Inspired by research. Driven by compassion.
Inspiré par la recherche. Guidé par la compassion.

Better evaluation of binary diagnostic tests and classifiers with a concordant partial area under the ROC curve

André Carrington PhD PEng
Doug Manuel MD MSc FRCPC

UPDATED JANUARY 7, 2020
ORIGINALLY PRESENTED NOVEMBER 7, 2019

The Ottawa Hospital | L'Hôpital d'Ottawa
RESEARCH INSTITUTE | INSTITUT DE RECHERCHE

www.ohri.ca | Affiliated with • Affilié à uOttawa

Executive summary (simple)

To evaluate a classifier or binary diagnostic test:

- The area under the ROC curve (AUC) is standard but includes decision thresholds which are unrealistic, not clinically relevant.
- The partial AUC measures avg sensitivity without specificity, while the partial area index measures avg specificity without sensitivity. The standardized partial area is also flawed.
- We devise the (proper) concordant partial AUC and its equal, the (first) partial *c* statistic[i,ii] for ROC as the only partial measures interpretable as a c statistic and with a clear relation to both avg. sensitivity and avg. specificity.

i. except for survival regression, where Harrell's C-index (sometimes called a *c* statistic) differs from classification's *c* statistic
ii. existing partial *c* is a different concept and purpose[30-33]

Executive summary (technical)

To evaluate a classifier or binary diagnostic test:

- AUC is standard/recommended[26,27,28], but flawed[4,5,6,8,17,18,34]. AUC equals the *c* statistic*, which provides interpretation.
- Partial AUC[8,16] is better, e.g., focuses on a clinically relevant region, but biased to positives[23] and flawed[17,22]. sPA[17] resolves that but has a flaw[22,23] and shortcoming[23], which we resolve. Alternatives[9,17-21] lack AUC's three key interpretations, until…
- We devise the (proper) concordant partial AUC[23], and its equal, the partial *c* statistic for ROC[23] (the first**) as generalizations of AUC and c

* except for survival regression, where Harrell's C-index (sometimes called a *c* statistic) has continuous targets, fewer ties, and multiple time-dependent ROC/AUC, different from classification's *c* statistic

** existing partial *c* is a different concept and purpose[30-33]

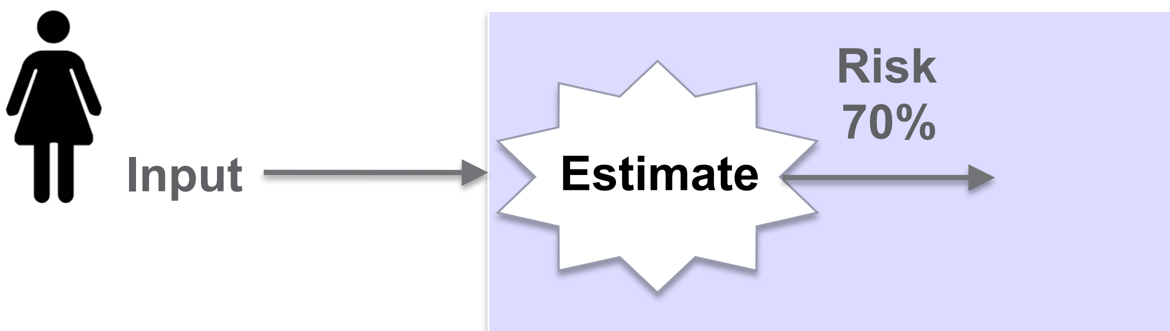
The Ottawa Hospital RESEARCH INSTITUTE | L'Hôpital d'Ottawa INSTITUT DE RECHERCHE

Affiliated with • Affilié à uOttawa

Testing or predicting binary outcomes

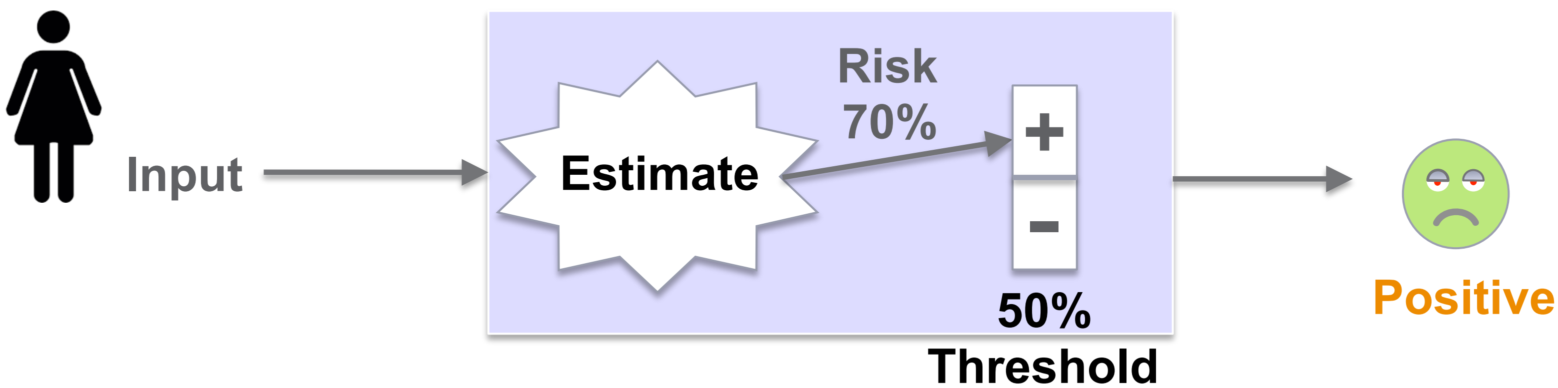
- Strep throat
- Breast cancer remission within 1 year of treatment[1]
- Lung cancer tumor malignancy[2]
- Hospital readmission within 1 year[3]

Tests/classifiers estimate risk as a continuous value*



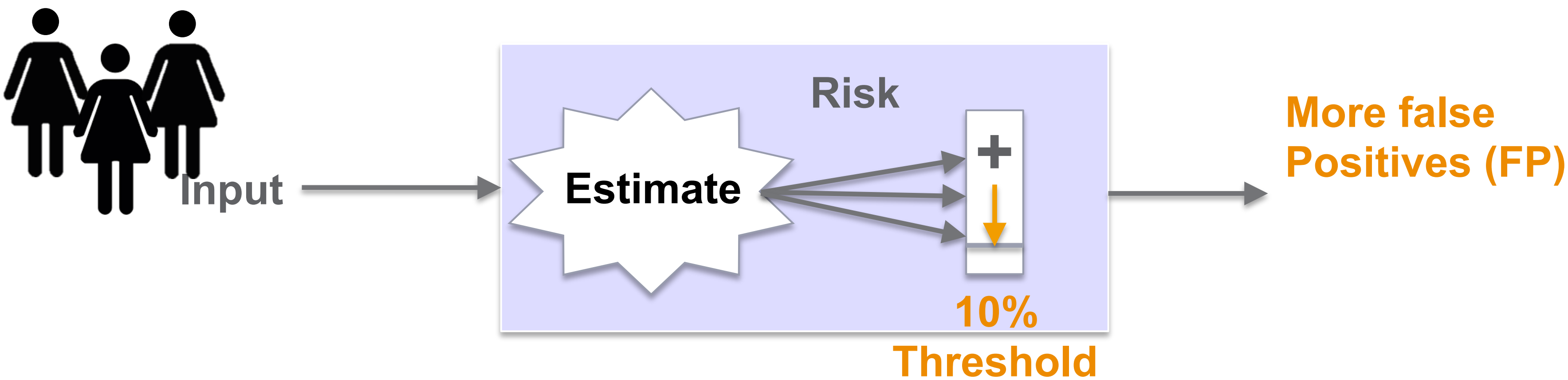
* Exceptions: k-NN, Decision trees, rule-based expert systems

Tests/classifiers estimate risk as a continuous value* **and threshold it**

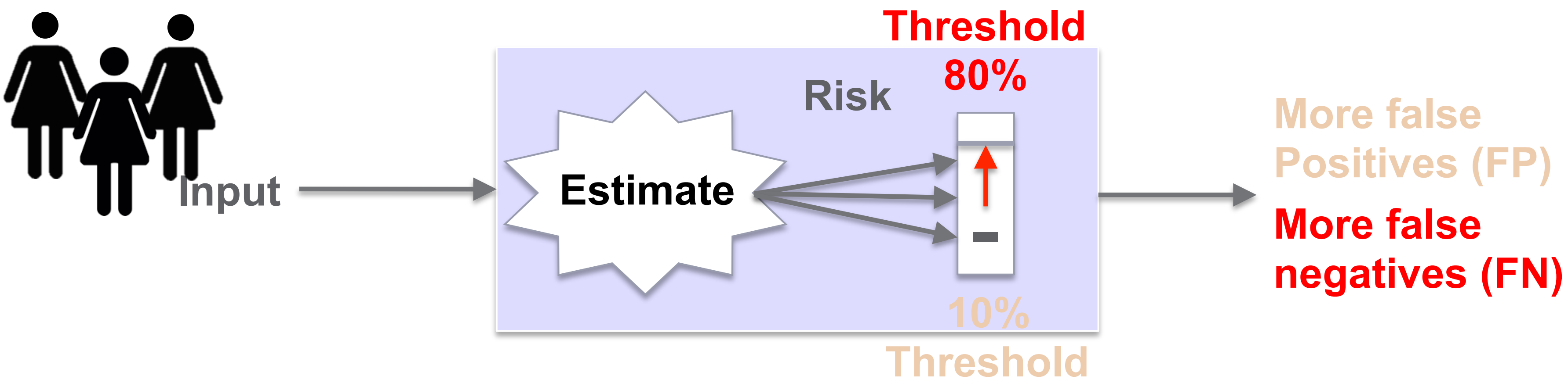


* Exceptions: k-NN, Decision trees, rule-based expert systems

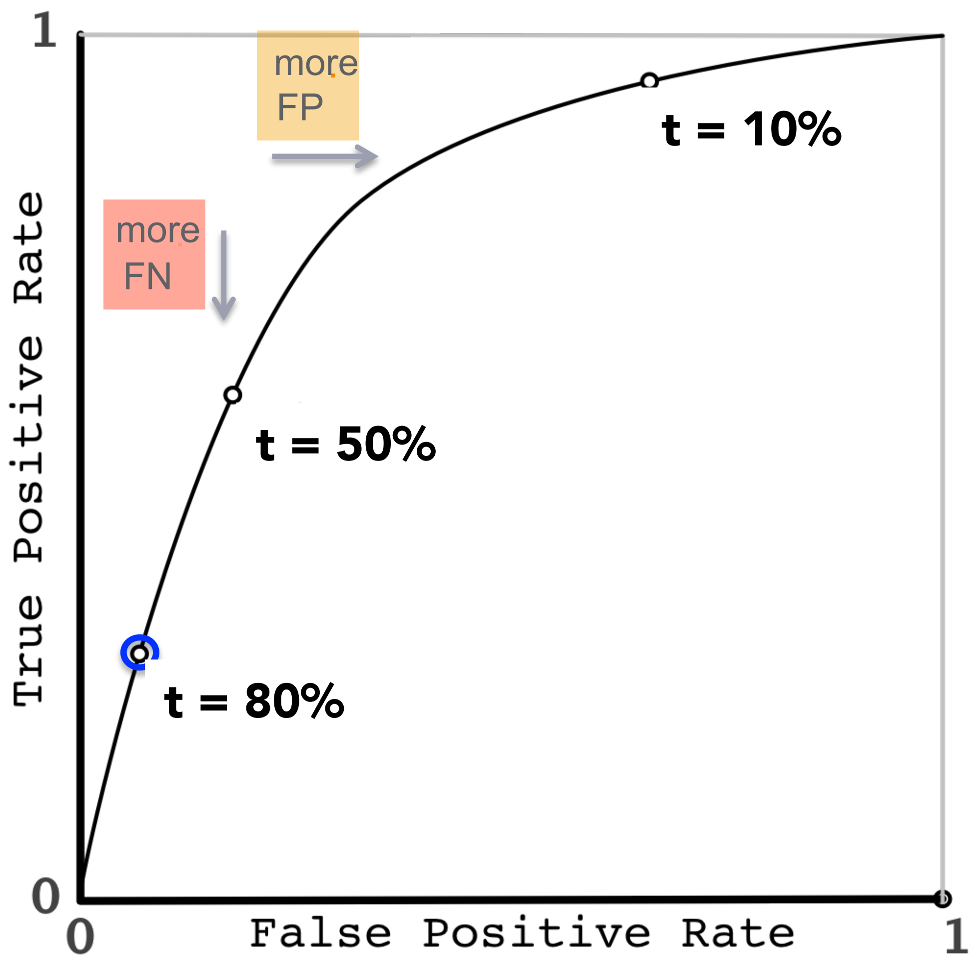
Lower thresholds cause more false positives



