

# Statistical Smorgasbord: ROC analysis, Power and Sample Size Analysis

Kathleen Torkko  
November 20, 2019

---

# Objectives

Learn how to perform and interpret an ROC analysis

Learn how to compare two or more ROC curves

Understand the components of a power and sample size analysis

# Receiver Operating Characteristic (ROC) Curve

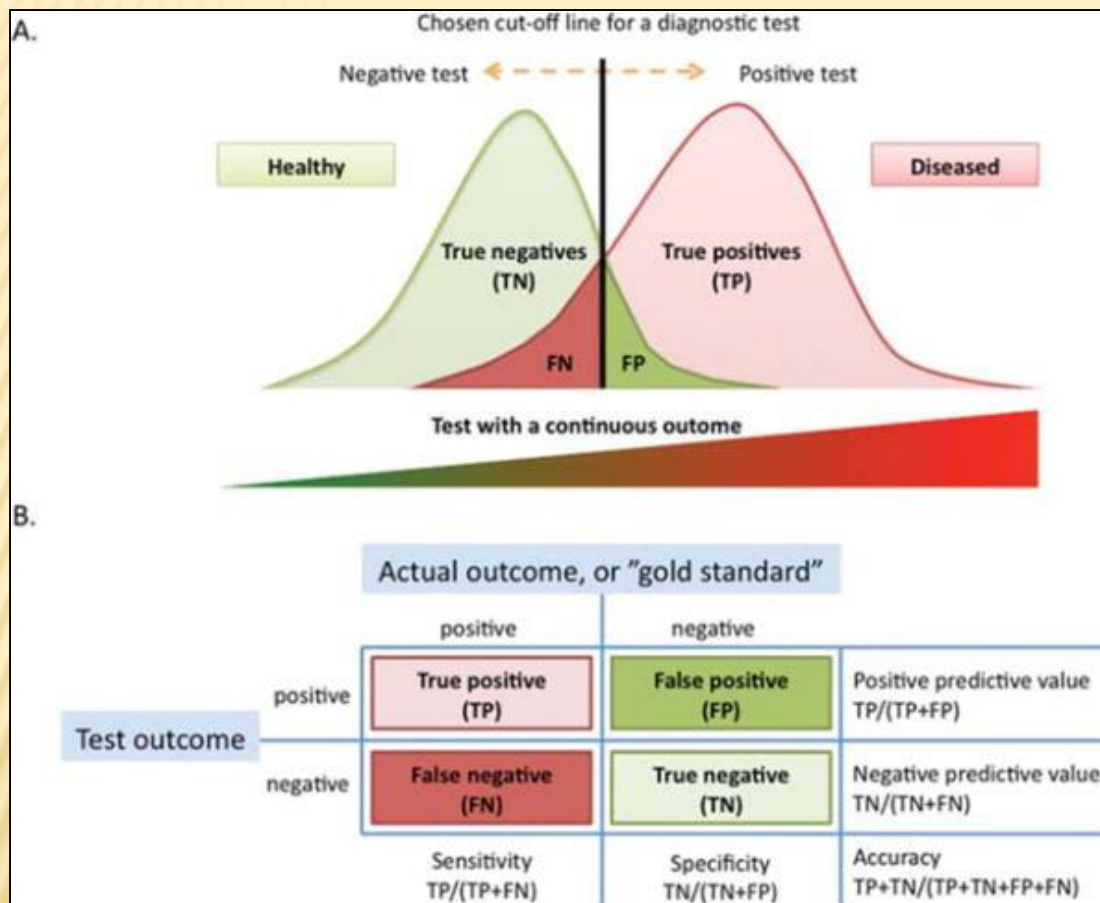
Using sensitivity and specificity to evaluate the ability of a diagnostic test to differentiate between people with disease and those without

Why the funny name?

These plots were originally used in WWII to determine whether blips on a radar screen were ships, planes, or birds or fish





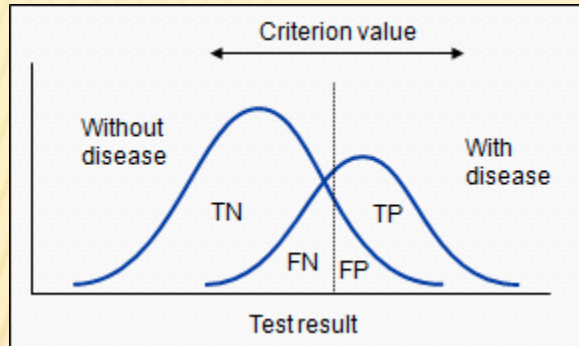


## Diagnostic Accuracy and Receiver-Operating Characteristics Curve Analysis in Surgical Research and Decision Making.

Soreide K et al Annals of Surgery. 2011; 253(1):27-34.

Figure 1 . A, A discriminatory test with fair discriminatory ability between the diseased and healthy population. The two populations show some overlap in test spectrum, which influence the discriminatory ability of the test. B, Depicts a 2 x 2 table with the definitions of the diagnostic descriptors. TP denotes true positives, TN true negatives; FN false negatives; FP false positives.

