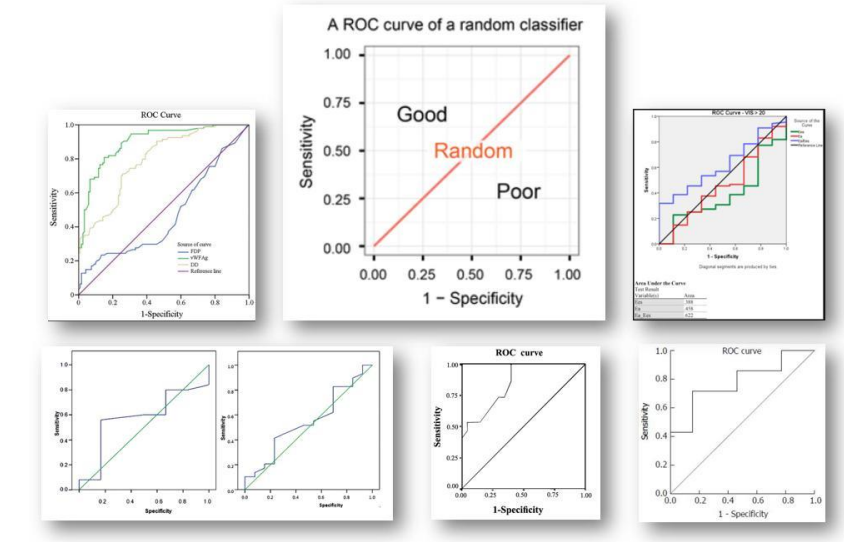
ORIGINAL THREAD: https://twitter.com/cecilejanssens/status/1104134423673479169

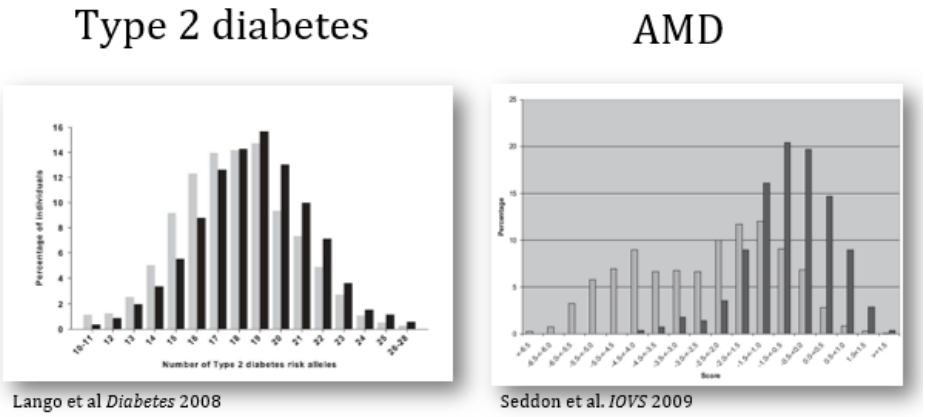
Cecile Janssens, Professor of epidemiology, Emory University, Atlanta USA

The area under the ROC curve (AUC) is so frequently criticized and misunderstood that I often wonder whether I am the metric's only fan. Let me explain why how I see and value the AUC.



(don't blame the metric)

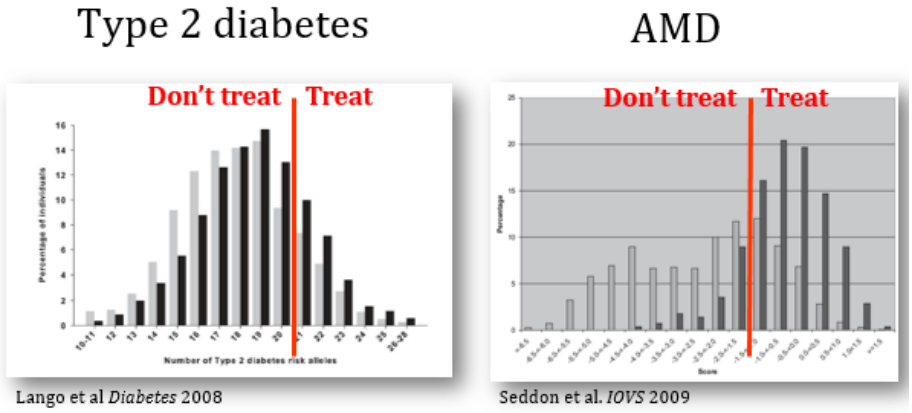
When we make a risk model/score and calculate predicted risks for all people in the study, we can plot distributions of risks for people who will develop the disease and those who will not (let's call them patients and nonpatients): scores on x-axis, frequency on y-axis.