Higher thresholds cause more false negatives



We plot those three thresholds (t) together



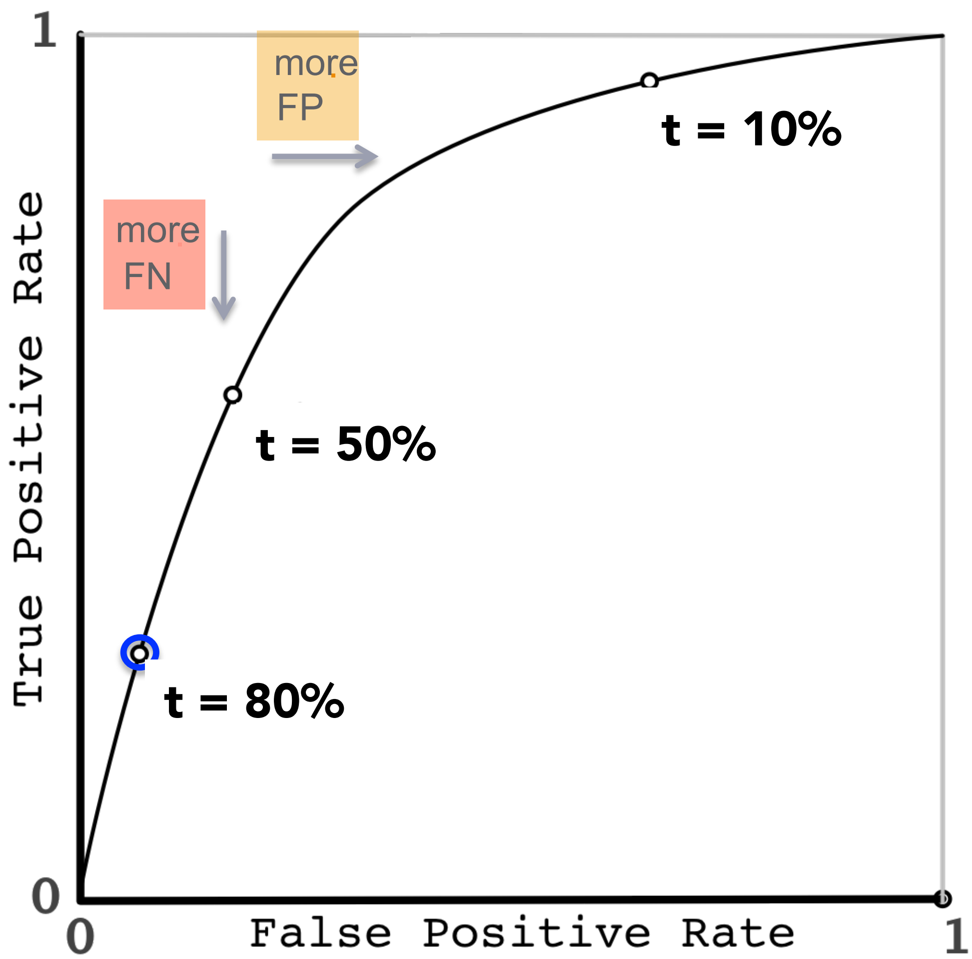
	True=P	True=N
Pred=P	95	25
=N	5	15

In a receiver operating characteristic (ROC) plot[35-37].

55	10
45	30

25	4
75	36

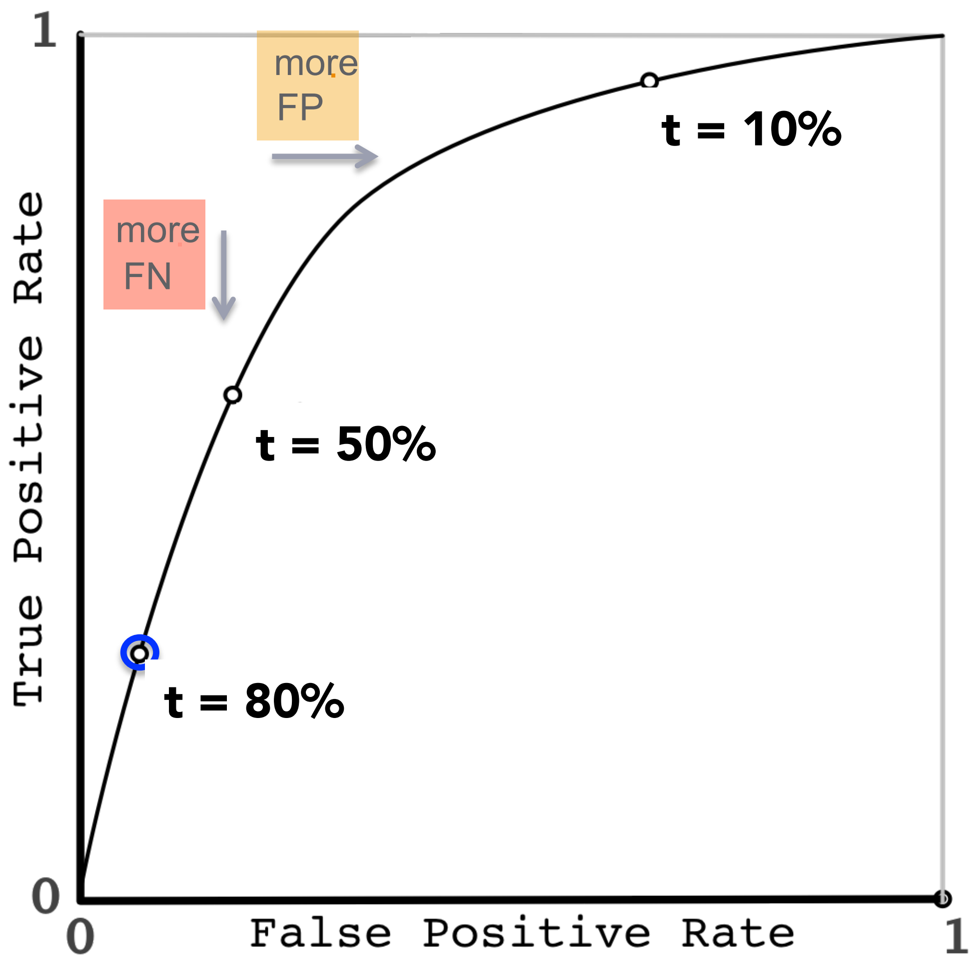
We plot those three thresholds (t) together



But a receiver operating characteristic (ROC) plot[35-37] **is unlike any other 2D plot!**

Normally, a 2D plot takes coordinates (TPR,FPR) as input, **but that is a SROC[10]!**

We plot those three thresholds (t) together



But a receiver operating characteristic (ROC) plot[35-37] **is unlike any other 2D plot!**

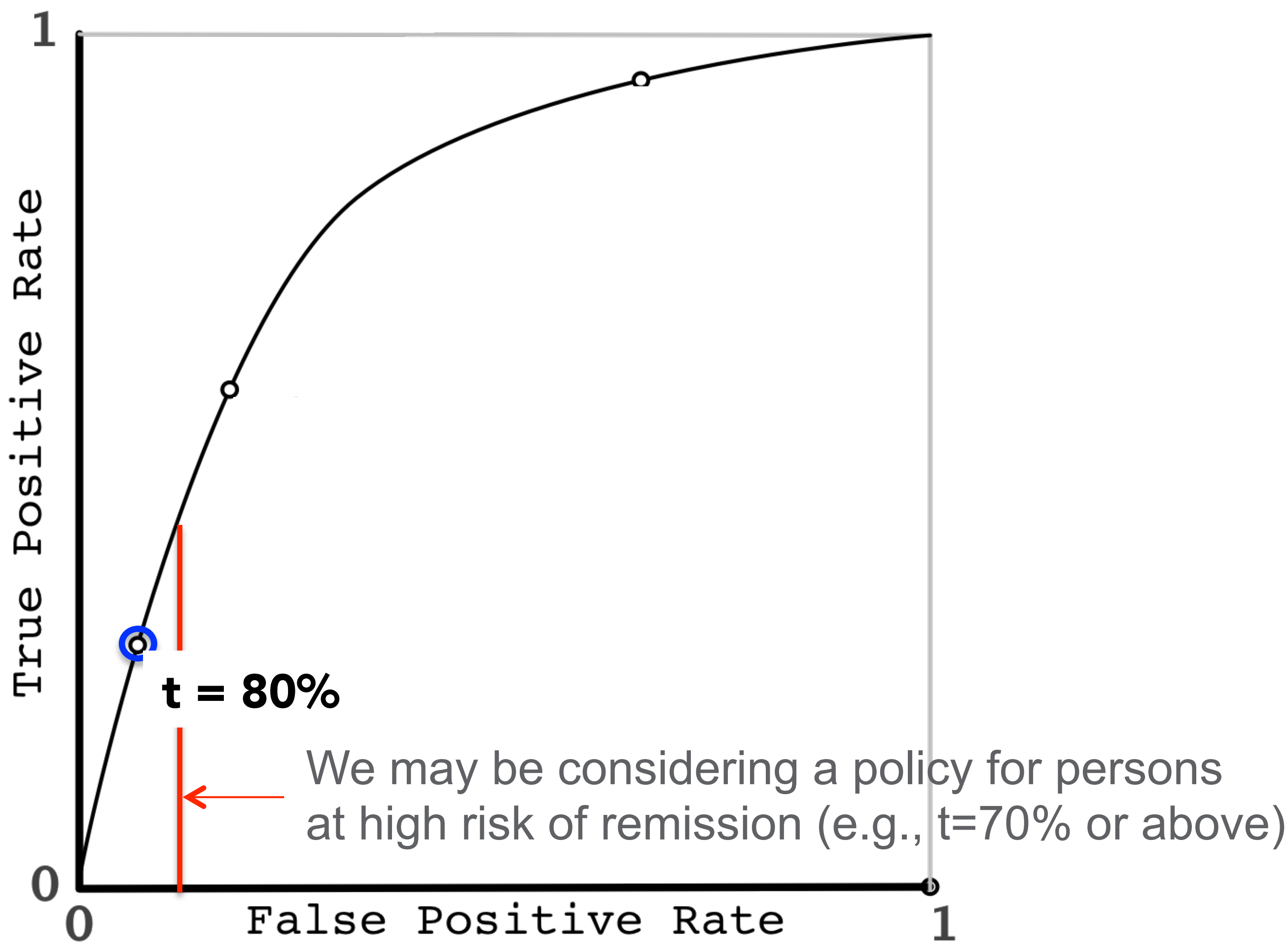
Normally, a 2D plot takes coordinates (TPR,FPR) as input, **but that is a SROC[10]!**

ROC plots take (score,label) inputs, following a procedure that sweeps a threshold across scores.

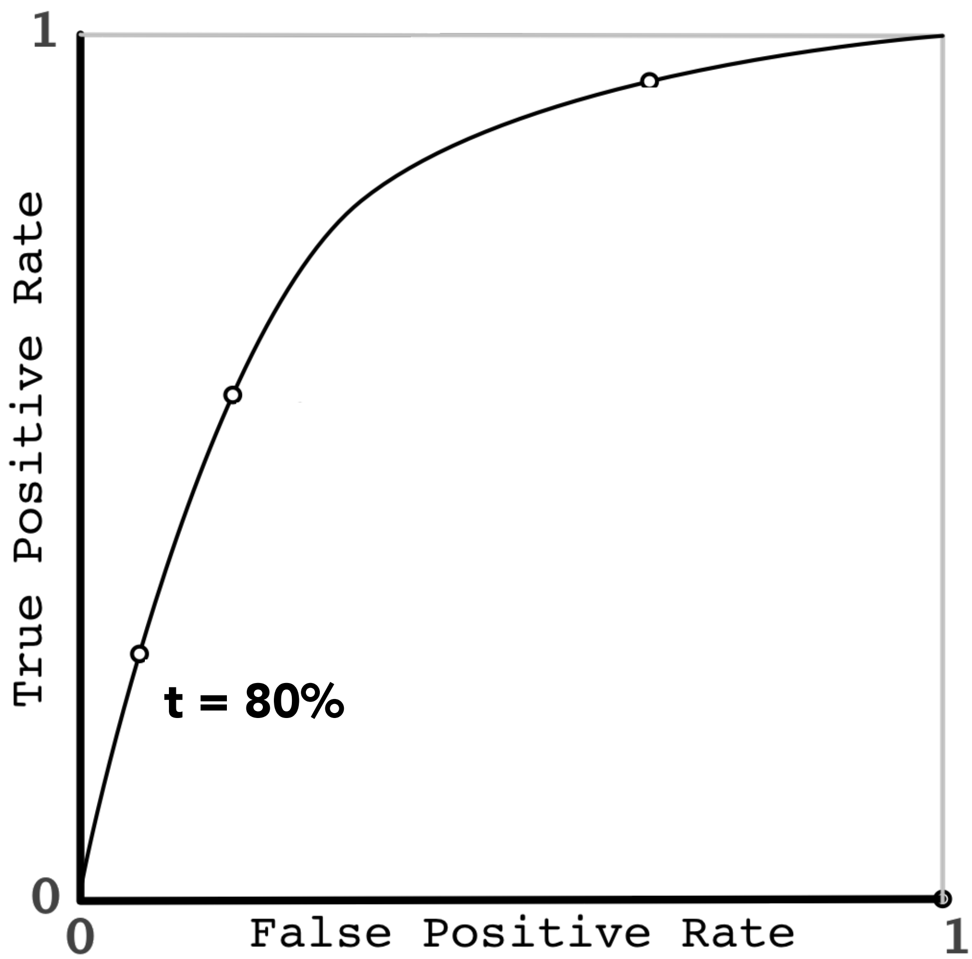
The Ottawa Hospital | L'Hôpital d'Ottawa
RESEARCH INSTITUTE | INSTITUT DE RECHERCHE