# Changing Threshold of a Test Changes Sensitivity and Specificity



Numbers of false positives and false negatives also change

If the threshold is too high, then some individuals with disease will be missed

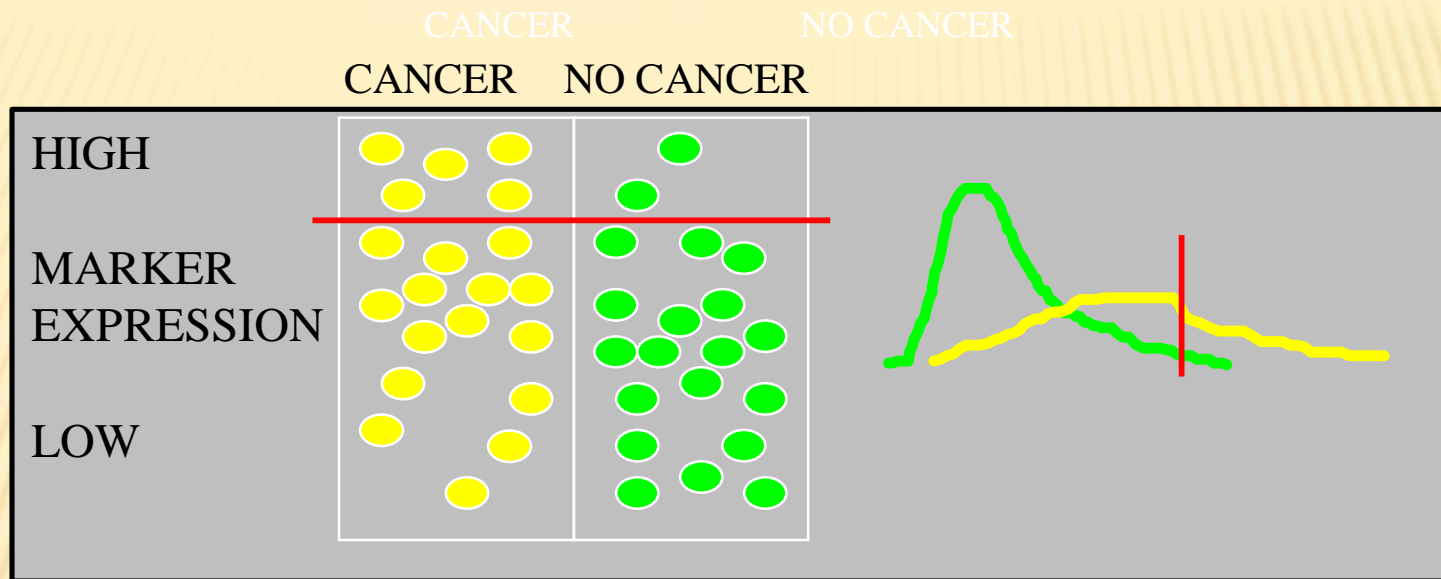
low sensitivity, high specificity

missing some cases = false negatives

If the threshold is too low, most individuals with disease will be detected, but many people without the disease will also have positive tests.