Lango et al *Diabetes* 2008

Seddon et al. *IOVS* 2009

We want these risk distributions to be separated so that, when we use the risk model to select high-risk individuals for treatment, we are more likely to select people who will benefit from it.



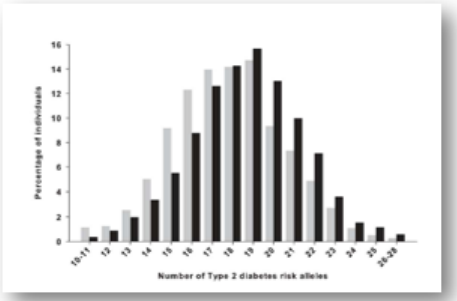
Lango et al *Diabetes* 2008

Seddon et al. *IOVS* 2009

More separation between the distributions means that patients tend to have higher risks than nonpatients, or … that a random patient has a higher risk than a random nonpatient. This pair must sound familiar.

AUC is a metric for the degree of separation between these two risk distributions. Nothing more and nothing less.

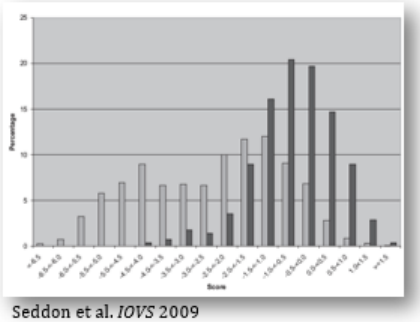
Type 2 diabetes



Lango et al *Diabetes* 2008

AUC = 0.60

AMD



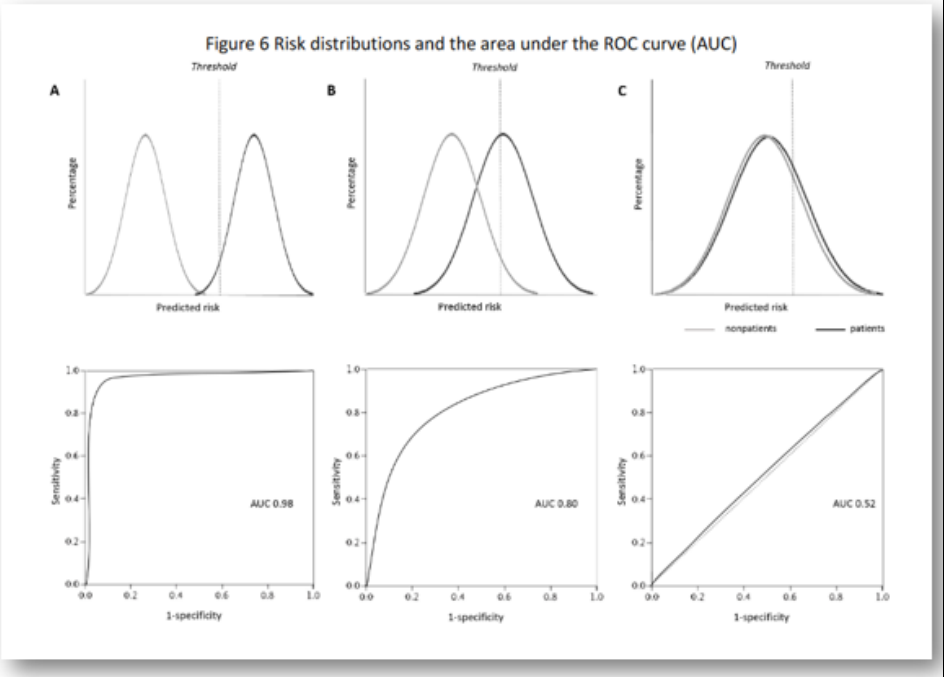
Seddon et al. *IOVS* 2009

AUC = 0.76

AUC = degree of separation between risk distributions of affected and unaffected individuals—nothing more, nothing less

0.50: complete overlap ~ random prediction
1.0: complete separation ~ perfect prediction