Affiliated with • Affilié à uOttawa

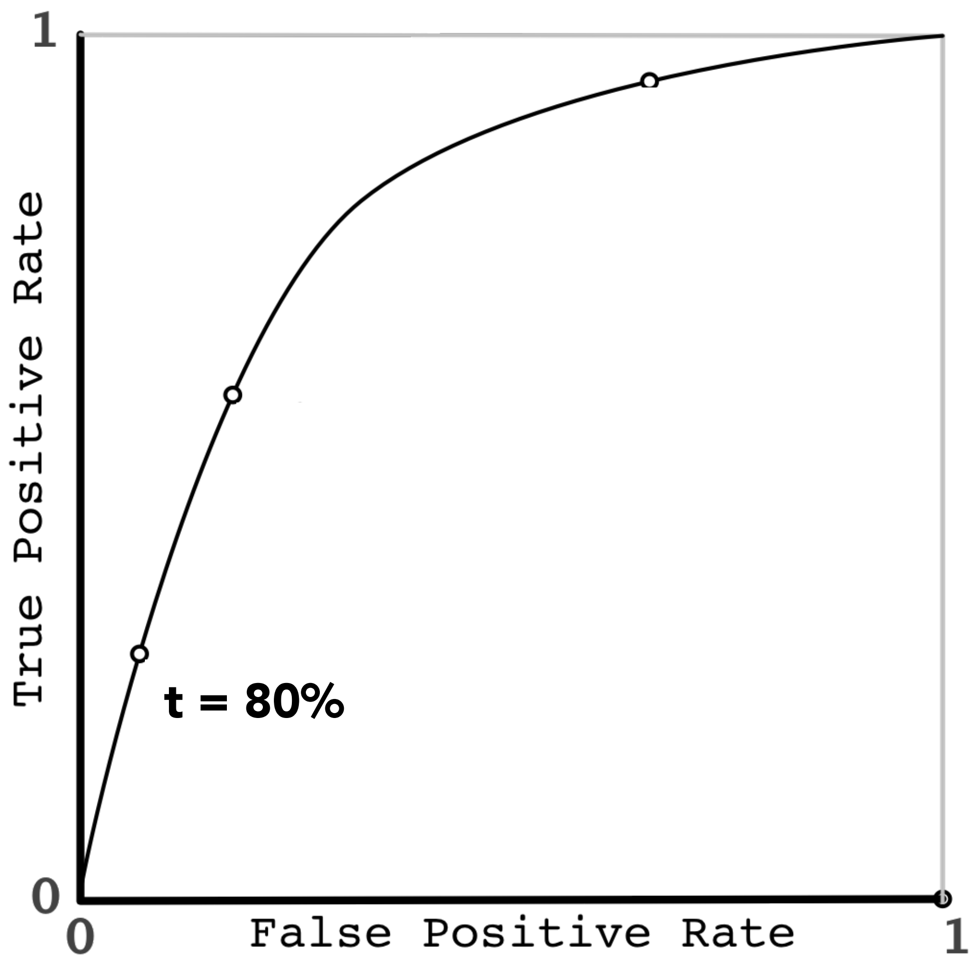
For our breast cancer remission example



What do we do with an ROC plot?



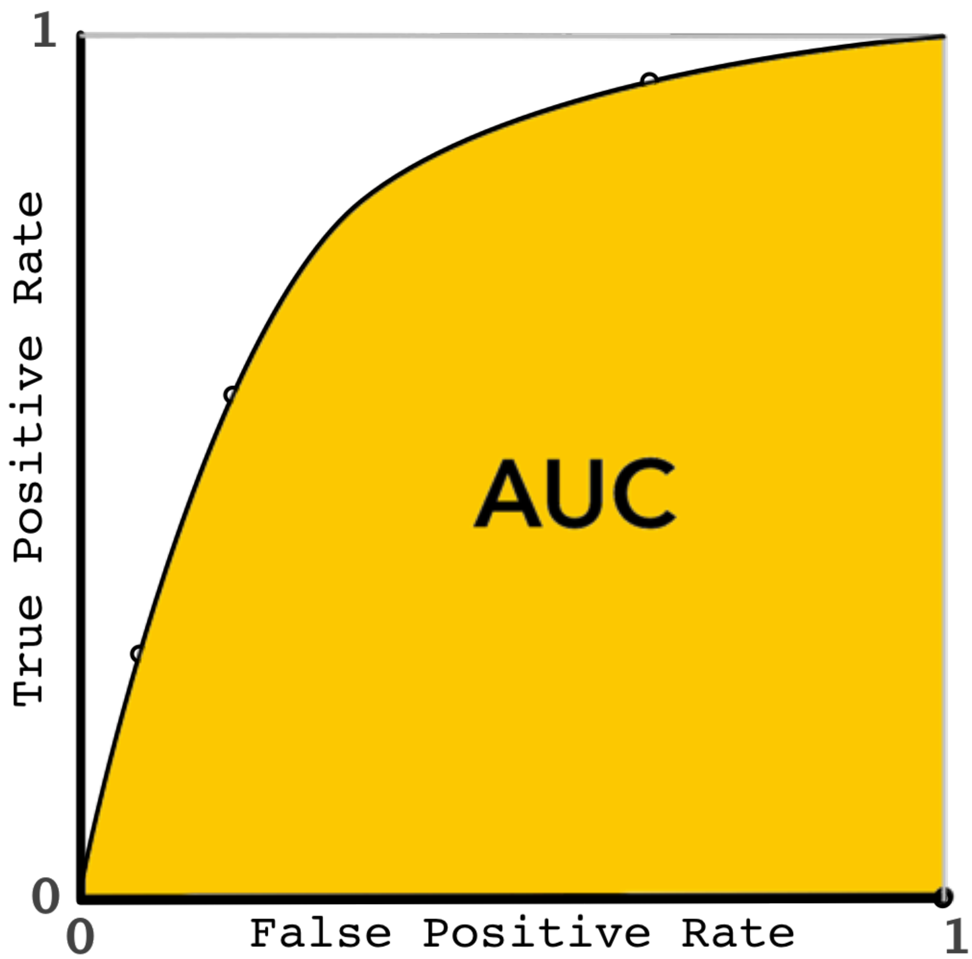
What do we do with an ROC plot?



1. Numerically report/compare performance
2. Pick a threshold to use

etc.

1. We report or compare performance

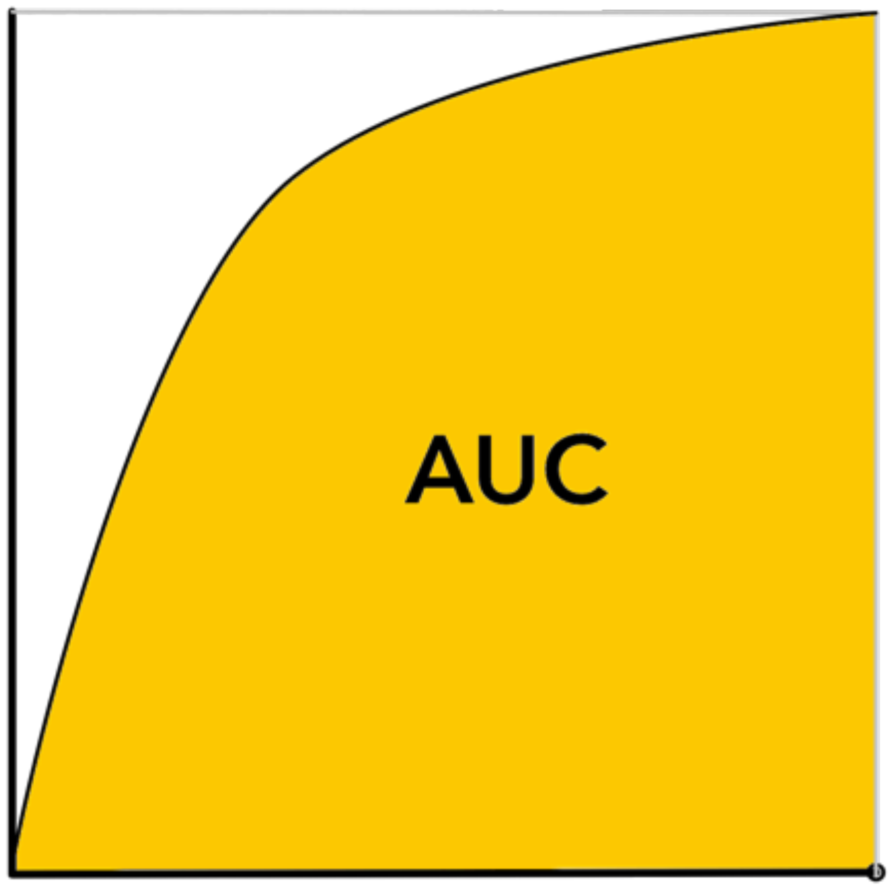


Using the area under the ROC curve (AUC)[29].

The Ottawa Hospital | L'Hôpital d'Ottawa
RESEARCH INSTITUTE | INSTITUT DE RECHERCHE

Affiliated with • Affilié à uOttawa

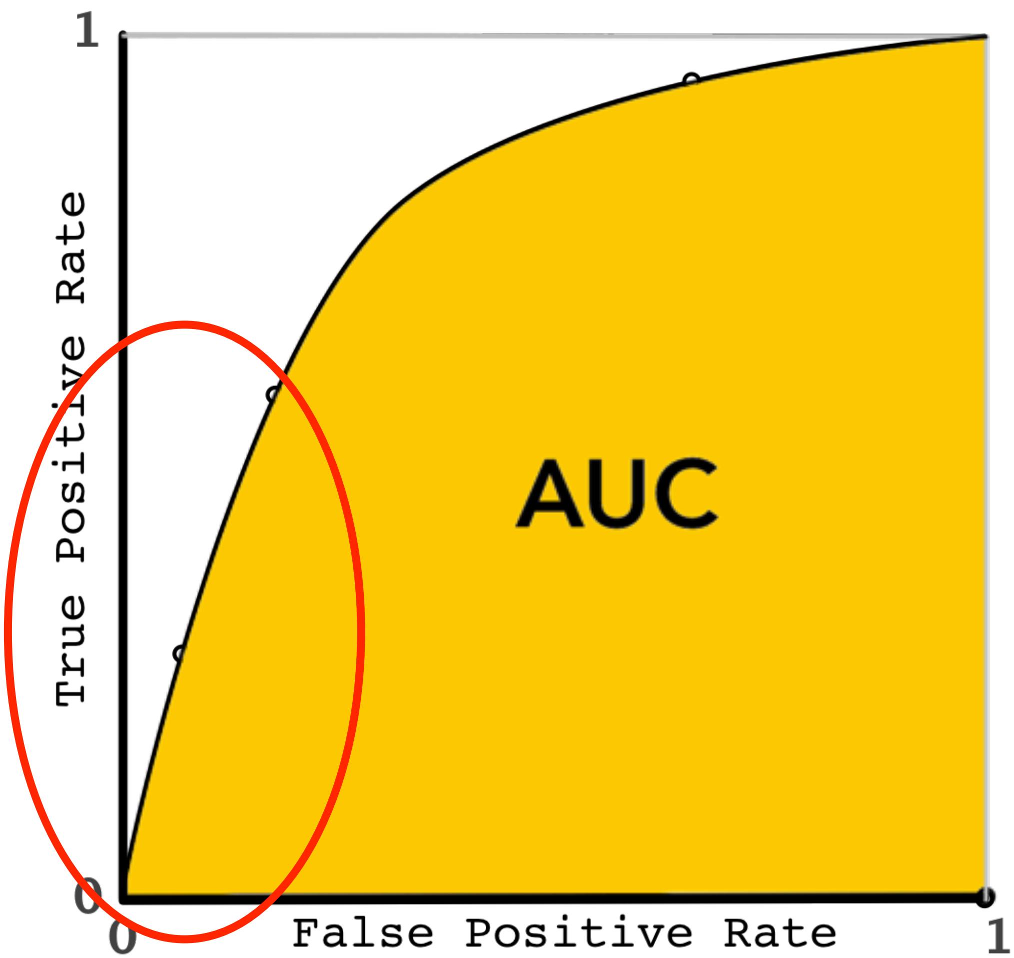
1. We report or compare performance



Using the area under the ROC curve (AUC)[29]:

- **"Intrinsic accuracy"**
- **Average sensitivity over all thresholds and risk groups.**
- **Average specificity over all thresholds and risk groups.**
- **Concordance: % agreement of rank in scores {0.7,0.4} with labels {pos,neg}, for every possible pos/neg pair.**

1. We report or compare performance

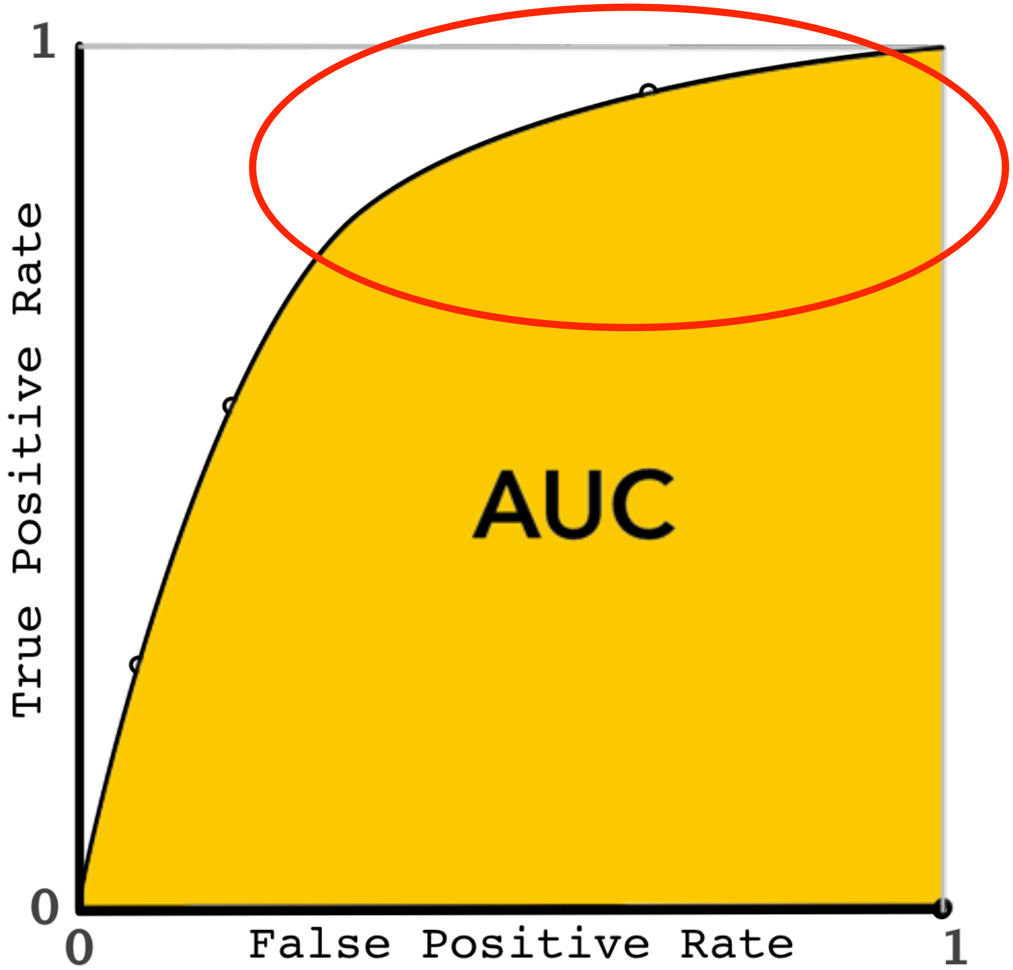


Using the area under the ROC curve (AUC)[29].

But only some regions are relevant![4,5,6]

For low prevalence, the region of interest is at left[7,8].

1. We report or compare performance



Using the area under the ROC curve (AUC)[29].

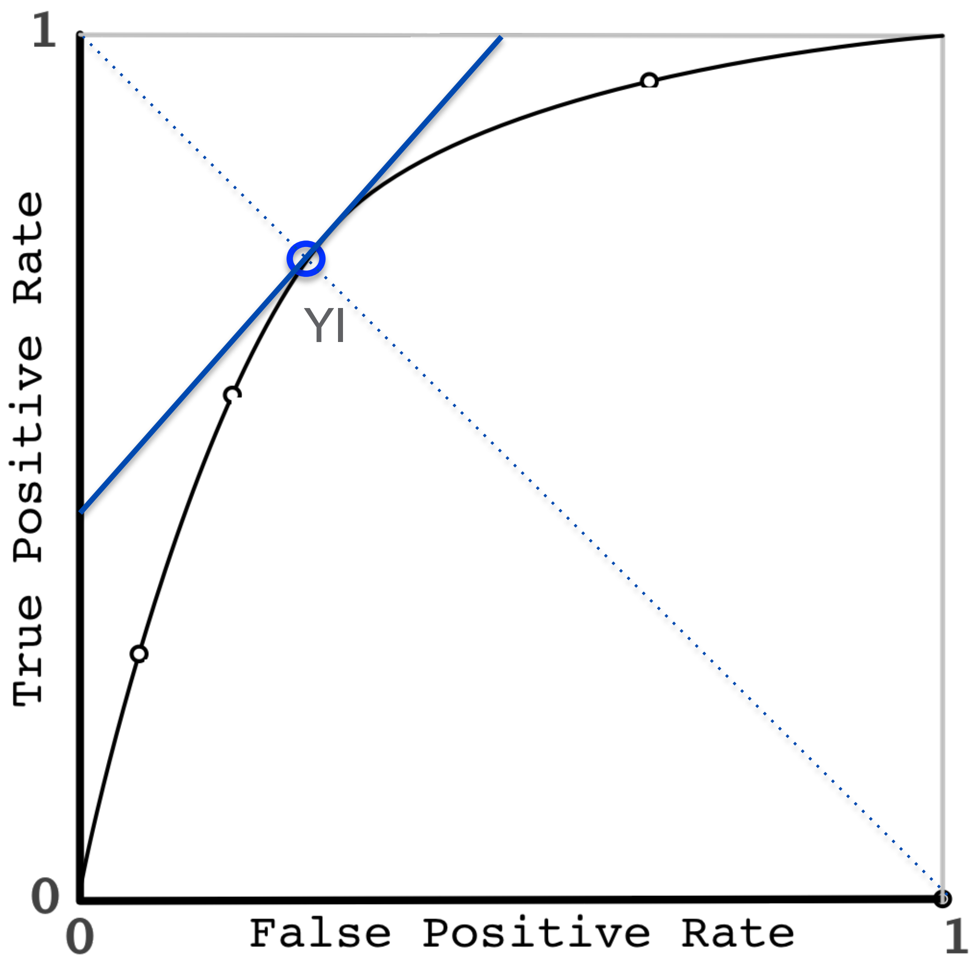
But only some regions are relevant![4,5,6]

For low prevalence, the region of interest is at left[7,8].

For high prevalence, the top is the region of interest[7,9,10].

So AUC has flaws[4,5,6,8]

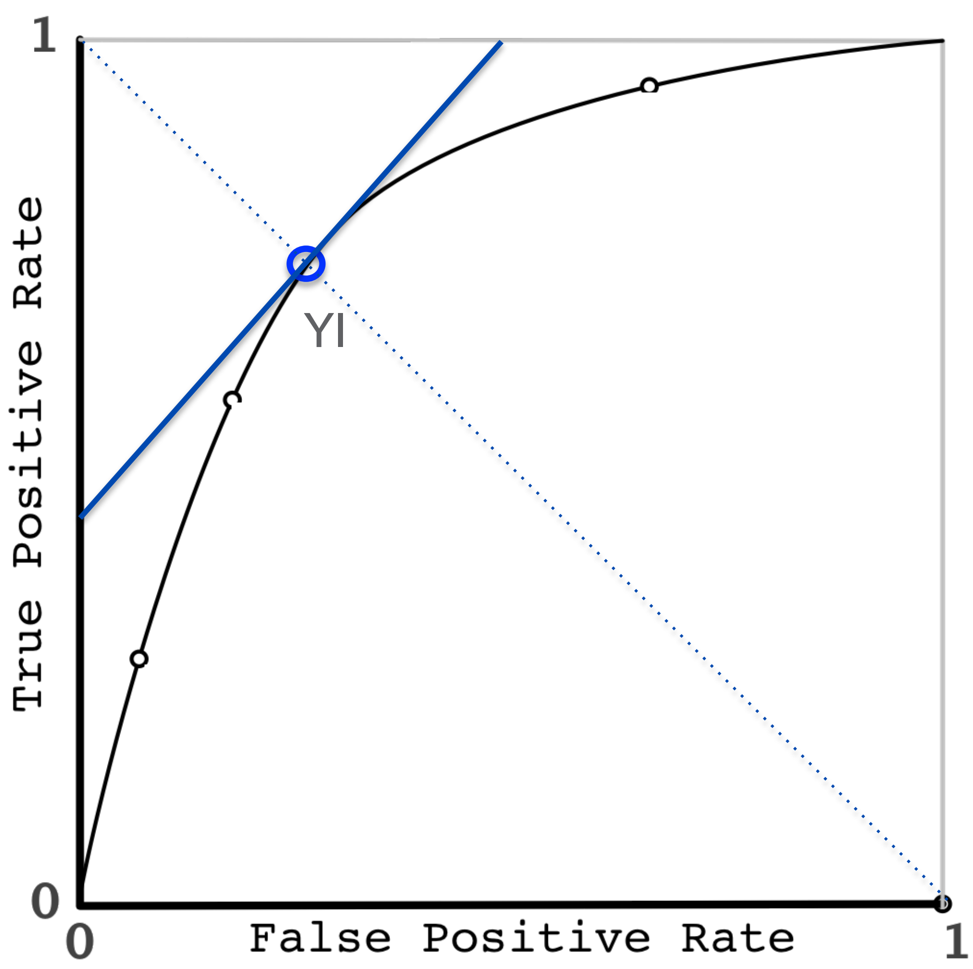
2. We can also pick a threshold to use



Youden's index[11] is typical.

The Ottawa Hospital | L'Hôpital d'Ottawa
RESEARCH INSTITUTE | INSTITUT DE RECHERCHE

2. We can also pick a threshold to use

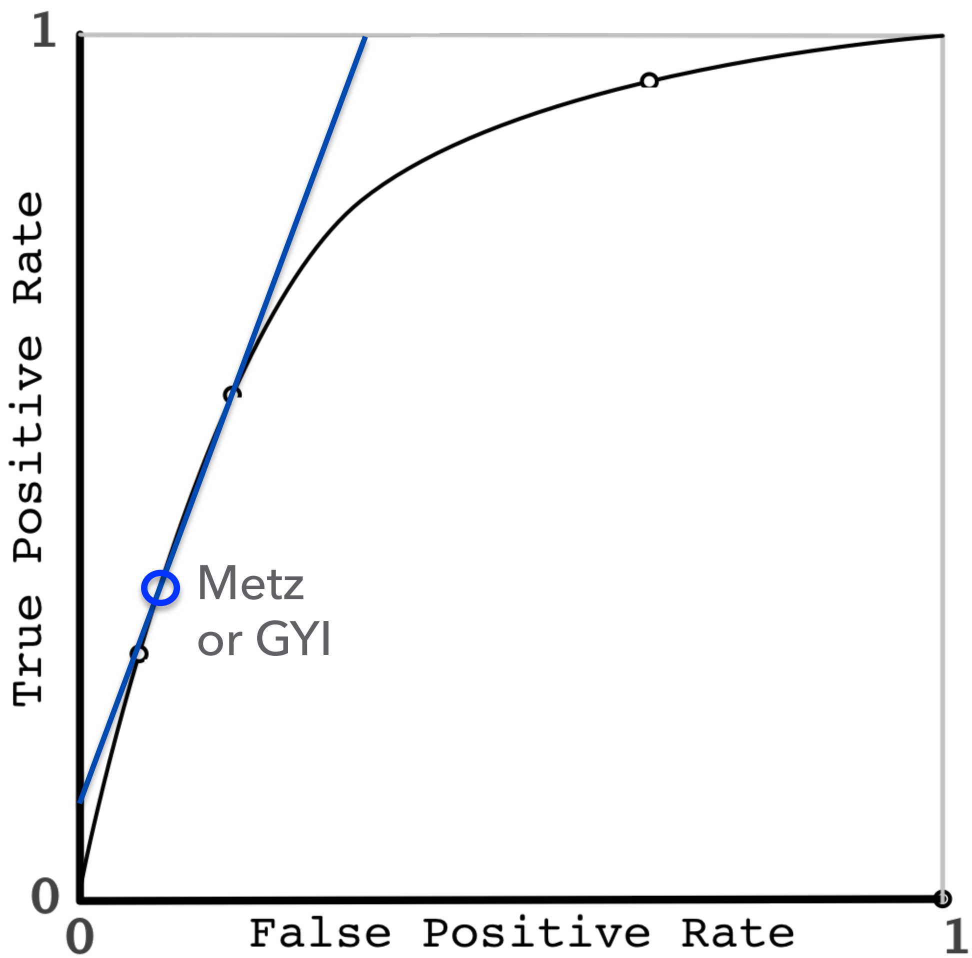


~~Youden's index~~[11]
~~is typical.~~

But Youden's Index assumes costFN = costFP !
and ignores prevalence!