high sensitivity, low specificity

more non-cases picked up = false positives

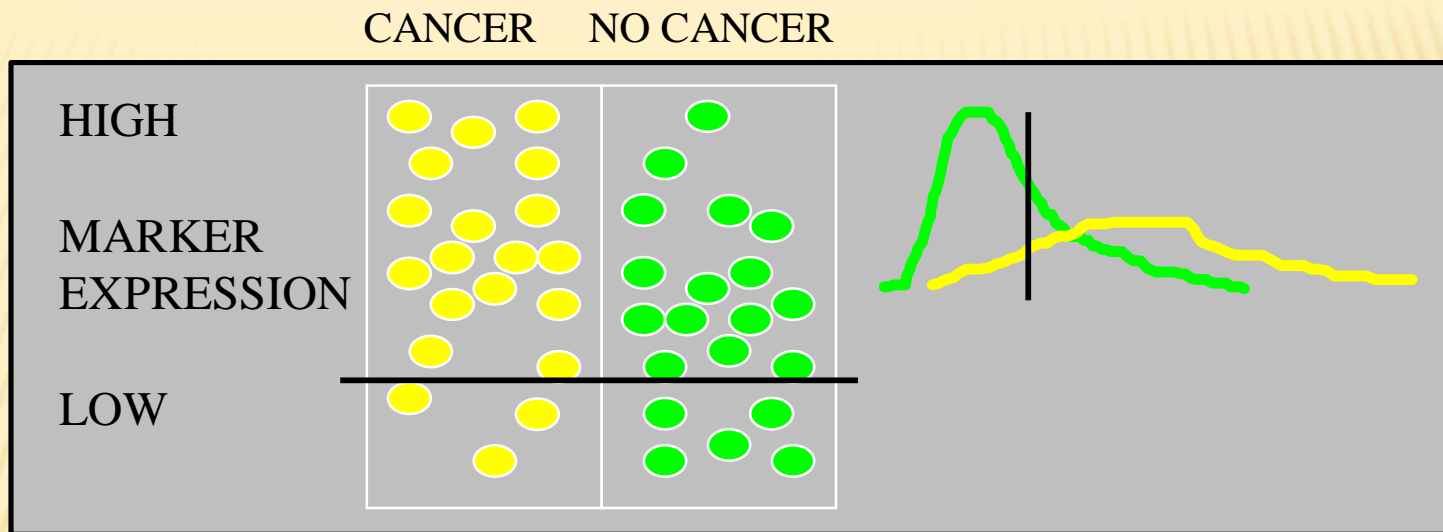


	CANCER	NO CANCER
HIGH	5	2
LOW	15	18
	20	20

$$\text{Sensitivity} = 5/20 = 25\%$$

$$\text{Specificity} = 18/20 = 90\%$$



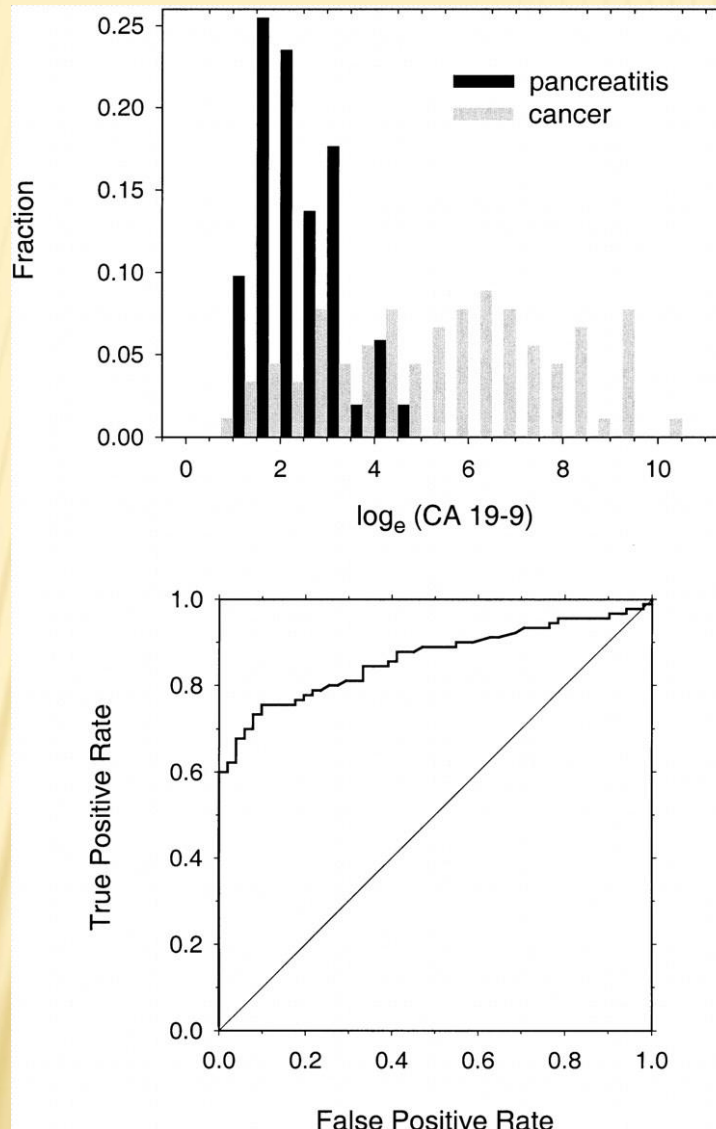


	CANCER	NO CANCER
HIGH	17	15
LOW	3	5
	20	20

$$\text{Sensitivity} = 17/20 = 85\%$$

$$\text{Specificity} = 5/20 = 25\%$$

# Histograms and ROC curve for a cancer antigen, CA 19-9, as a biomarker for pancreatic cancer





# Receiver Operating Characteristic (ROC) Curve

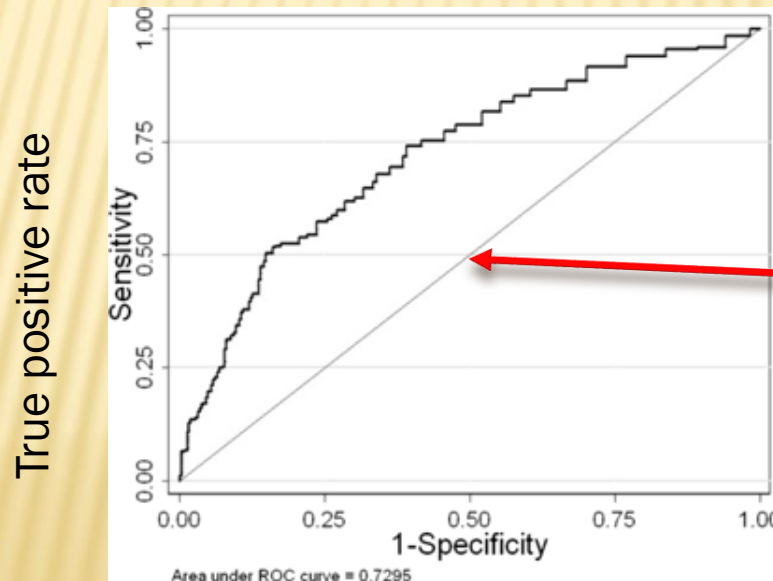
The ROC Curve is a plot of the sensitivity vs. 1-specificity for each potential threshold (cut-off) for a diagnostic test

You need

- a binary outcome (*i.e.*, have disease or not)

- a test with continuous values

- but it is a non-parametric test so data distribution is not important



If your ROC line is lines close to this line, your test is no better than a coin toss at diagnosing disease

False negative rate

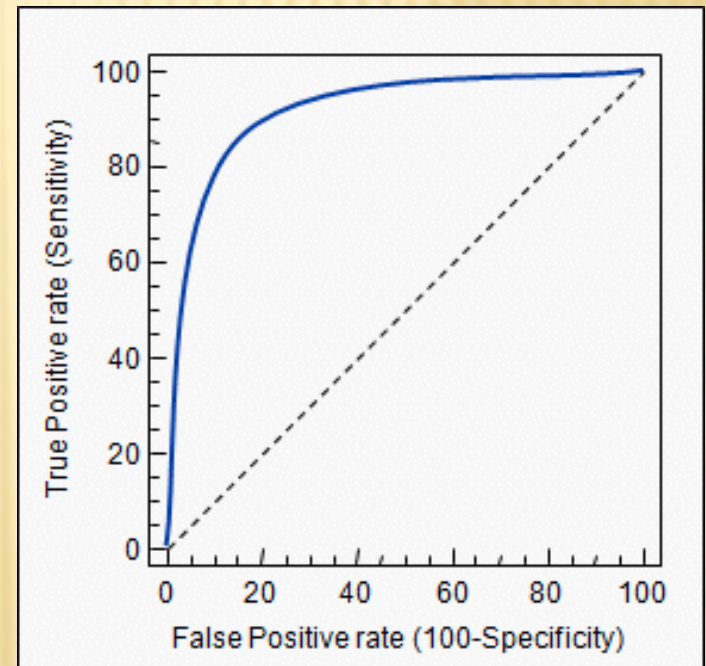
# Receiver Operating Characteristic (ROC) Curve

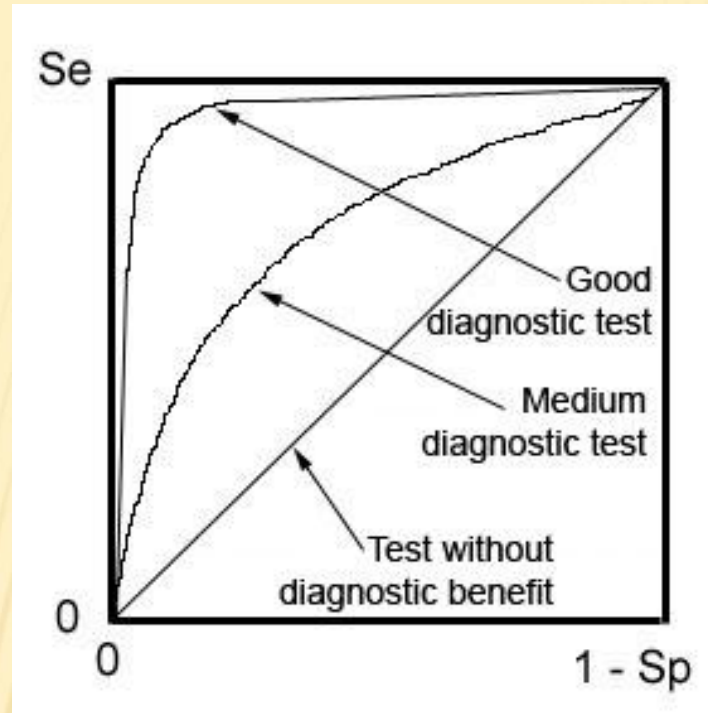
The bottom left of the curve is one silly extreme when the test will never return a diagnosis of disease

0% sensitivity, 100% specificity

At the top right, everyone is diagnosed with the disease

100% sensitivity, 0% specificity



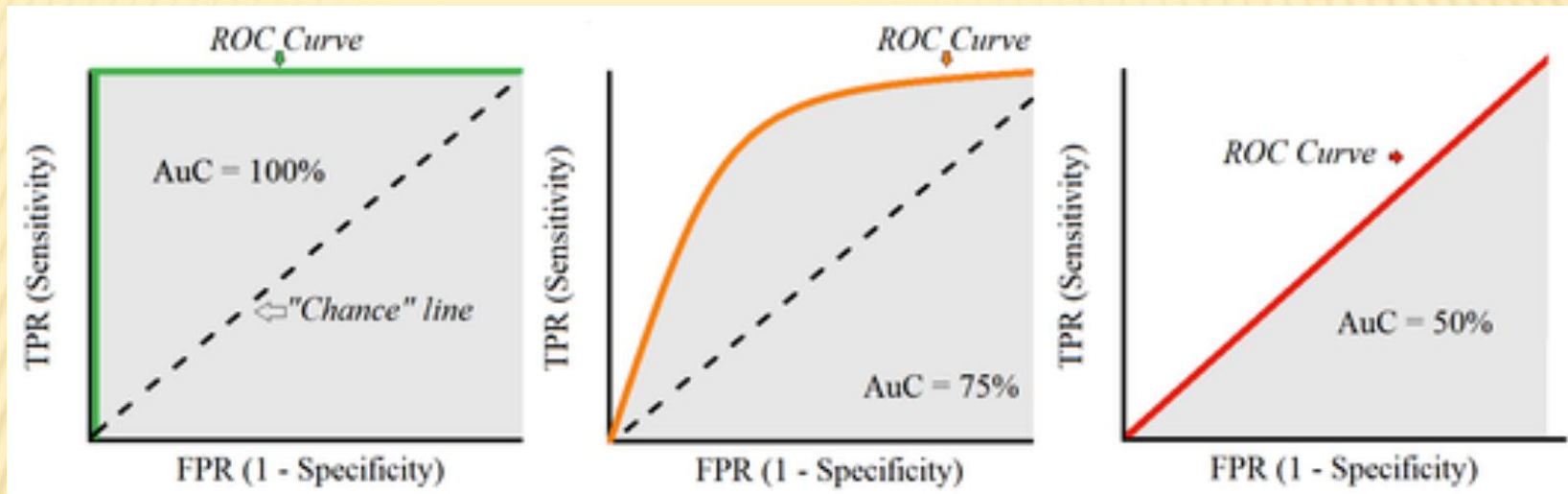


Curves closer to the top left and further from the diagonal line indicate more accurate tests

This is often described numerically with the area under the curve (AUC).