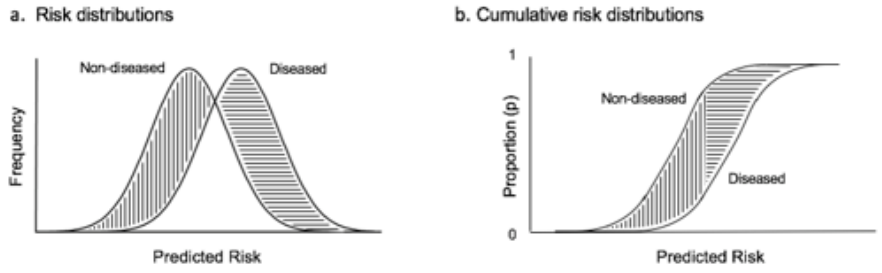
When distributions largely overlap, AUC is close to 0.5, when they are largely separated AUC approaches 1. More separation between the risk distributions means a larger area under the ROC curve. Why does that make sense?

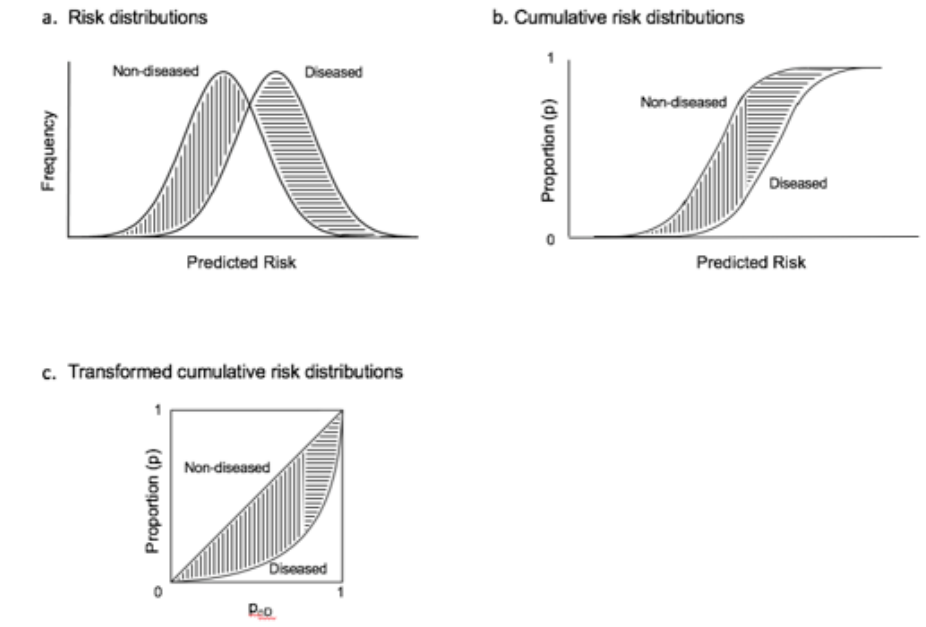
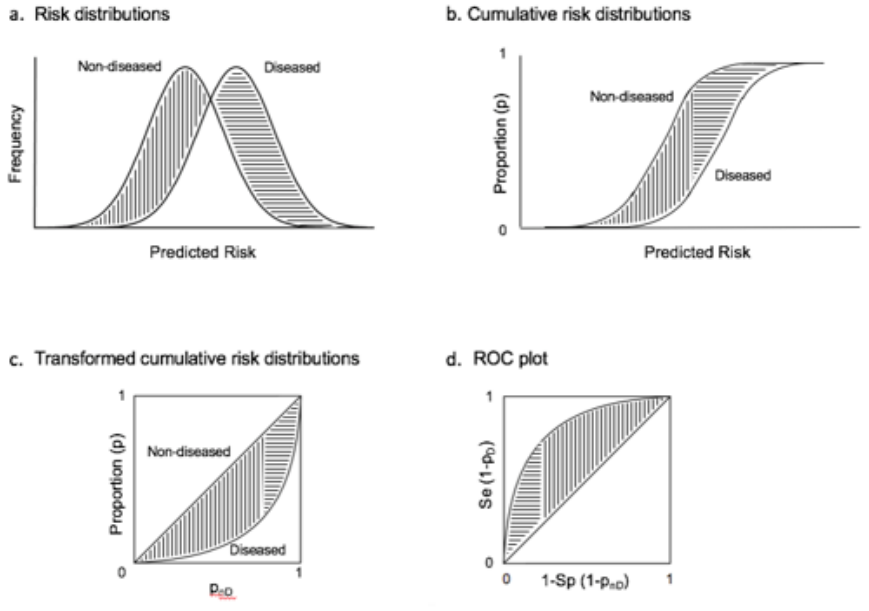


