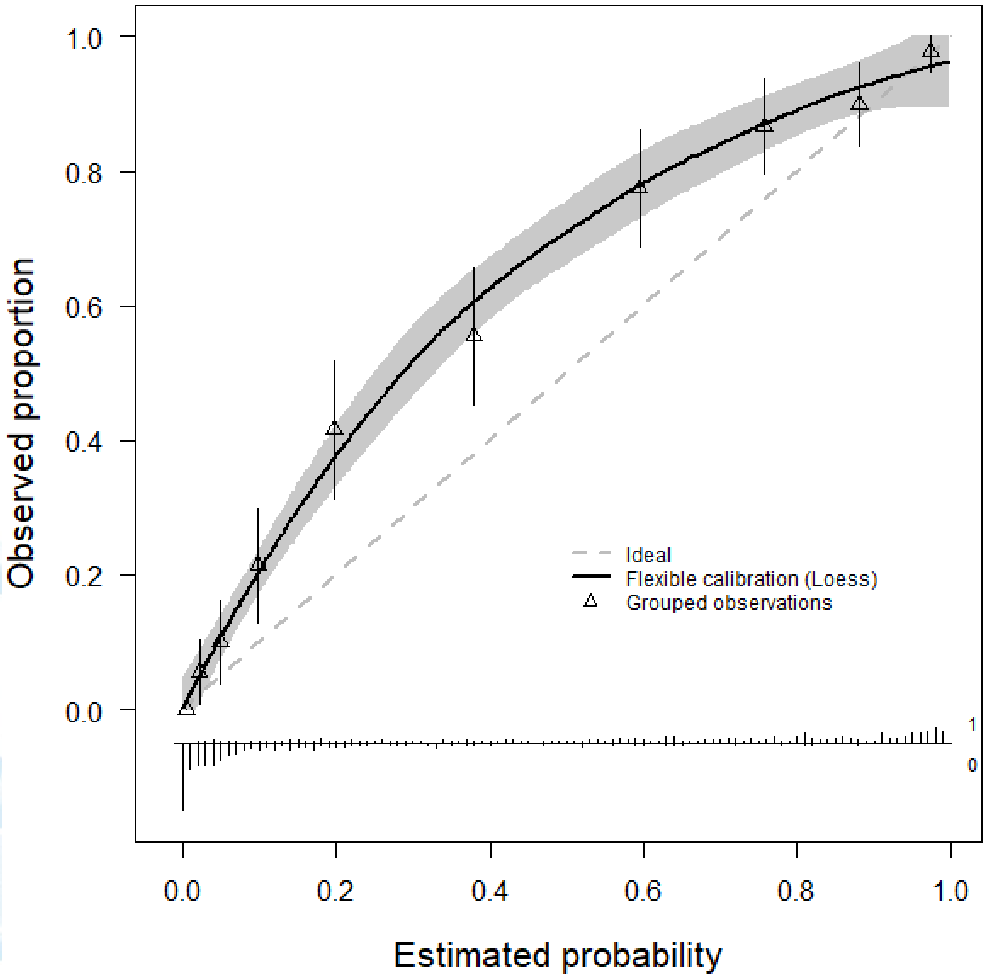
Case study: risk distributions



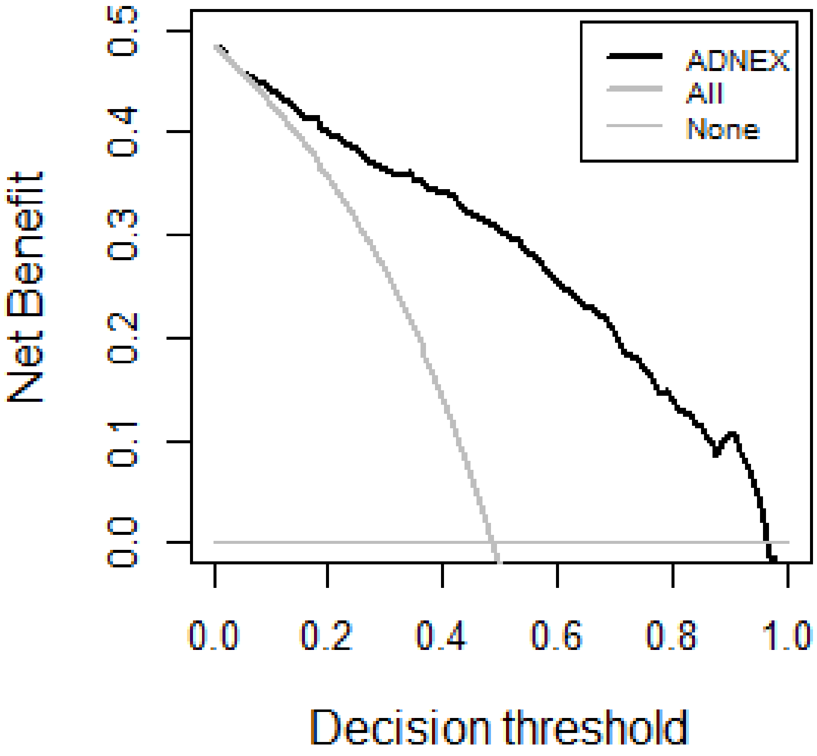
Case study: ROC and PR curves

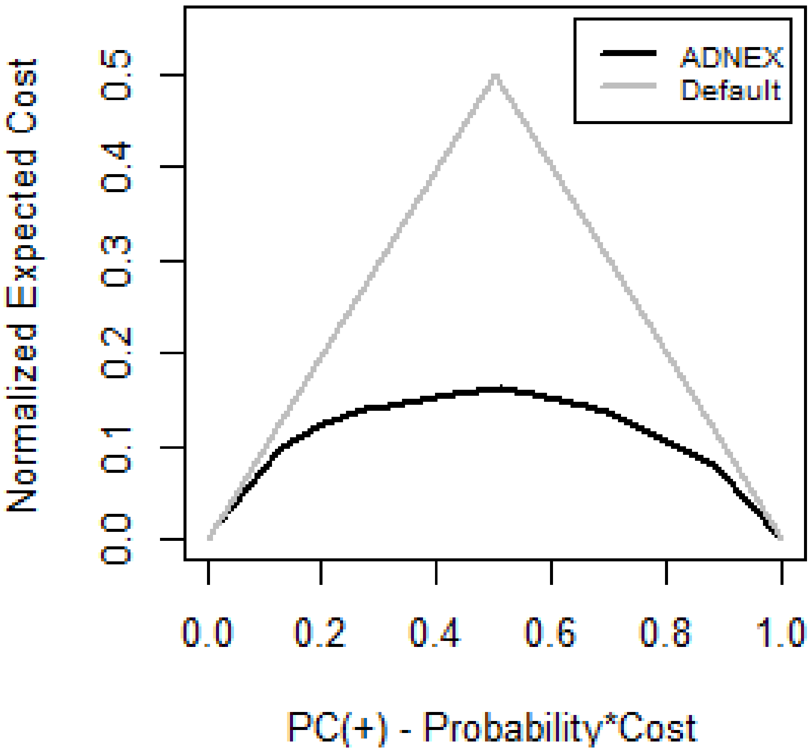


Case study: calibration plot

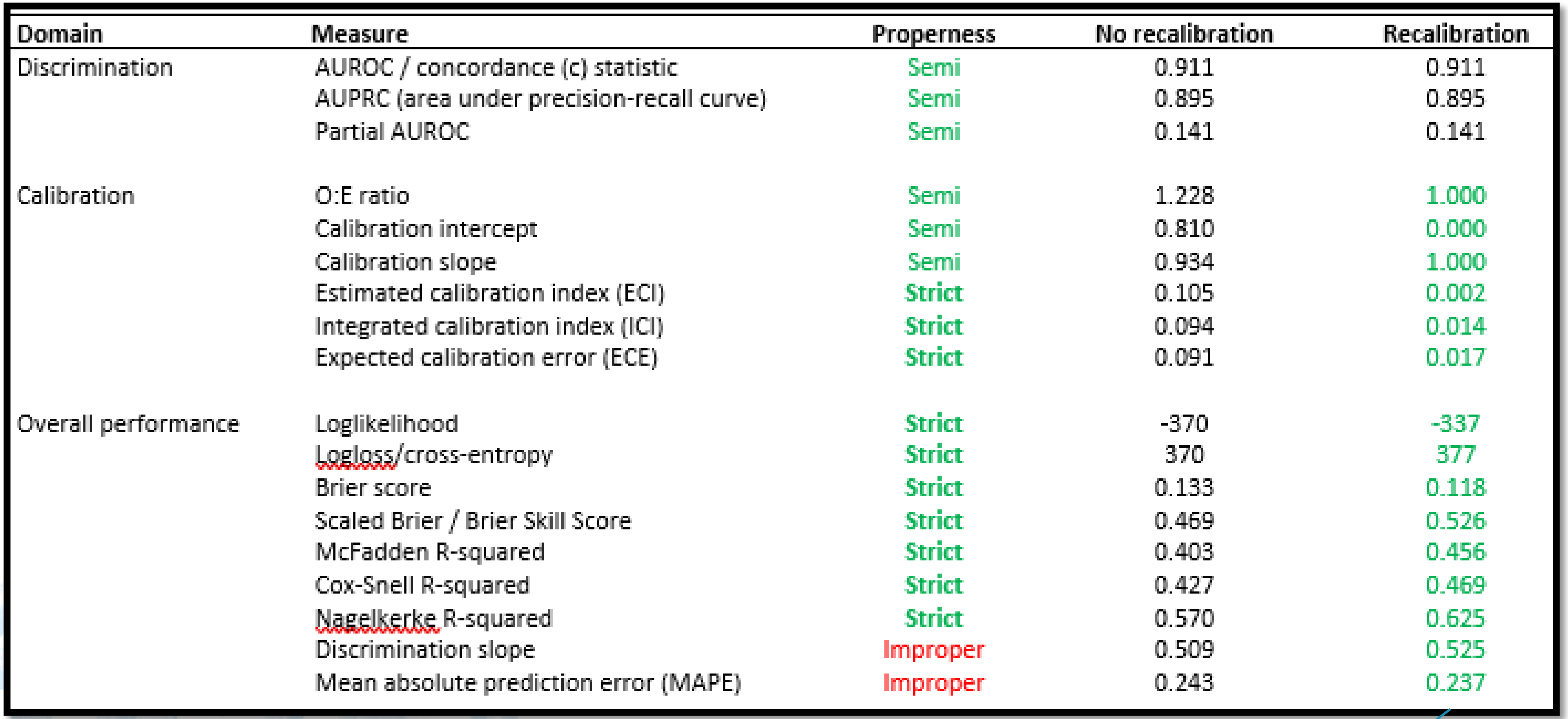


Case study: decision and cost curves





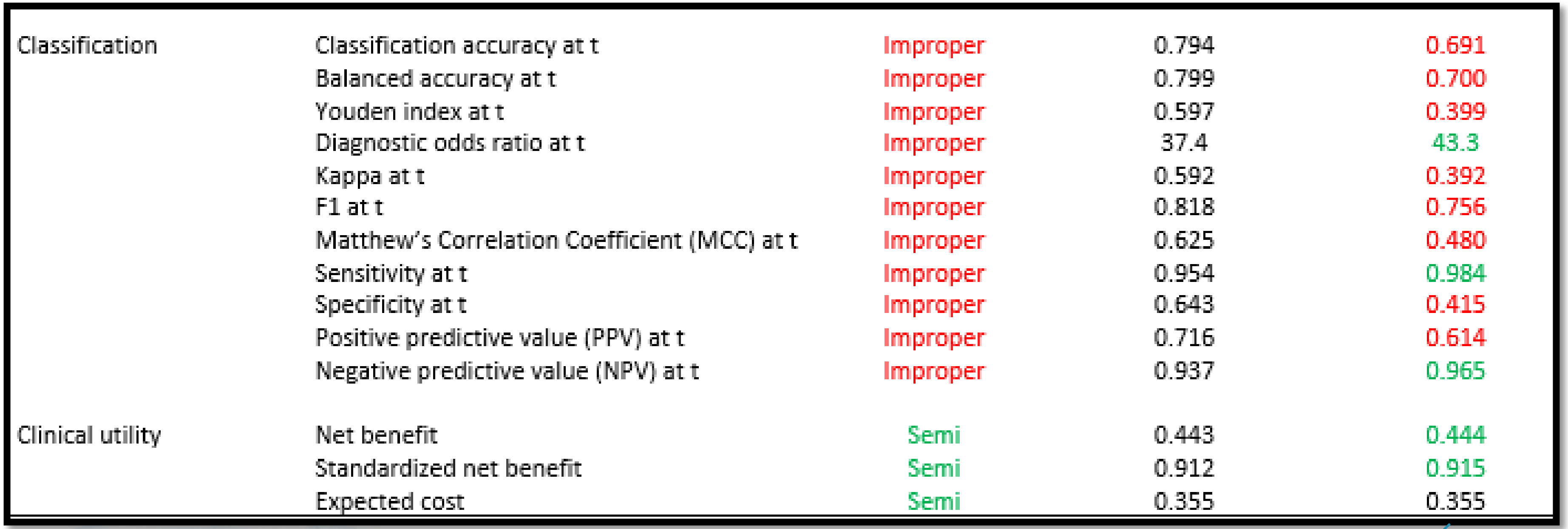
Case study: before/after updating



Domain	Measure	Properness	No recalibration	Recalibration
Discrimination	AUROC / concordance (c) statistic	Semi	0.911	0.911
	AUPRC (area under precision-recall curve)	Semi	0.895	0.895
	Partial AUROC	Semi	0.141	0.141
Calibration	O:E ratio	Semi	1.228	1.000
	Calibration intercept	Semi	0.810	0.000
	Calibration slope	Semi	0.934	1.000
	Estimated calibration index (ECI)	Strict	0.105	0.002
	Integrated calibration index (ICI)	Strict	0.094	0.014
	Expected calibration error (ECE)	Strict	0.091	0.017
Overall performance	Loglikelihood	Strict	-370	-337
	Logloss/cross-entropy	Strict	370	377
	Brier score	Strict	0.133	0.118
	Scaled Brier / Brier Skill Score	Strict	0.469	0.526
	McFadden R-squared	Strict	0.403	0.456
	Cox-Snell R-squared	Strict	0.427	0.469
	Nagelkerke R-squared	Strict	0.570	0.625