Usually costFN > costFP
and data imbalanced

2. We can also pick a threshold to use

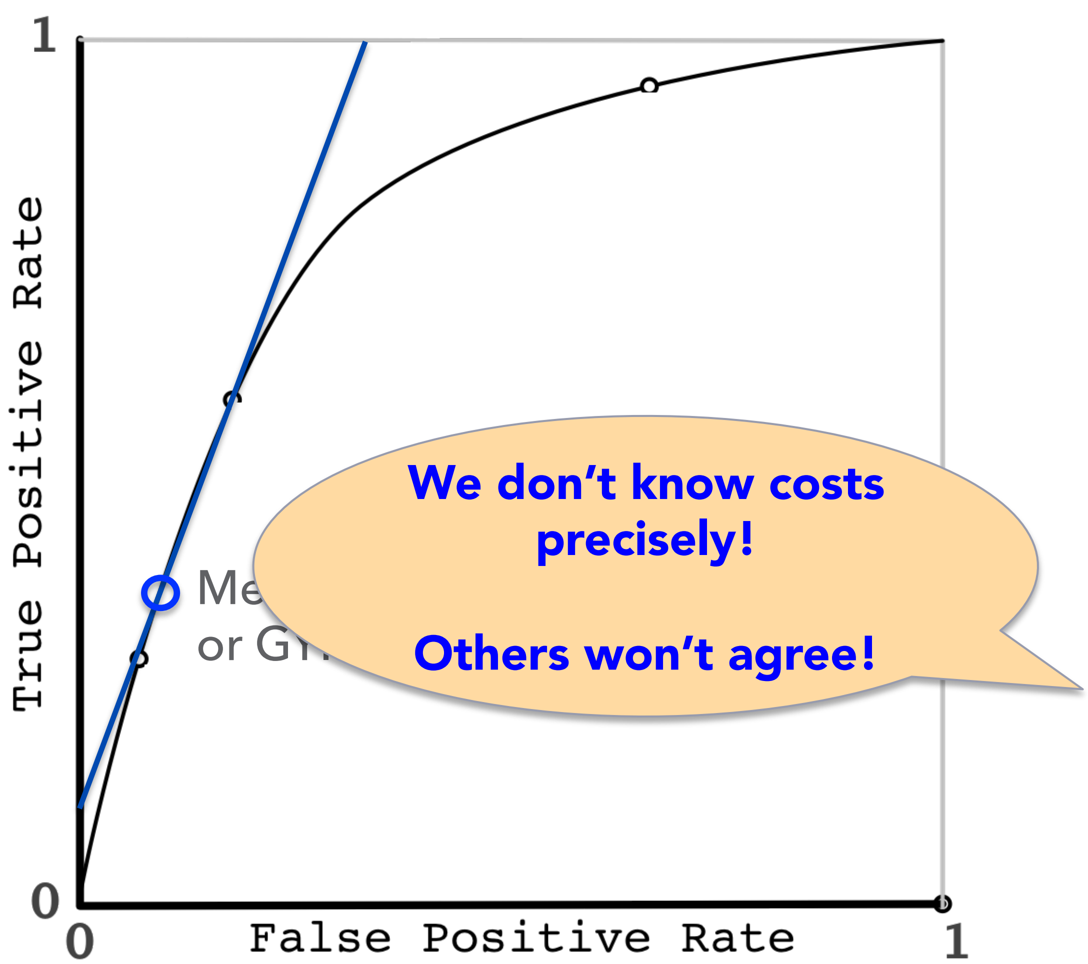


~~Youden's index~~[11]
~~is typical.~~

Use Metz's formula[7,12] or a (fully) generalized Youden's index[13,14,15].

But using correct formulas with default equal costs is still wrong!

2. We can also pick a threshold to use

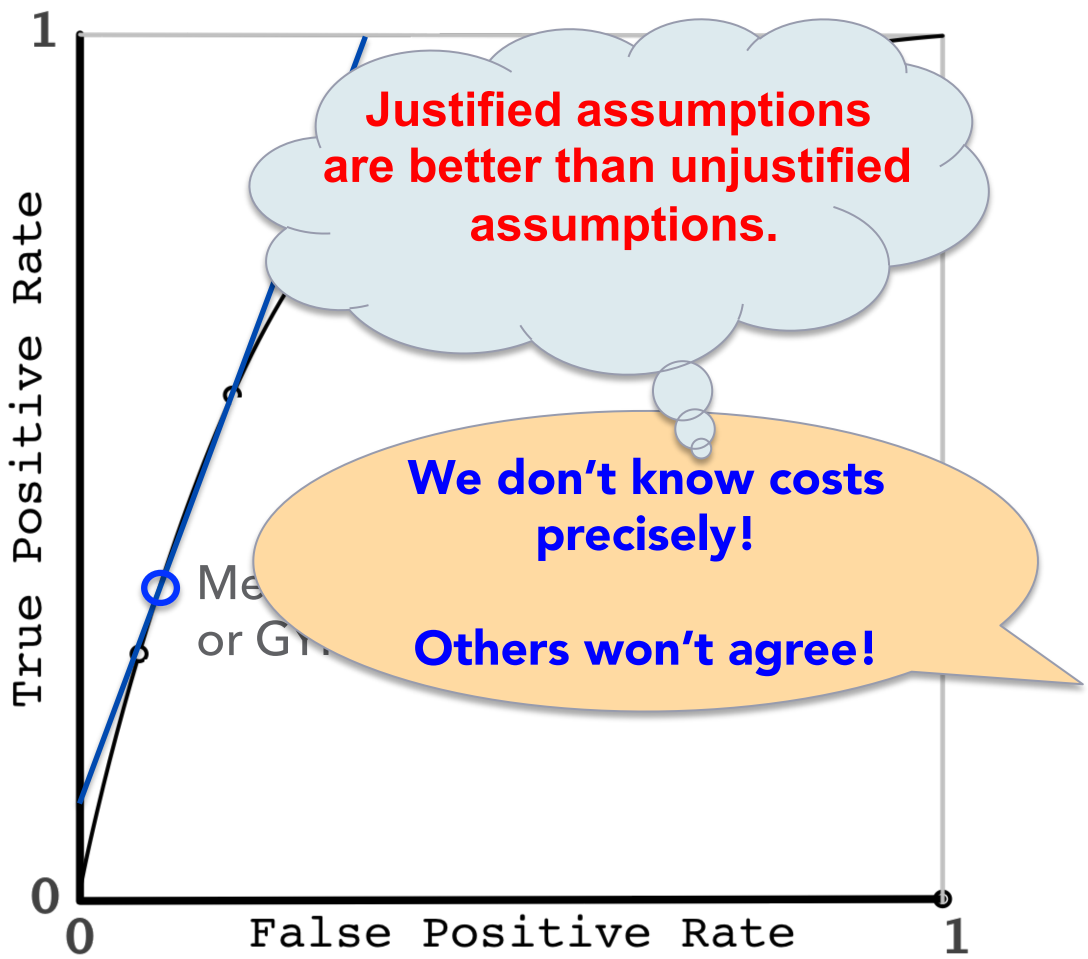


~~Youden's index~~[11] ~~is typical.~~

Use Metz's formula[7,12] or a (fully) generalized Youden's index[13,14,15].

But using correct formulas with default equal costs is still wrong!

2. We can also pick a threshold to use



~~Youden's index~~[11]
~~is typical.~~

Use Metz's formula[7,12] or a (fully) generalized Youden's index[13,14,15].

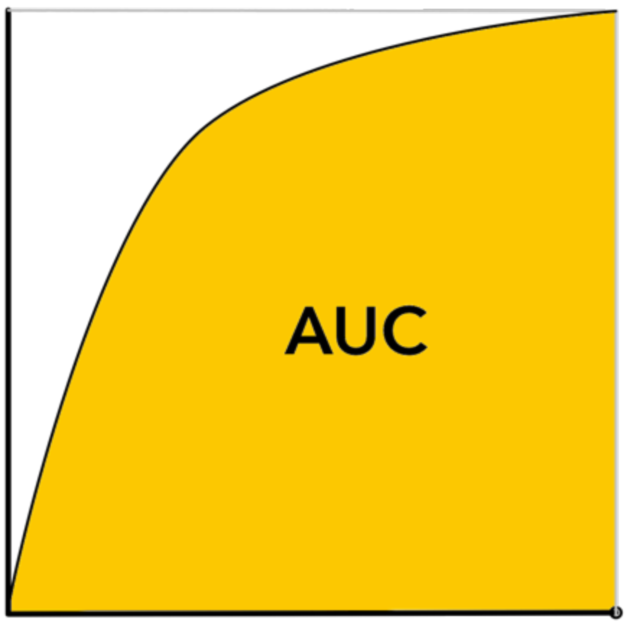
But using correct formulas with default equal costs is still wrong!

How do we fix those problems and concerns?

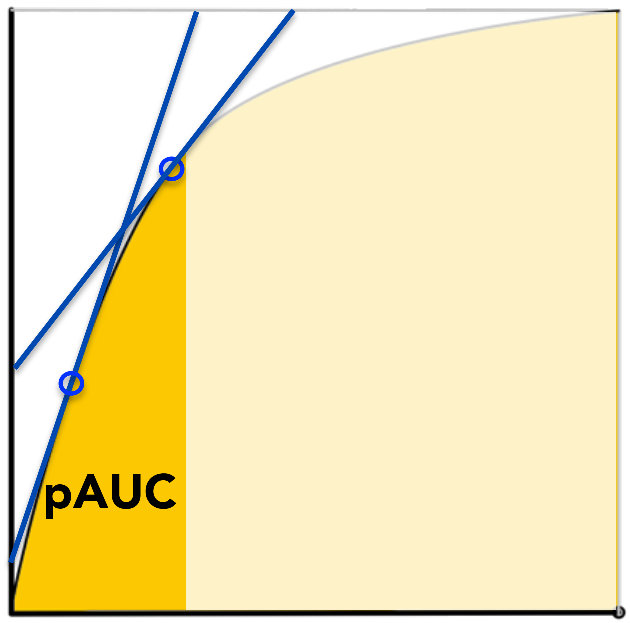
Instead of assuming **one** cost (risk) for all patients

We need something in between!

Or assuming all possible choices of cost (risk) are relevant



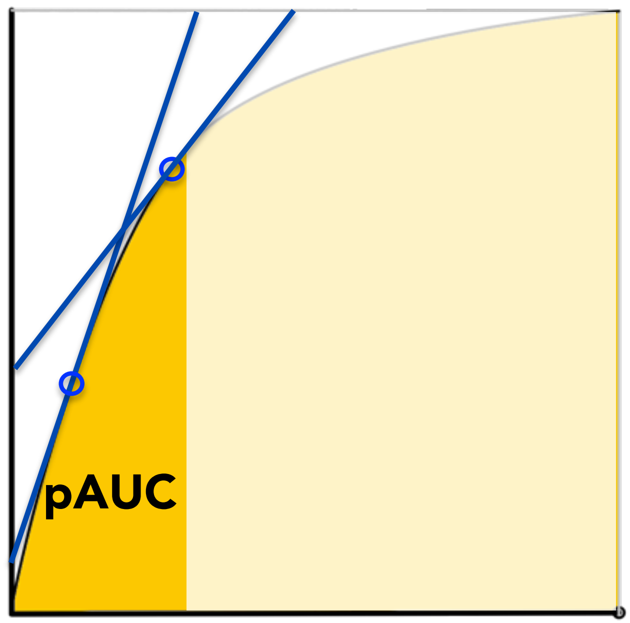
Partial AUC[8,16] (pAUC) in between is better!



Focuses on a clinically relevant region (not a single point).