# Area Under the Curve (AUC)



A perfect test will have an AUC of 1.0, while a completely useless test (one whose curve falls on the diagonal line) has an AUC of 0.5.

Excellent test	AUC 0.9-1.0
Good	AUC 0.8-0.9
Fair	AUC 0.7-0.8
Poor	AUC 0.6-0.7
Fail	AUC 0.5-0.6

## The ROC curve has three purposes

Determine the cut-off point at which optimal sensitivity and specificity are achieved

Assess the diagnostic accuracy of a test

The higher the area under the curve (AUC) the better the test

Compare two or more diagnostic tests

The test with the higher AUC is the better test

# Assumptions of an ROC Curve Analysis

The most common ROC analyses are nonparametric.

Nonparametric ROC methods do not require any assumptions about the diagnostic test result distributions and do not provide a smooth ROC curve.

However, parametric methods assume that some function of the diagnostic test measurements are normally distributed in both the diseased and non-diseased populations but with different means.

Null Hypothesis for a single curve: The AUC for the population curve = 0.5

Area under the ROC curve	
Area	0.7183
Std. Error	0.03208
95% confidence interval	0.6555 to 0.7812
P value	<0.0001



# UroScreen Study: Example of Prospective Cohort Design

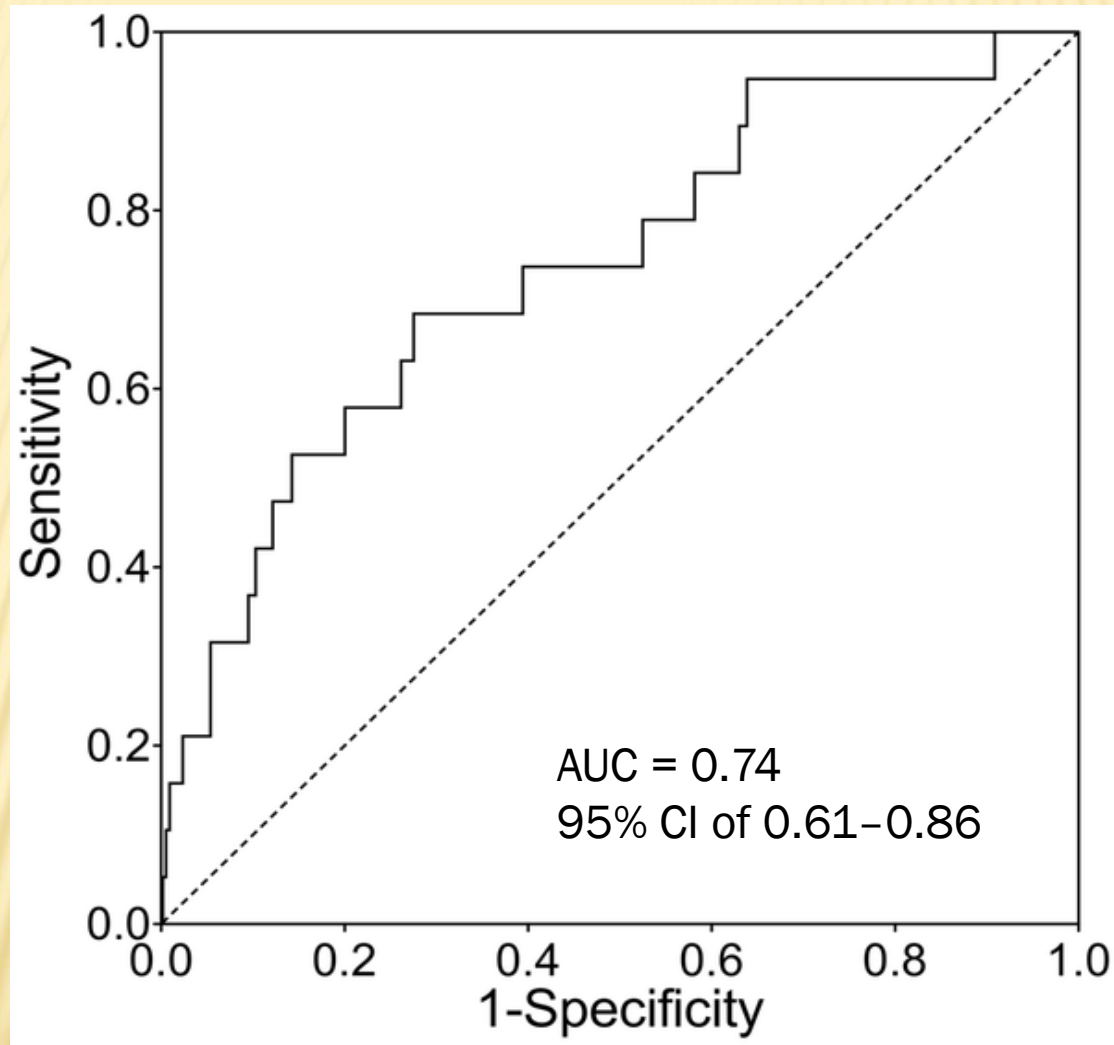
## Performance of survivin mRNA as a biomarker for bladder cancer in the prospective study UroScreen

Survivin was analyzed in 5,716 urine samples from 1,540 chemical workers previously exposed to aromatic amines

Surveillance program with yearly examinations between 2003 and 2010

RNA was extracted from urinary cells and survivin was determined by Real-Time PCR