	Discrimination slope	Improper	0.509	0.525
	Mean absolute prediction error (MAPE)	Improper	0.243	0.237

Green: better
Red: worse

Case study: before/after updating



Classification	Classification accuracy at t	Improper	0.794	0.691
	Balanced accuracy at t	Improper	0.799	0.700
	Youden index at t	Improper	0.597	0.399
	Diagnostic odds ratio at t	Improper	37.4	43.3
	Kappa at t	Improper	0.592	0.392
	F1 at t	Improper	0.818	0.756
	Matthew's Correlation Coefficient (MCC) at t	Improper	0.625	0.480
	Sensitivity at t	Improper	0.954	0.984
	Specificity at t	Improper	0.643	0.415
	Positive predictive value (PPV) at t	Improper	0.716	0.614
	Negative predictive value (NPV) at t	Improper	0.937	0.965
Clinical utility	Net benefit	Semi	0.443	0.444
	Standardized net benefit	Semi	0.912	0.915
	Expected cost	Semi	0.355	0.355

Green: better
Red: worse

“Confusion” matrix: it’s in the name

How many measures are there to summarize a 2x2 table?!

The 7 measures we evaluated are improper at threshold *t*

Reason: *t* implies specific misclassification costs, but these are ignored

Summary measures: no value to formally assess or compare performance