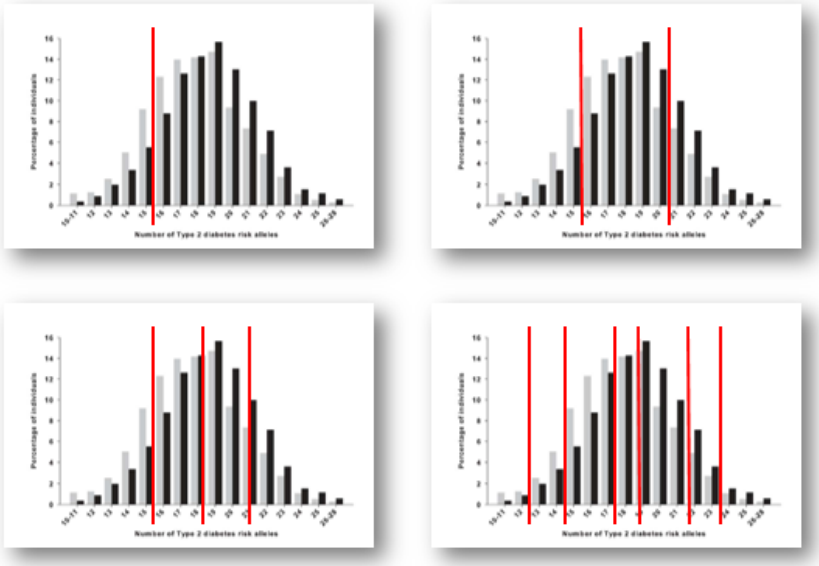
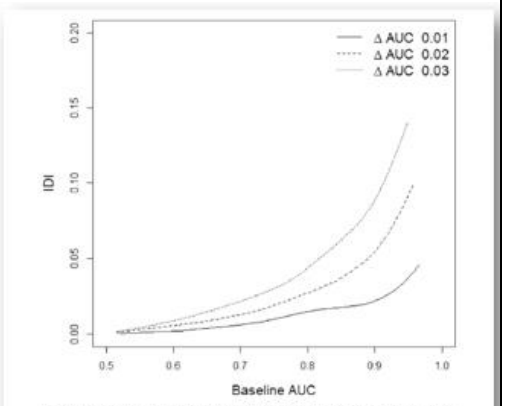
Janssens & Martens, Introduction to prediction research, 2018

Here's how to get from risk distributions to ROC curves in 3 steps. 1. We can transform both risk distributions into cumulative risk distributions (A-->B), where, at each predicted risk, p is the proportion of people with a predicted risk *lower than* that predicted risk.

Figure 1. From risk distributions to the ROC curve



2. Instead of predicted risk on the x-axis we can also present the cumulative proportion of nonpatients at each predicted risk (B-->C). Note that the diagonal line is the cumulative proportion of nonpatients---a straight line because the same proportion is also the x-axis.	**Figure 1.** From risk distributions to the ROC curve  Legend: see https://peerj.com/preprints/3468/
3. If we flip both axis (C-->D), we get the ROC plot. The diagonal line, often referred as the reference line, is still (based on) the cumulative proportion in nonpatients (for more detailed explanation see legend)	**Figure 1.** From risk distributions to the ROC curve 
So: the ROC plot is just an alternative way to present risk distributions and AUC is just a metric how much the distributions are separated. Seeing the ROC and AUC from this perspective shines a different light on some of the criticism.	

It is argued that AUC is only for binary risk categories (one threshold). This sounds true if predicted risks are considered as potential thresholds, but the risk distributions are what they are irrespective of the number of risk categories.	**Risk distributions don't change with more thresholds** 
It is said that AUC does not change even when predicted risks change, which is true. This particularly happens when baseline AUC is high. AUC is a rank test for discriminating between groups. Use other metrics (like IDI) if interest is in improving individual risks.	Mini-Review Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve Nancy R. Cook CONTENT: The ROC curve is typically used to evaluate clinical utility for both diagnostic and prognostic models. This curve assesses how well a test or model discriminates, or separates individuals into two classes, such as diseased and nondiseased. A strong risk predictor, such as lipids for cardiovascular disease, may have limited impact on the area under the curve, called the AUC or c-statistic, even if it alters predicted values. Calibration, measuring whether predicted probabilities agree with observed proportions, is another component of model accuracy important to assess. Reclas-  **Fig. 3.** Integrated discrimination improvement (IDI) for fixed increments in the area under the receiver operating characteristic curve (ΔAUC) by AUC value of the baseline prediction model. In all scenarios, the event rate was 10% and the frequency of the added risk factor was 20%. Martens et al. J Clin Epidemiol 2016

Individuals may move between risk categories when AUC hardly changes. True. This particularly happens for risk thresholds in the center of the risk distribution, where minor changes in predicted risks can move many people to the neighboring category, yet in both directions.