## ROC curve for survivin



Fair

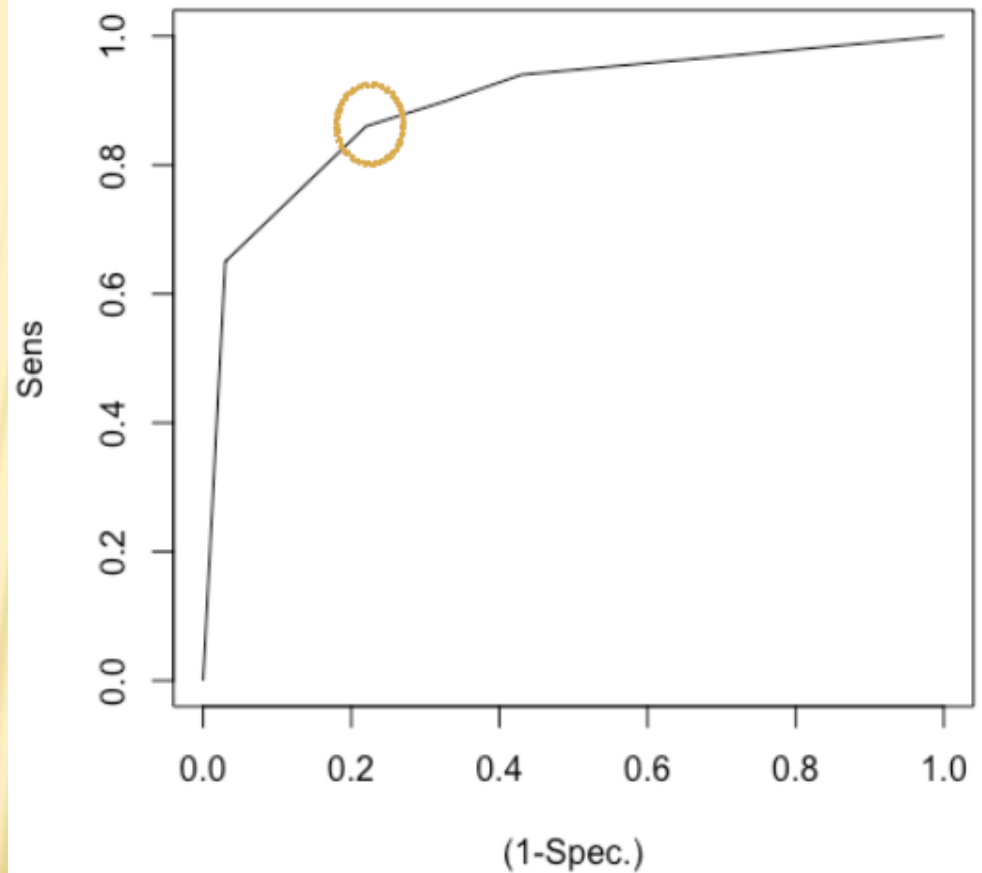
AUC 0.7-0.8

## The ROC: Determining Test Threshold

This corresponds to giving positive test results to those with a score of 4 or higher for this particular test