Allows choice of costs (risk) or uncertainty of costs in a region:
- specific to a patient
- specific to a doctor

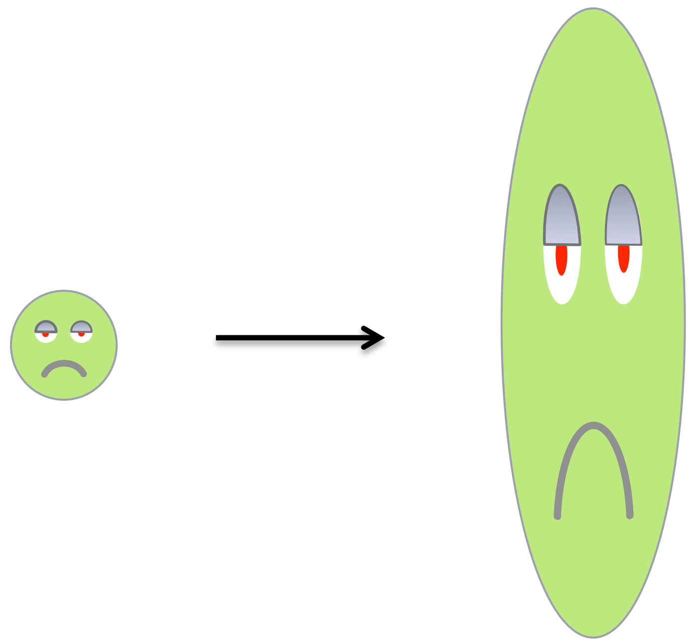
Partial AUC[8,16] (pAUC)



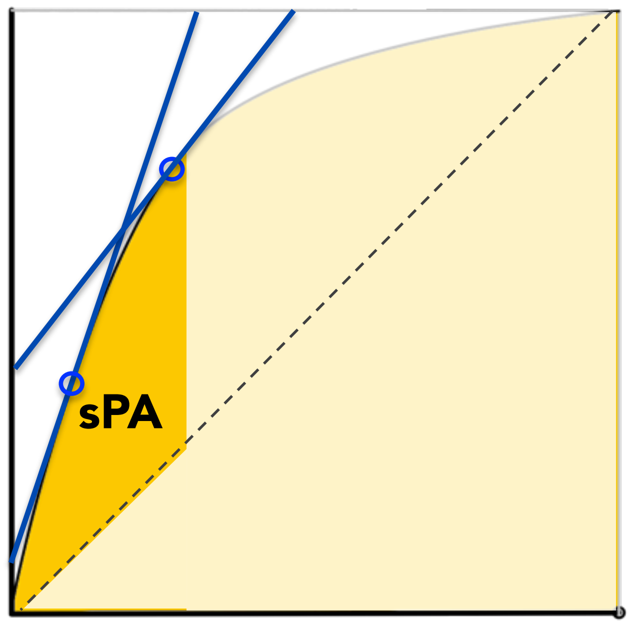
But partial AUC is flawed[21-23]:

- Biased to positives (the vertical axis)
- not interpretable as a c statistic
- increases monotonically with FPR

The bias of pAUC on the vertical (positives) is like a magnifying glass that distorts



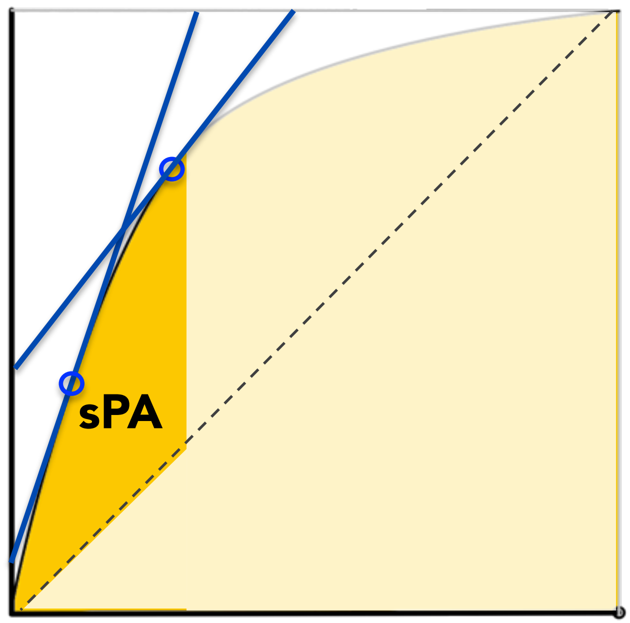
The standardized partial area[17] (sPA) is better but has one flaw[21,23] and one shortcoming[23]



sPA[17] fixes some problems with the partial AUC:

- **Avoids** positive only (vertical) focus*
- not interpretable as a c statistic[23]
- **Fixes** monotonic increase with FPR
- fails for improper ROC curves[21,23]

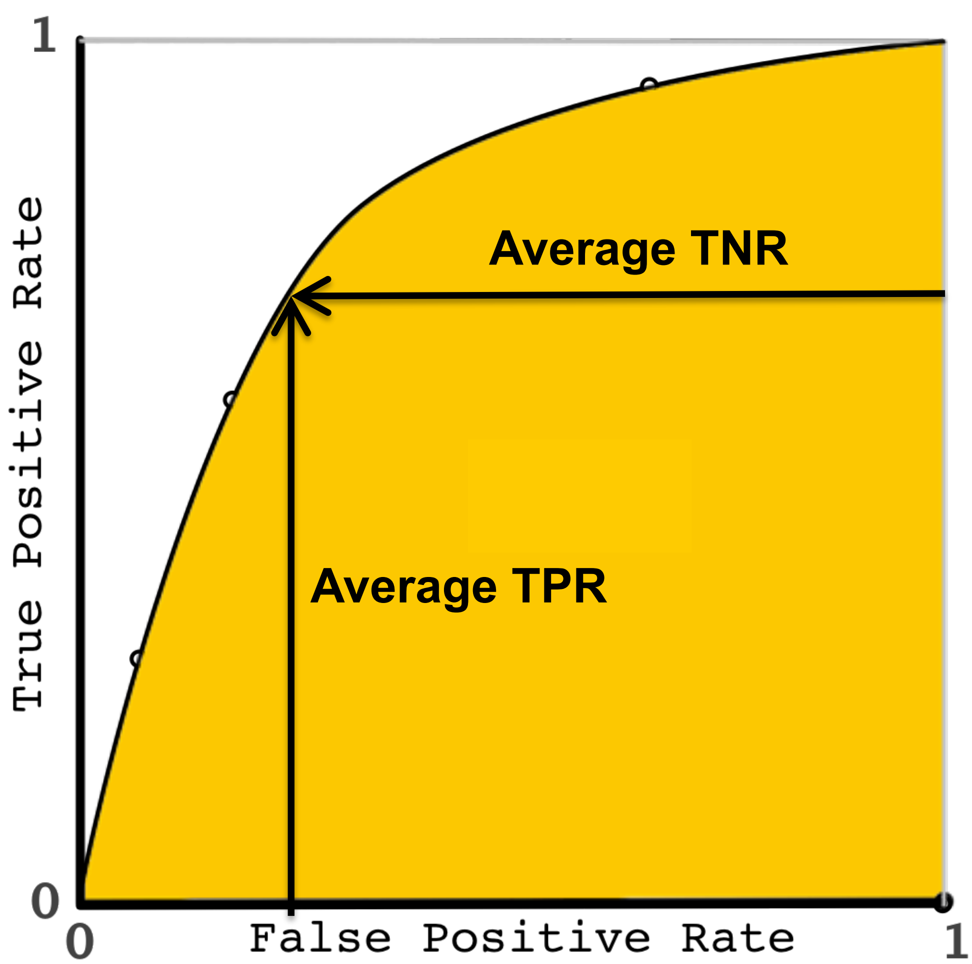
Other partial area measures[9,18-21] also fall short.



Of existing alternatives:
PAI[9], PAI_m[17], sPA[17], half-AUC[18], two-way AUC[19], novel PAI[20] and tighter sPA[21], **none can be interpreted as a c statistic. None are proper analogies to AUC and c.**

Some also have fixed bounds or only address specific questions.

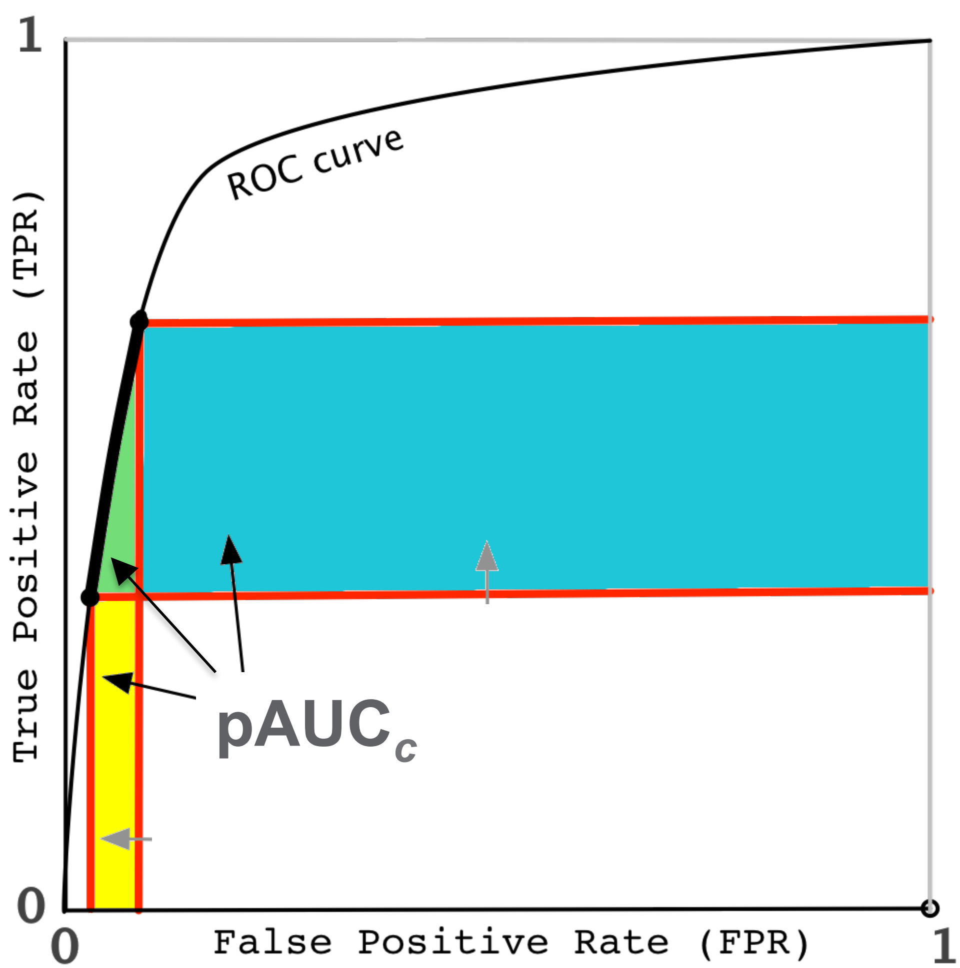
A partial measure analog to AUC must recognize its horizontal, vertical and *c* statistic interpretations[23]



AUC = average TPR[29]
= average TNR[29]
= *c* statistic[29]

AUC computation only uses average TPR since it is redundant for the whole curve, **<u>but not a partial curve!</u>**[23]

We propose a concordant partial AUC[23], $pAUC_c$



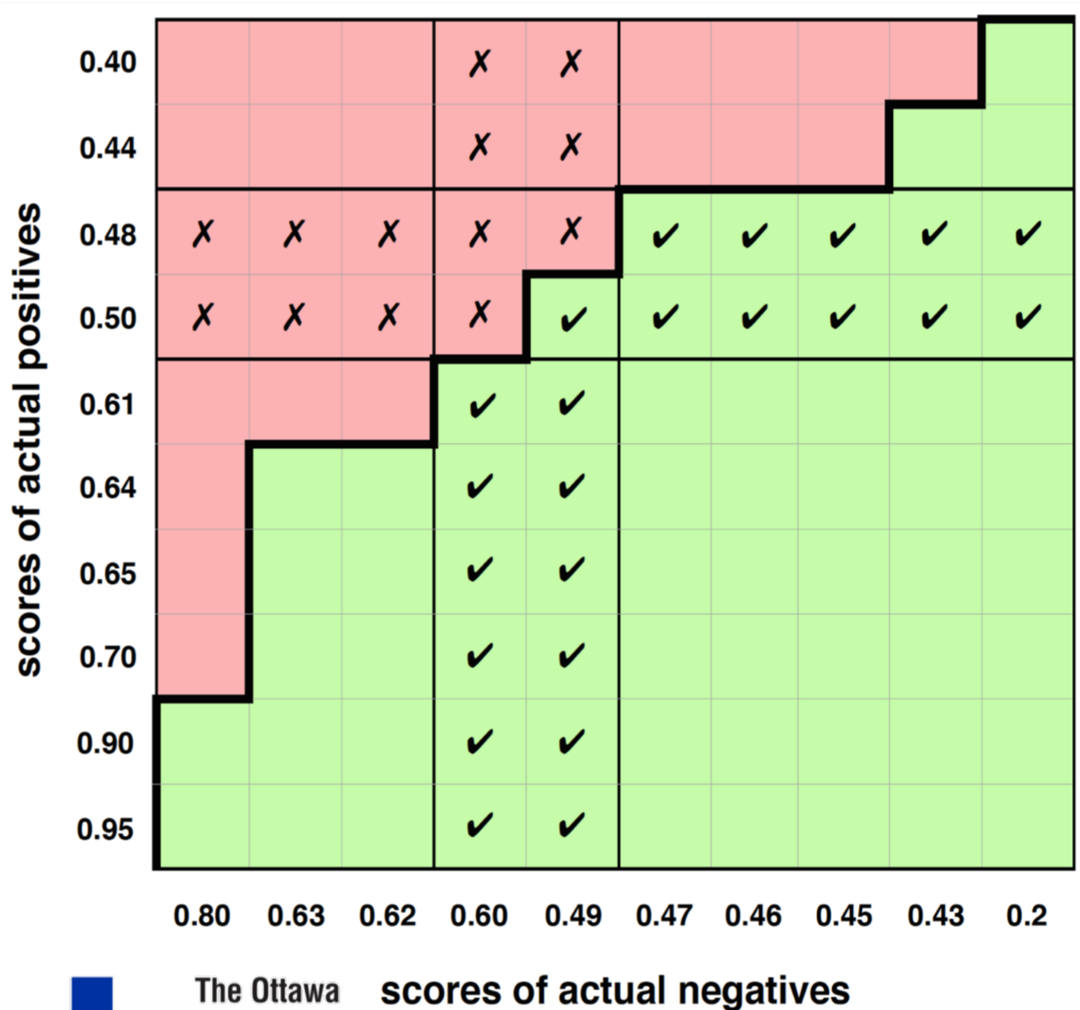
In a partial curve the horizontal and vertical areas are not the same.