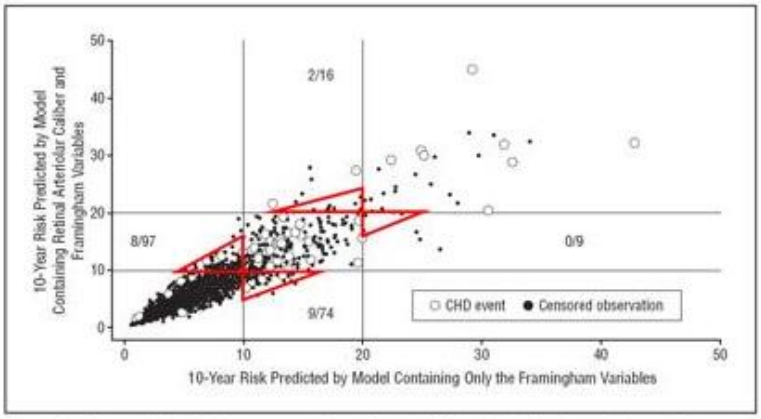


Figure 2. Ten-year risk of incident coronary heart disease (CHD) predicted by the model containing retinal arteriolar caliber and Framingham variables against risk predicted by model containing only the Framingham variables. Data are given as CHD events/total number of events and censored observations (total = 207/4912).