Sens: 0.86

Spec: 0.78





# ROC Curve for Prostate Specific Antigen and Prostate Cancer

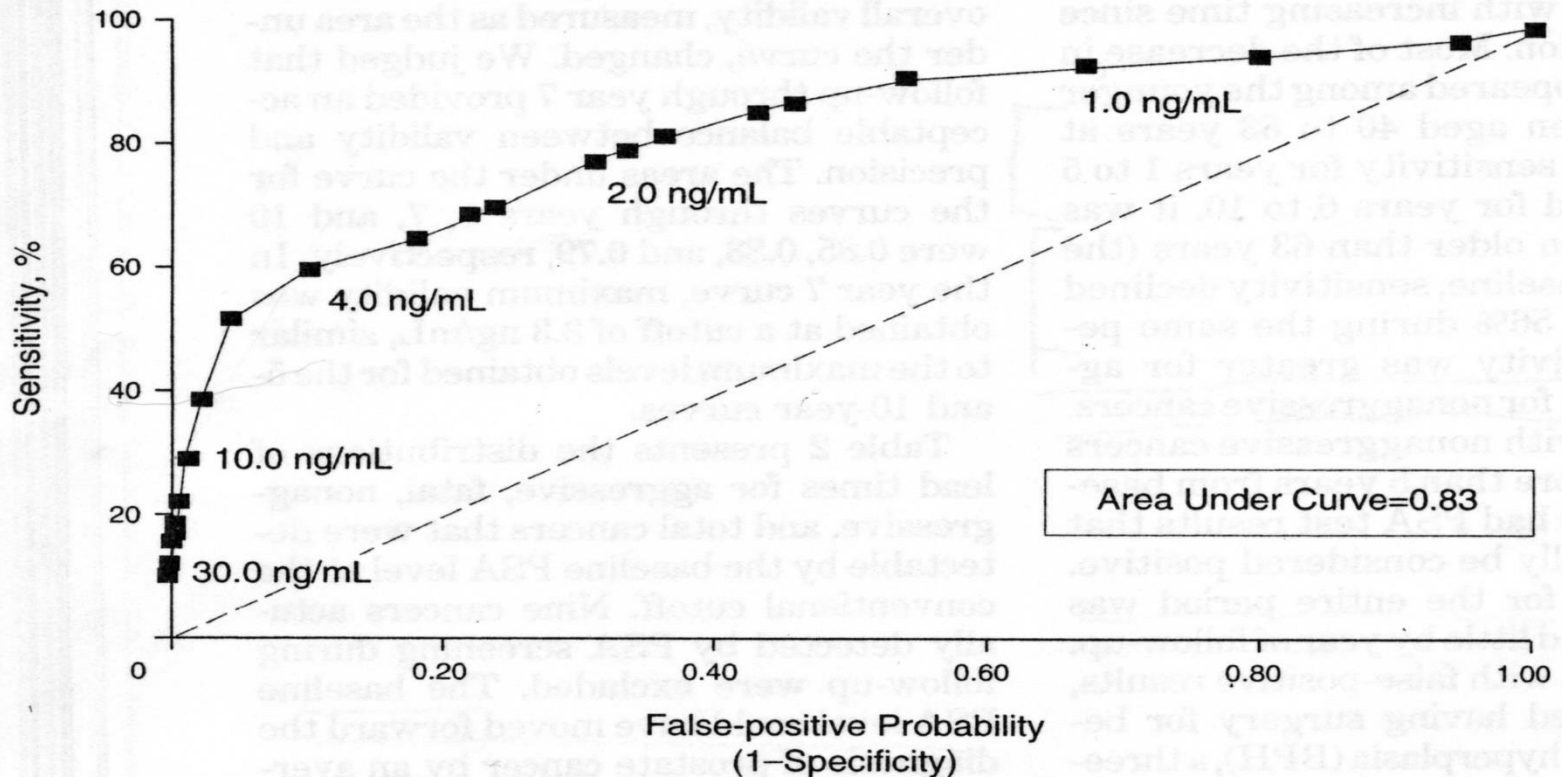


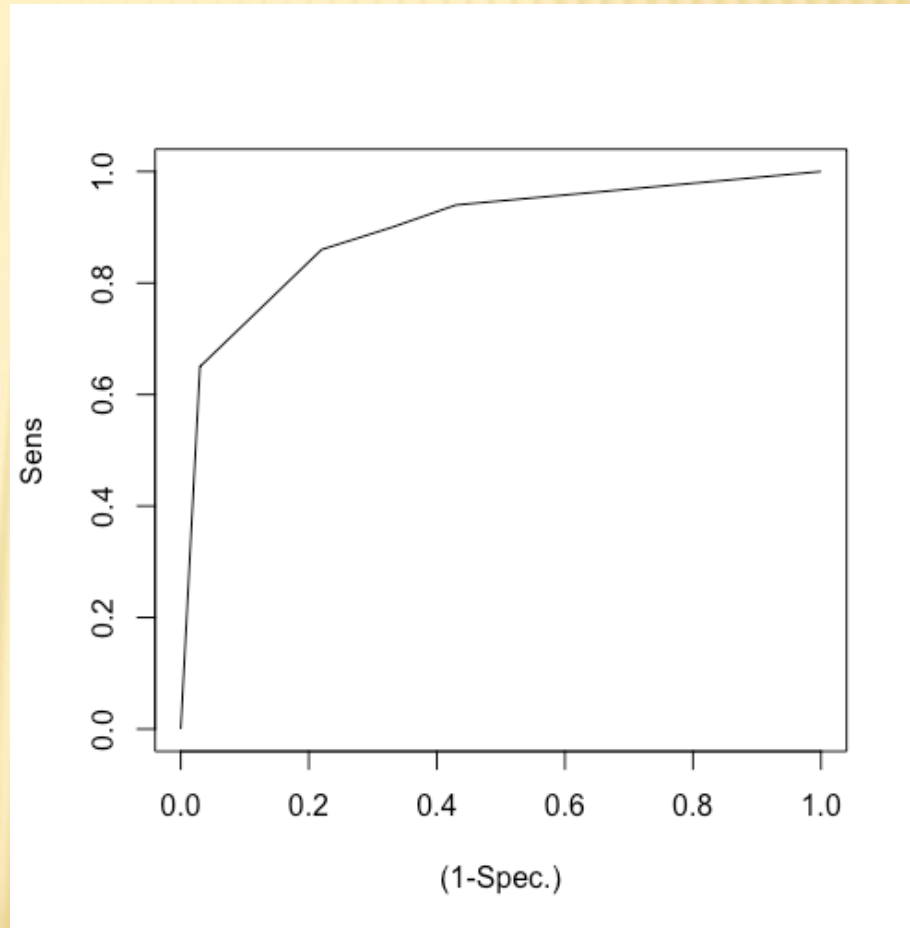
Figure 2.—Sensitivity and specificity for prostate-specific antigen and prostate cancer at various cutoff points, during 7 years of follow-up: Physicians' Health Study (203 cases and 609 controls).

# The ROC: Determining Test Threshold

Where we ultimately decide to put the threshold for a positive test will depend on the costs of false positives vs. false negatives

If we think the costs are similar, then we would like to strike a balance between FP and FN

Choose the threshold closest to upper left corner of the plot



## How can I choose a cut-off?

When cost of a false positive or false negative are similar, two common methods for finding optimal cut-off points are:

1. Find point on curve closest to upper left corner of ROC curve

Smallest value for  $[(100 - \text{sensitivity}\%)^2 + (100 - \text{specificity}\%)^2]$

the cut-off point that best differentiates between people with disease and those without disease

2. Youden Index ( $J$ )

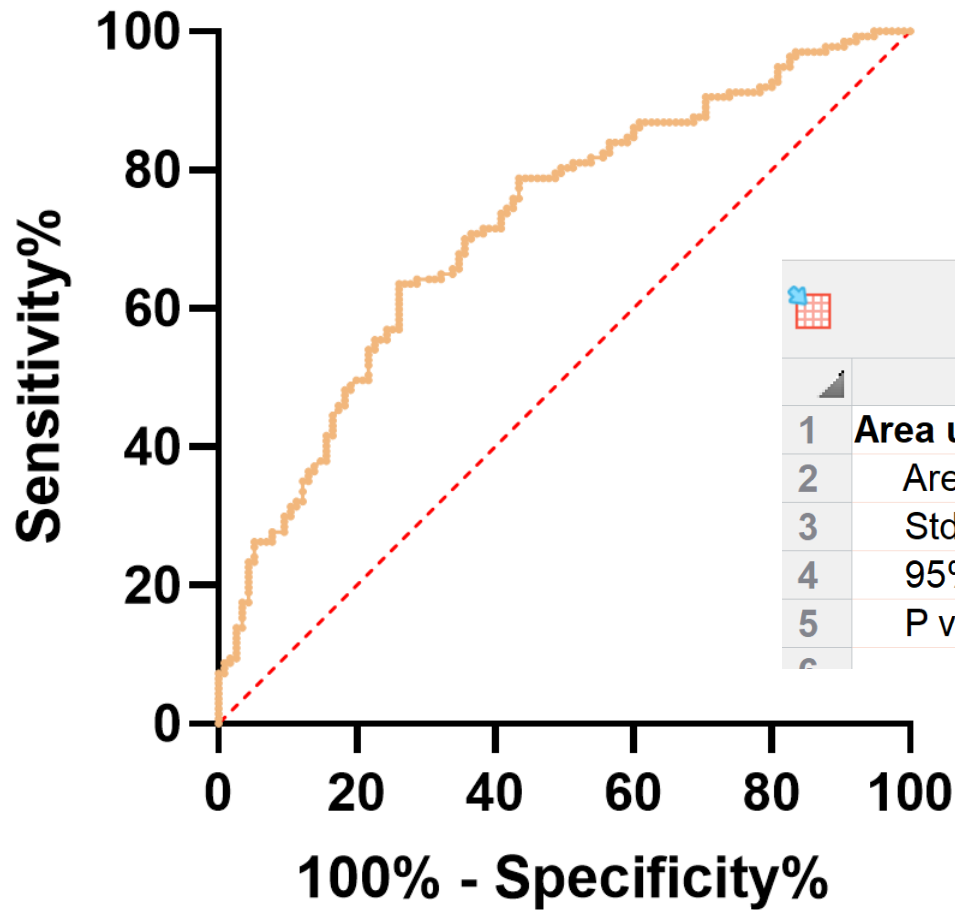
Maximum vertical distance between the ROC curve and the diagonal (random chance) line

$J = \text{Largest value for } (\text{sensitivity}\% + \text{specificity}\% - 100)$

Corresponds to the point on the curve farthest from random chance line



# Example



ROC Area		
1	Area under the ROC curve	
2	Area	0.7183
3	Std. Error	0.03208
4	95% confidence interval	0.6555 to 0.7812
5	P value	<0.0001

		Sensitivity%	95% CI	Specificity%	95% CI
1	< 0.1655	0.7299	0.03744% to 4.019%	100.0	96.77% to 100.0%
2	< 0.1725	1.460	0.2594% to 5.166%	100.0	96.77% to 100.0%
3	< 0.1849	2.190	0.5969% to 6.240%	100.0	96.77% to 100.0%
4	< 0.1966	2.920	1.141% to 7.266%	100.0	96.77% to 100.0%
5	< 0.2028	3.650	1.569% to 8.259%	100.0	96.77% to 100.0%
6	< 0.2075	4.380	2.022% to 9.225%	100.0	96.77% to 100.0%
7	< 0.2114	5.109	2.497% to 10.17%	100.0	96.77% to 100.0%
8	< 0.2143	5.839	2.988% to 11.10%	100.0	96.77% to 100.0%
9	< 0.2164	6.569	3.494% to 12.01%	100.0	96.77% to 100.0%
10	< 0.2178	7.299	4.013% to 12.92%	100.0	96.77% to 100.0%
11	< 0.2193	7.299	4.013% to 12.92%	99.13	95.24% to 99.96%
12	< 0.2227	8.029	4.542% to 13.81%	99.13	95.24% to 99.96%
13	< 0.2260	8.759	5.082% to 14.69%	99.13	95.24% to 99.96%
14	< 0.2269	8.759	5.082% to 14.69%	98.26	93.88% to 99.69%
15	< 0.2275	9.489	5.629% to 15.56%	98.26	93.88% to 99.69%
16	< 0.2293	9.489	5.629% to 15.56%	97.39	92.61% to 99.29%