Partial measures (sens, spec, PPV, NPV): good for description

F1

Harmonic mean of PPV (precision) and sensitivity (recall)

Fierce defenders

"Furthermore, previous studies[18,21,31] 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, npj Precis Oncol 2024)

- F1 ignores true negatives
- F1 absolute value changes by switching the outcome labels
- F1 value cannot be interpreted
- F1 at threshold *t* is improper, like all classification measures

Precision-Recall curve

Alternative for ROC

Plots PPV (aka precision) by sensitivity (aka recall)

“The PR curve overcame the optimism of the ROC curve in rare diseases”

(Ozenne, JCE 2015)

AUPRC: ignores TN, depends on prevalence

AUROC: comprehensive and interpretable (~ Mann-Whitney)

it does not depend on prevalence (≠ overestimating)

AUROC does not tell the full story, but AUPRC does not solve this

Matthew’s Correlation Coefficient (MCC)

Pearson correlation of classifications and outcomes.

(cf phi correlation)

Interpretation?

It will not help.

Utility: net benefit or expected cost

NB uses the link between threshold and misclassification costs

EC does not, it rather does logistic recalibration behind the scenes

NB: “misclassification costs imply t=0.1, so how useful is model at t=0.1?”

EC: “OK, misclassification costs imply t=0.1, but cost minimized at t=0.06”

Class imbalance is not a problem

Class imbalance: the two outcome classes are not equally common

Claims that some measures (AUROC, accuracy) are invalid/misleading because imbalance not considered

AUPRC/F1 often recommended to ‘overcome’ this ‘problem’

But:

- Class imbalance is not proportional to cost imbalance (conflation)
- Some measures are just improper (eg accuracy)
- We have utility measures to address this appropriately

Imbalance is a fact of life rather than a problem, so just deal with it

What measures/plots to use?

(Being discussed ATM)

Do not use measures that do not meet the 2 characteristics

Generally, these seem the key ones:

- Show risk distributions (~ overall)
- Discrimination: AUROC
- Calibration: provide a calibration plot

For the modeler:
How can we improve the model?

(if the model is intended to support decisions)

- Classification: descriptive measures
- Clinical utility: NB or EC with plot

For the decision maker:
Is the model potentially useful?

KU LEUVEN