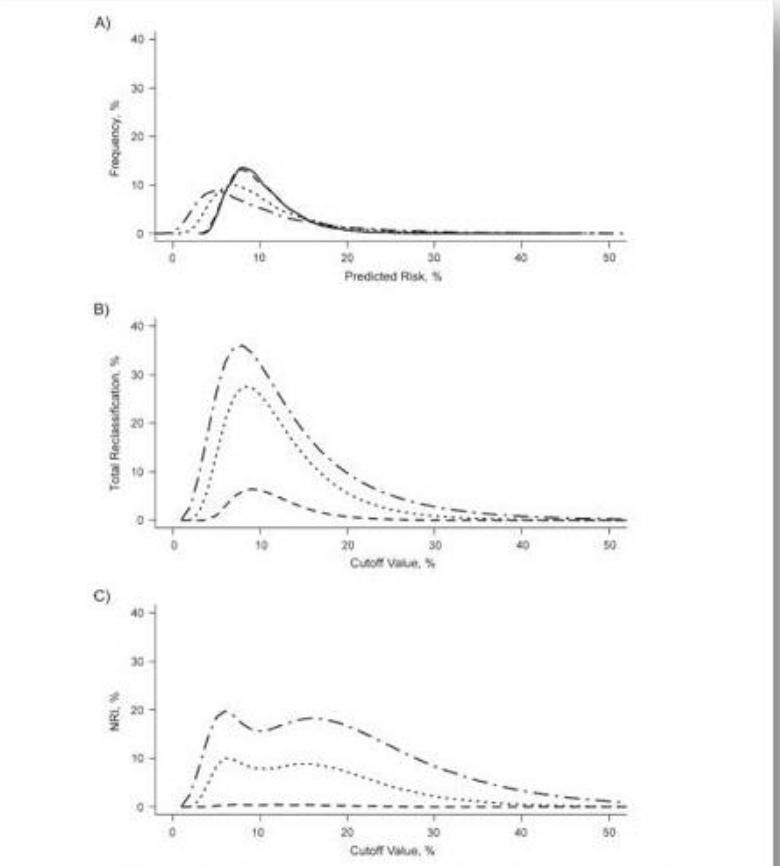


Figure 2. Risk distribution and reclassification measures for prediction models in the simulation study. A) Distribution of predicted risks for the model based on 20 polymorphisms (continuous line) and for the updated models when change in the area under the receiver operating characteristic curve (ΔAUC) is 0.005 (dashed line), 0.05 (dotted line), and 0.10 (dashed-and-dotted line). B) Percentage of total reclassification per cutoff value for risk stratification when ΔAUC is 0.005 (dashed line), 0.05 (dotted line), and 0.10 (dashed-and-dotted line). C) Net reclassification improvement (NRI) per cutoff value for risk stratification when ΔAUC is 0.005 (dashed line), 0.05 (dotted line), and 0.10 (dashed-and-dotted line). Each polymorphism has an odds ratio of 1.2 and a frequency of 20%. Disease risk is 10%, and sample size is 10,000. Percentage of total reclassification and NRI are presented as mean values obtained from 100 simulations. Cutoff values are used to define 2 risk categories; one included individuals with a risk higher than or equal to the cutoff value, and the other included those with a risk lower than the cutoff value. NRI was calculated as the sum of differences in the proportion of individuals moving up minus the proportion moving down for cases, and the proportion of individuals moving down minus the proportion moving up for noncases.

Mihaescu et al. Am J Epidemiol 2010

It occurs to me that researchers often criticize AUC when the metric fails to deliver. If you don't trust AUC, don't use it. Use a metric which results you are willing to accept (and show risk distributions before and after adding variables to the risk model---we know).

Assessment of Improved Prediction Beyond Traditional Risk Factors
When Does a Difference Make a Difference?

A. Cecile J.W. Janssens, PhD; Muin J. Khoury, MD, PhD

(*Circ Cardiovasc Genet.* 2010;3:3-5.)

Reclassification metrics are often used to compensate for disappointing results from AUC analysis

Correct interpretation:

Reclassification ↑	AUC ↑	: prediction better
Reclassification ↑	AUC -	: different errors

Finally, and most importantly: AUC is *not* a measure of utility. What value of AUC is high enough depends entirely on what the risk model will be used for. For some applications 0.65 will be high enough, for others 0.90 might be too low.

Prediction does not be perfect to be useful

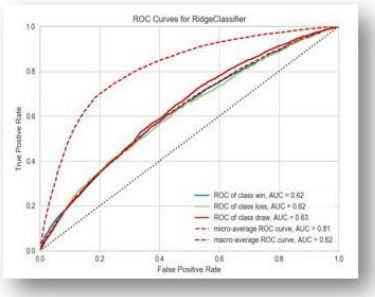
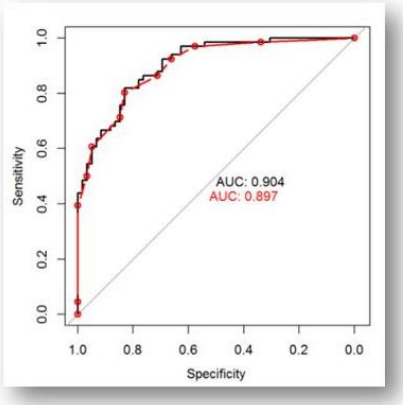
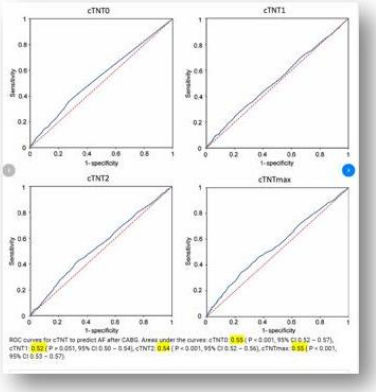
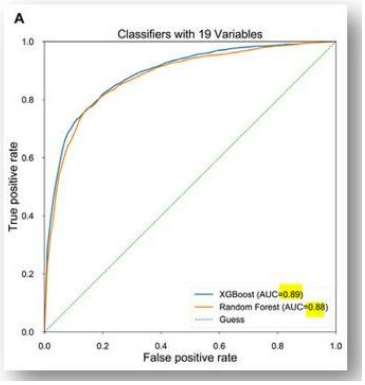
Key: what are you going to do with prediction?

Example of breast cancer prevention:

Improving efficiency and effectiveness of mammography screening by risk stratification: lower AUC may already lead to cost-effective improvement

Recommending prophylactic surgery: higher AUC may not be high enough to prevent unnecessary surgeries

Predictive ability is the beginning,
clinical utility decides what is the end.

So, my rule of thumb: AUC is close to 0.5 -> ROC curve close to diagonal -> risk distributions overlap. Adding variables does not change AUC -> ROC curves overlap -> risk distributions stay the same. There are some exceptions when I would not rely on AUC, but these are rare	   
Research	http://www.cecilejanssens.org/wp-content/uploads/2018/01/PredictionManual2.0.pdf