Smallest value for  $[(100 - \text{sensitivity}\%)^2 + (100 - \text{specificity}\%)^2]$

	A	B	C	D	E	F
1	Cut-off	Sensitivity	Specificity	(100-sens)2	(100-spec)2	sum
2	< 0.1655	0.7299	100	9854.552754	0	9854.55
3	< 0.1725	1.46	100	9710.1316	0	9710.13
4	< 0.1849	2.19	100	9566.7961	0	9566.80
5	< 0.1966	2.92	100	9424.5264	0	9424.53
102	< 0.2914	59.12	73.91	1671.1744	680.6881	2351.86
103	< 0.2923	59.85	73.91	1612.0225	680.6881	2292.71
104	< 0.2936	60.58	73.91	1553.9364	680.6881	2234.62
105	< 0.2941	61.31	73.91	1496.9161	680.6881	2177.60
106	< 0.2945	62.04	73.91	1440.9616	680.6881	2121.65
107	< 0.2949	62.77	73.91	1386.0729	680.6881	2066.76
108	< 0.2952	63.5	73.91	1332.25	680.6881	2012.94
109	< 0.2965	63.5	73.04	1332.25	726.8416	2059.09
110	< 0.2978	63.5	72.17	1332.25	774.5089	2106.76
111	< 0.2982	63.5	71.3	1332.25	823.69	2155.94
112	< 0.2986	64.23	71.3	1279.4929	823.69	2103.18
113	< 0.2987	64.23	69.57	1279.4929	925.9849	2205.48
114	< 0.2990	64.23	68.7	1279.4929	979.69	2259.18
115	< 0.2996	64.23	67.83	1279.4929	1034.9089	2314.40
116	< 0.3002	64.96	67.83	1227.8016	1034.9089	2262.71
117	< 0.3005	64.96	66.09	1227.8016	1149.8881	2377.69



## Youden Index ( $J$ )

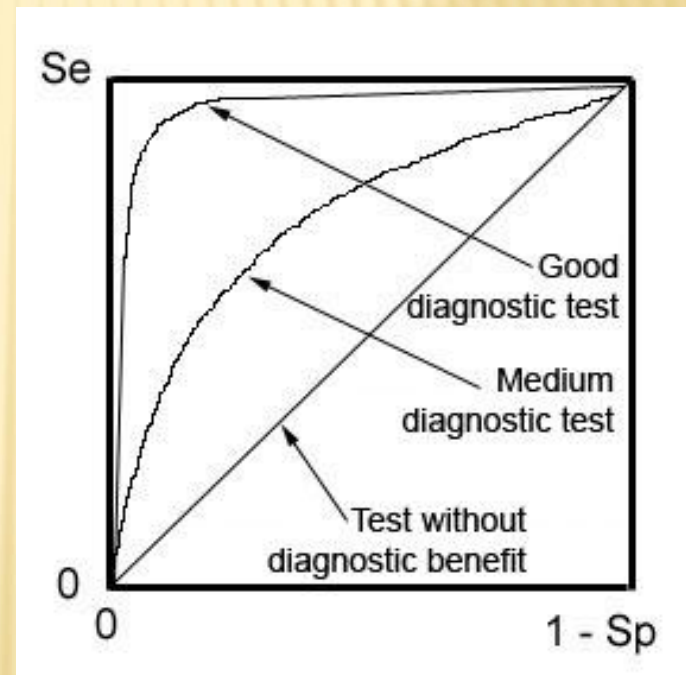
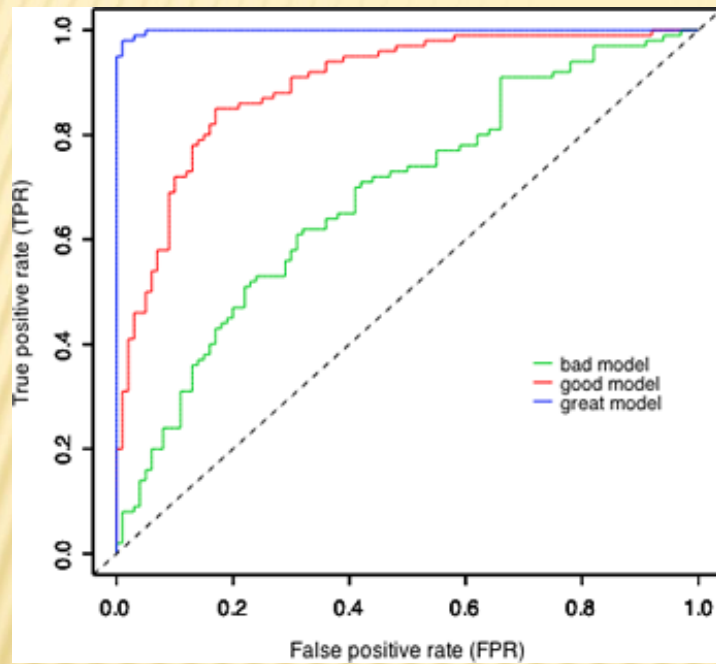
$J$  = Largest value for (sensitivity% + specificity% – 100)

	A	B	C	D
1	Cut-off	Sensitivity	Specificity	Youden
2	< 0.1655	0.7299	100	0.7299
3	< 0.1725	1.46	100	1.46
4	< 0.1849	2.19	100	2.19
5	< 0.1966	2.92	100	2.92

103	< 0.2923	59.85	73.91	33.76
104	< 0.2936	60.58	73.91	34.49
105	< 0.2941	61.31	73.91	35.22
106	< 0.2945	62.04	73.91	35.95
107	< 0.2949	62.77	73.91	36.68
108	< 0.2952	63.5	73.91	37.41
109	< 0.2965	63.5	73.04	36.54
110	< 0.2978	63.5	72.17	35.67
111	< 0.2982	63.5	71.3	34.8
112	< 0.2986	64.23	71.3	35.53
113	< 0.2987	64.23	69.57	33.8
114	< 0.2990	64.23	68.7	32.93