Our measure has all three interpretations: horizontal, vertical and *c* statistic as a proper analogy to AUC.

The Ottawa Hospital RESEARCH INSTITUTE | L'Hôpital d'Ottawa INSTITUT DE RECHERCHE

Affiliated with • Affilié à uOttawa

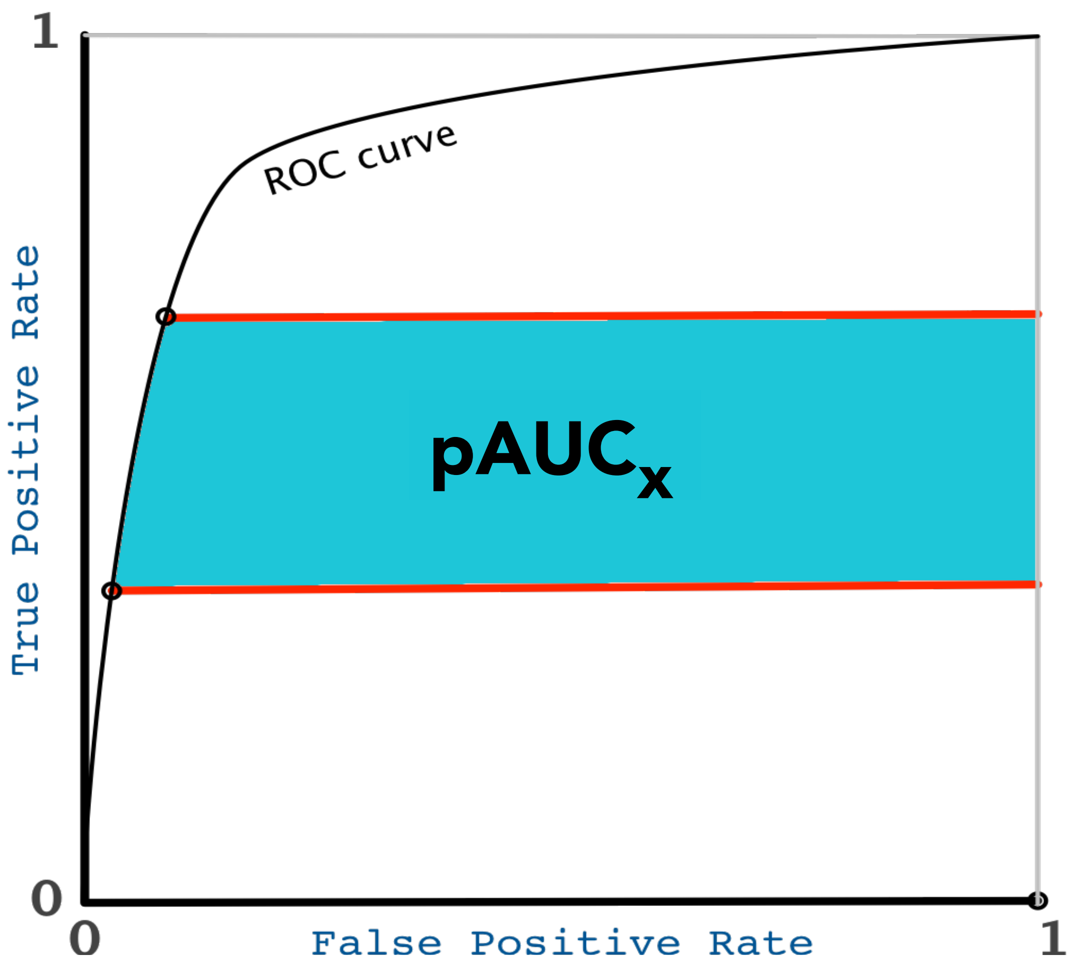
It equals a proposed partial *c* statistic[23] c_Δ for ROC data



First concordance interpretation for a partial ROC curve.

The meanings of the axes and instances (patients) along them are important for interpretation.

We also need (and propose) a horizontal partial AUC[23], pAUC$_x$, with free bounds



The partial area index[9] has one boundary fixed at TPR=1.0

As part of measuring pAUC$_c$ we need a horizontal partial AUC with free bounds per Walter's suggestion[10].

Before our work

Measure Type	Positives	Negatives	Positives and Negatives
Point	True positive rate (Sensitivity)	True negative rate (Specificity)	Accuracy
Whole Area, *c* statistic	AUC = Avg TPR	AUC = Avg TNR	AUC = c
Partial Area	Partial AUC = Local Avg TPR	Partial Area Index*	Partial AUC
Partial *c* statistic	Placement values (positive)	Placement values (negative)	?

Existing partial *c* is for other purposes[30-33]

*PAI[9] requires a fixed right boundary of FPR=1

The Ottawa Hospital RESEARCH INSTITUTE | L'Hôpital d'Ottawa INSTITUT DE RECHERCHE

Affiliated with • Affilié à uOttawa

After our work[23]

Measure Type	Positives	Negatives	Positives and Negatives
Point	True positive rate (Sensitivity)	True negative rate (Specificity)	Accuracy
Whole Area, *c* statistic	AUC = Avg TPR	AUC = Avg TNR	AUC = c
Partial Area	Partial AUC = Local Avg TPR	Horizontal Partial AUC ✓*	Concordant Partial AUC ✓
Partial *c* statistic	Placement values (positive)	Placement values (negative)	Partial c for ROC ✓

*Both boundaries for the partial area are free.

This matters because…

Data may have implicit bias or cause bias, and we can compensate for that, but measures should not have bias unless paired and clearly understood as positive/negative

Unbiased measure

Positives and Negatives
Accuracy
AUC = c
Concordant Partial AUC ✓
Partial c for ROC ✓

We tested our measures with data

- Ljubljana breast cancer data (remission at 1 year)
- Wisconsin breast cancer data
- Classic ROC example data from Fawcett
- Classic ROC example modified for imbalance

Differences in AUC are minimal when AUC ≈ 97% (Wisconsin)---Ljubljana provides better examples.

Results

- We validated the theory of our measures on all 4 data sets. Results for three partial curves add up to the whole.
  - $\Sigma \text{ pAUC}_c = \text{AUC}$
  - $\Sigma \text{ c}_\Delta = \text{AUC}$
  - $\text{pAUC}_c = \text{c}_\Delta$
  - $\widetilde{\text{pAUC}}_c = \tilde{\text{c}}_\Delta$

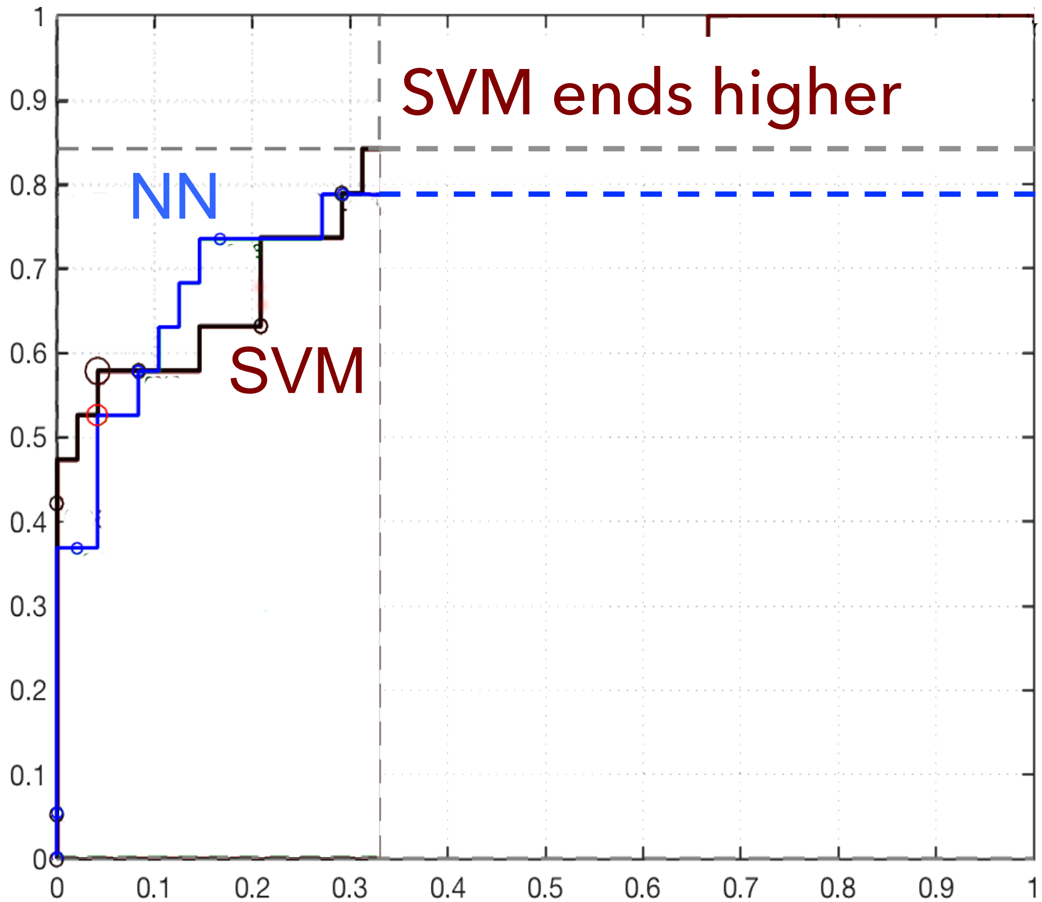
- We also interpreted our measures in comparison to other measures (next).

The Ottawa Hospital | L'Hôpital d'Ottawa
RESEARCH INSTITUTE | INSTITUT DE RECHERCHE

Affiliated with • Affilié à uOttawa

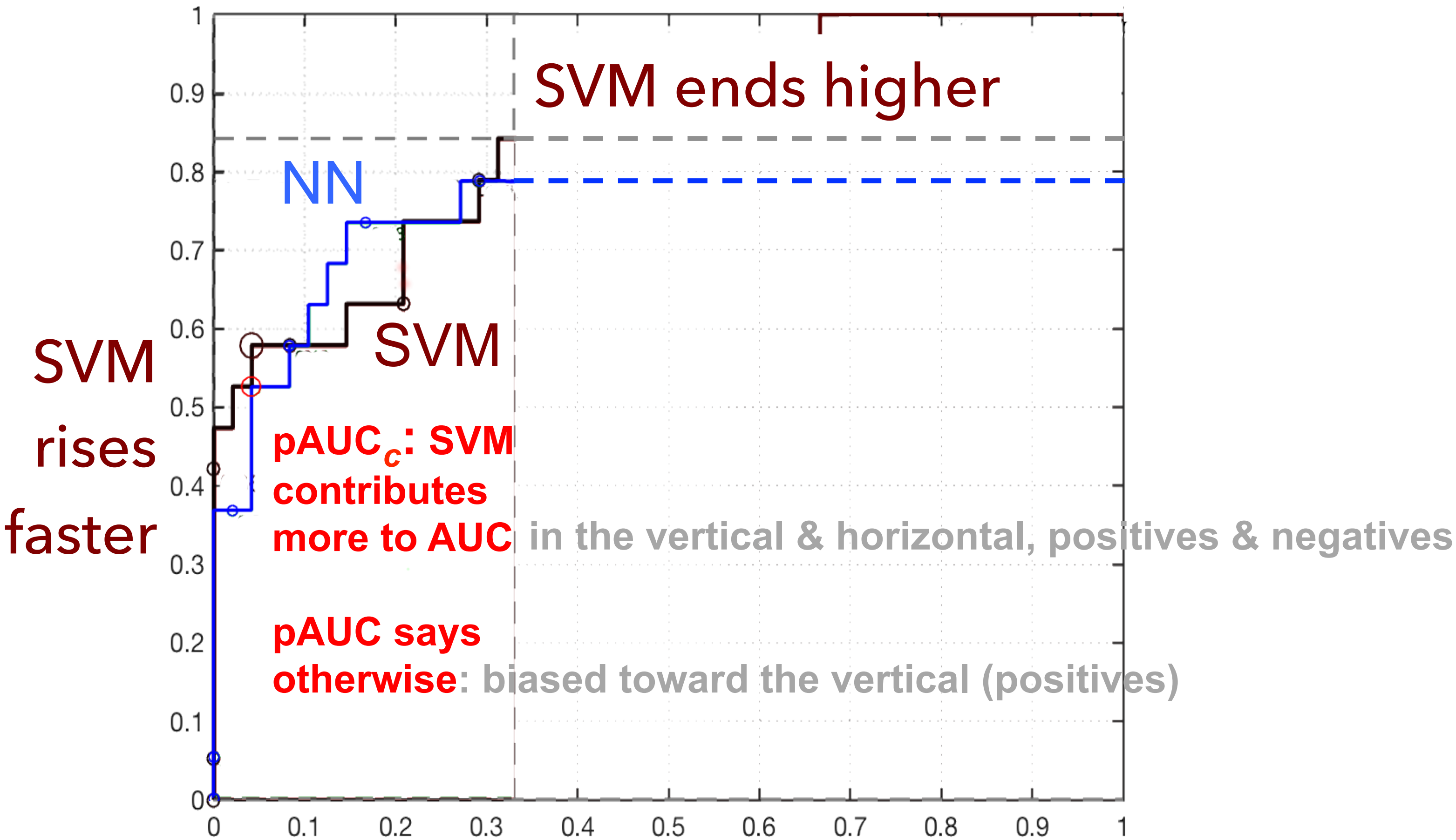
In Ljubljana results the NN* and SVM* ROC curves cross twice in the first partial curve FPR=[0,0.33]

SVM rises faster



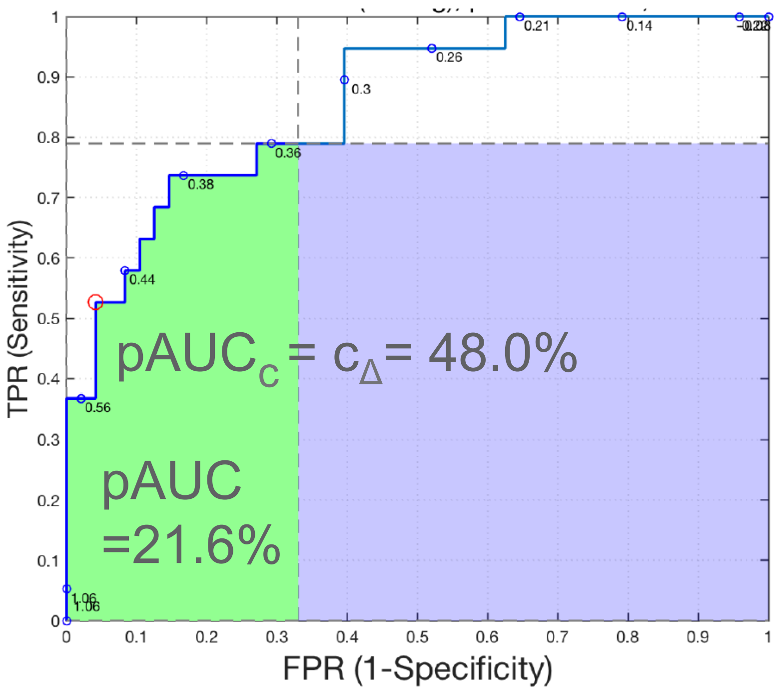
*NN=neural network and SVM=support vector machine

In Ljubljana results the NN* and SVM* ROC curves cross twice in the first partial curve FPR=[0,0.33]

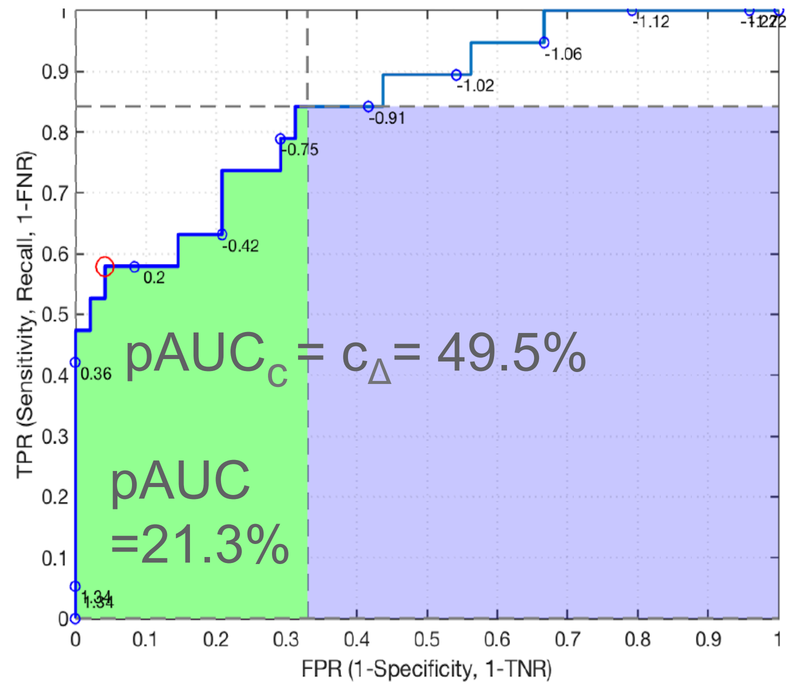


*NN=neural network and SVM=support vector machine

Another view of that



(a) A neural network (NN)



(b) A support vector machine (SVM)

In the first partial curve $pAUC_c \sim AUPRC_+$

Measures	LDA*	LogR*	SVM	NN
Whole Area				
AUC	82.9%	77.1%	<u>84.8%</u>	**86.0%**
$AUPRC_+$	60.9%	53.5%	**72.2%**	<u>71.0%</u>
$AUPRC_-$	<u>54.5%</u>	**56.7%**	53.7%	53.3%
Partial Area $i = 1$				
sPA	75.0%	69.2%	<u>78.8%</u>	**79.2%**
$pAUC$	19.2%	16.0%	<u>21.3%</u>	**21.6%**
$pAUC_c$	47.5%	37.2%	**49.5%**	<u>48.0%</u>

*LDA=linear discriminant analysis, LogR=logistic regression

Summary

- Our measures[23]:
  - are generalizations of AUC and c
  - have no bias re positives vs negatives (people)
  - improve interpretation & measures of partial curves
  - Improve understanding of AUC and c equivalence

Toward equitable, explainable and optimal AI

Future work

- A <u>short</u> paper with more <u>clinical</u> examples
- Use of the measure in a research study
- Studying, benchmarking decision-making thresholds

References

1. Michalski RS, Mozetic I, Hong J, Lavrac N. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence. 1041-1045. 1986.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019 Jan;25(1):44-56.
3. El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. Journal of medical Internet research. 2012;14(1):e33.
4. Lobo JM, Jiménez‐Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography. 2008 Mar;17(2):145-51.
5. Wagstaff K. Machine learning that matters. arXiv preprint arXiv:1206.4656. 2012 Jun 18.
6. Peterson AT, Papeş M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological modelling. 2008 Apr 24;213(1):63-72.
7. Metz CE. Basic principles of ROC analysis. In Seminars in nuclear medicine 1978 Oct 1 (Vol. 8, No. 4, pp. 283-298). WB Saunders.
8. McClish DK. Analyzing a portion of the ROC curve. Medical Decision Making. 1989 Aug;9(3):190-5.
9. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology. 1996 Dec;201(3):745-50.

References

10. Walter SD. The partial area under the summary ROC curve. Statistics in medicine. 2005 Jul 15;24(13):2025-40.
11. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.
12. Bettinger R. Cost-sensitive classifier selection using the ROC convex hull method. SAS Institute. 2003:1-2.
13. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's Index. Statistics in medicine. 1996 May 30;15(10):969-86.
14. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Preventive veterinary medicine. 2000 May 30;45(1-2): 23-41.
15. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology. 2005 Jan 1:73-81.
16. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. Statistics in Medicine. 1989 Oct;8(10):1277-90.
17. McClish DK. Evaluation of the accuracy of medical tests in a region around the optimal point. Academic radiology. 2012 Dec 1;19(12):1484-90.
18. Bradley AP. Half-AUC for the evaluation of sensitive or specific classifiers. Pattern Recognition Letters. 2014 Mar 1;38:93-8.

References

19. Yang H, Lu K, Lyu X, Hu F. Two-way partial AUC and its properties. Statistical methods in medical research. 2019 Jan;28(1):184-95.
20. Wu T, Huang H, Du G, Sun Y. A novel partial area index of receiver operating characteristic (ROC) curve. In Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment 2008 Mar 6 (Vol. 6917, p. 69170B). International Society for Optics and Photonics.
21. Vivo JM, Franco M, Vicari D. Rethinking an ROC partial area index for evaluating the classification performance at a high specificity range. Advances in Data Analysis and Classification. 2018 Sep 1;12(3):683-704.
22. Ma H, Bandos AI, Rockette HE, Gur D. On use of partial area under the ROC curve for evaluation of diagnostic performance. Statistics in medicine. 2013 Sep 10;32(20):3449-58.
23. Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen H, Mayr F, Manuel DG. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Medical Informatics and Decision Making 2020 20(4) doi:10.1186/s12911-019-1014-6
24. Sharma SK, Saini N, Chwla Y. Hepatitis B virus: inactive carriers. Virology Journal. 2005 Dec;2(1):82.
25. McCarthy M. US melanoma prevalence has doubled over past 30 years, BMJ 2015;350:h3074
26. Steyerberg EW, Kattan MW, Gonen M, Obuchowski N, Pencina MJ, Vickers AJ, Gerds T, Cook NR. Assessing the Performance of Prediction Models: a Framework for Some Traditional and Novel Measures. Epidemiology. 2009 21(1):128-138. doi:10.1097/ede.0b013e3181c30fb2

References

27. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. European Heart Journal. 2014 35(29):1925-1931. doi:10.1093/eurheartj/ehu207. arXiv:1011.1669v3
28. Steyerberg EW. Clinical Prediction Models. Springer, 2009.
29. Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. John Wiley & Sons; 2009 Sep 25.
30. Hu Y-C, Chen C-J. A promethee-based classification method using concordance and discordance relations and its application to bankruptcy prediction. Information Sciences. 2011 181(22):4959-4968.
31. Joerin F, Musy A. Land management with GIS and multicriteria analysis. International transactions in operational research. 2000 7(1):67-78.
32. Legendre P. Species associations: the kendall coefficient of concordance revisited. Journal of agricultural, biological, and environmental statistics. 2005 10(2):226
33. Mendas A, Delali A. Integration of multicriteria decision analysis in GIS to develop land suitability for agriculture: Application to durum wheat cultivation in the region of Mleta in Algeria. Computers and Electronics in Agriculture 2012 83:117-126
34. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. British Medical Journal. 2012 Jul 2;345:e3999.

References

35. Lusted LB. Introduction to medical decision making. American Journal of Physical Medicine & Rehabilitation. 1970 Oct 1;49(5):322.
36. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging. 1989 Jan 1;29(3):307-35.
37. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006 Jun 1;27(8):861-74.

André Carrington

Doug Manuel

Doug Manuel Lab

Clinical Epidemiology

Ottawa Hospital Research Institute

Thanks and acknowledgments to

Co-authors on the related paper:

Paul Fieguth, Hammad Qazi, Andreas Holzinger, Franz Mayr, Helen Chen, Doug Manuel

André Carrington was generally funded by: (2015-19)

Data are often imbalanced

Imbalanced classes: #positives ≠ #negatives

1:3	Breast cancer[1]
5:100	Hepatitis B[24]
2:10 000	Melanoma[25], fraud

Ignoring negatives is like considering

Sensitivity ***without*** Specificity

Ignoring negatives is like considering

Sensitivity	***without***	Specificity
Positive predictive value (PPV)	***"***	Negative predictive value (NPV)

Ignoring negatives is like considering

Sensitivity	***without***	Specificity
Positive predictive value (PPV)	"	Negative predictive value (NPV)
Likelihood ratio positive (LR+)	"	Likelihood ratio negative (LR−)

Ignoring negatives is like considering

Sensitivity	***without***	Specificity
Positive predictive value (PPV)	"	Negative predictive value (NPV)
Likelihood ratio positive (LR+)	"	Likelihood ratio negative (LR−)
Average precision (AP = AUPRC+)	"	AUPRC−

Which we cannot do in medicine