The same cut-off as the other method

ROC curves can be used to compare tests



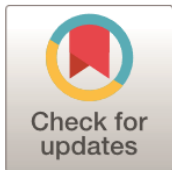
The larger the AUC the better the test

RESEARCH ARTICLE

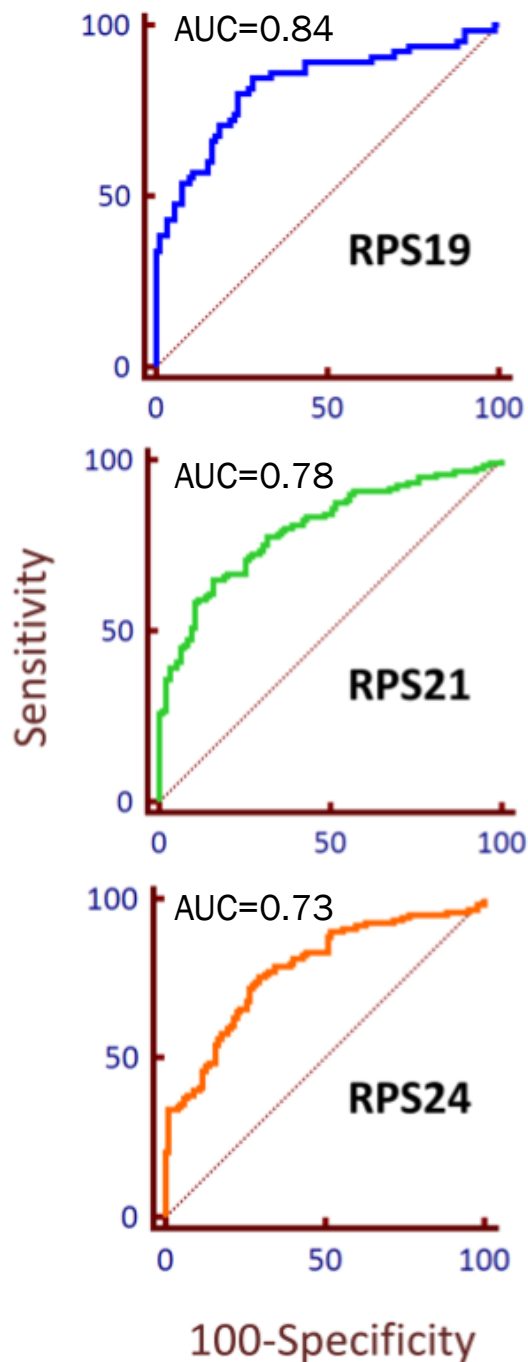
# Expression of ribosomal proteins in normal and cancerous human prostate tissue

**Callum Arthurs<sup>1</sup>, Bibi Nazia Murtaza<sup>1,2\*</sup>, Calum Thomson<sup>3</sup>, Kerry Dickens<sup>1</sup>, Rui Henrique<sup>4,5</sup>, Hitendra R. H. Patel<sup>6,7</sup>, Mariana Beltran<sup>8</sup>, Michael Millar<sup>9</sup>, Christopher Thrasivoulou<sup>10</sup>, Aamir Ahmed<sup>1,2\*</sup>**

**1** Prostate Cancer Research Centre at the Centre for Stem Cells and Regenerative Medicine, King's College London, London, United Kingdom, **2** Division of Surgery, University College London, London, United Kingdom, **3** Dundee Imaging Facility, College of Life Sciences, University of Dundee, Dundee, United Kingdom, **4** Department of Pathology, Portuguese Oncology Institute, Porto, Portugal, **5** Department of Pathology and Molecular Immunology, Abel Salazar Institute of Biomedical Sciences, University of Porto, Porto, Portugal, **6** Division of Surgery, Oncology, Urology and Women's Health, University Hospital of Northern Norway, Tromsø, Norway, **7** Department of Urology, Princess Alexandra Hospital NHS Trust, Harlow, Essex, United Kingdom, **8** Aquila Biomedical, Edinburgh, United Kingdom, **9** Queen's Medical Research Institute, University of Edinburgh, Edinburgh, United Kingdom, **10** Research Department of Cell and Developmental Biology, The Centre for Cell and Molecular Dynamics, Rockefeller Building, University College London, London, United Kingdom







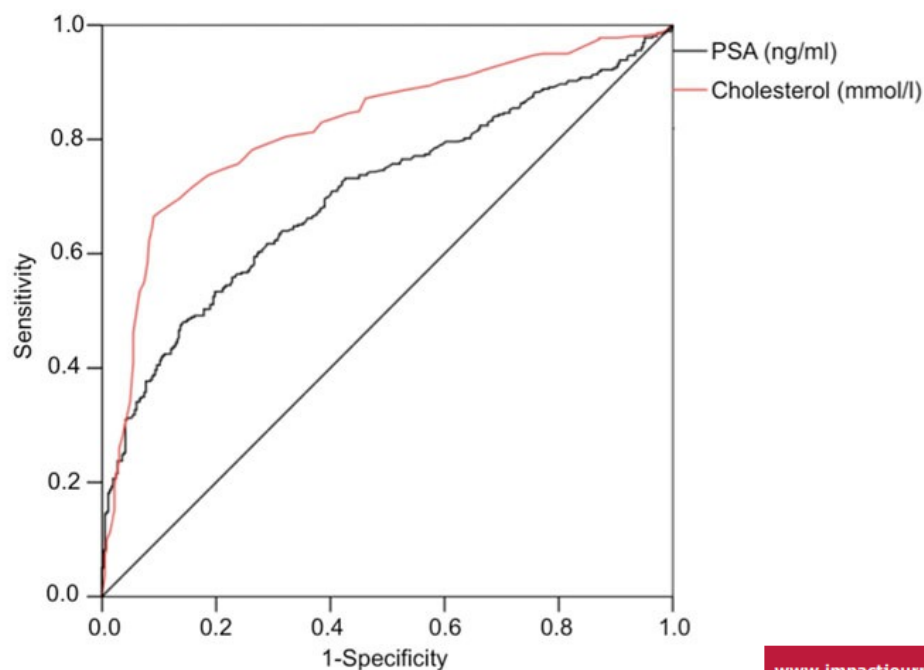
**Fig 3. ROC of RPS19, RPS21, and RPS24 in prostate tissue.**

ROC curve to demonstrate the potential ability of a label to differentiate between malignant and non-malignant prostate tissue. The probit function was used for area fraction per amount of tissue for ROC analysis of RPS19, RPS21, and RPS24. The dotted line is representative of an area under the curve of 0.5 which shows no distinction between the two classifiers (e.g. normal vs malignant).

*Which marker is better at diagnosing malignant from non-malignant tissue?*

## ROC curve for diagnosing high-risk prostate cancer

Preoperative total serum cholesterol values  
PSA (prostate specific antigen)



Cholesterol  
AUC=0.82, 95% CI 0.79-0.85,  $p<0.001$

PSA  
AUC=0.71, 95% CI 0.67-0.74,  $p<0.001$

### **Influence of serum cholesterol level and statin treatment on prostate cancer aggressiveness**

**Thomas J. Schnoeller<sup>1</sup>, Florian Jentzmik<sup>1,2</sup>, Andres J. Schrader<sup>3</sup> and Julie Steinestel<sup>3</sup>**

<sup>1</sup>Department of Urology, Ulm University Medical Center, Ulm, Germany

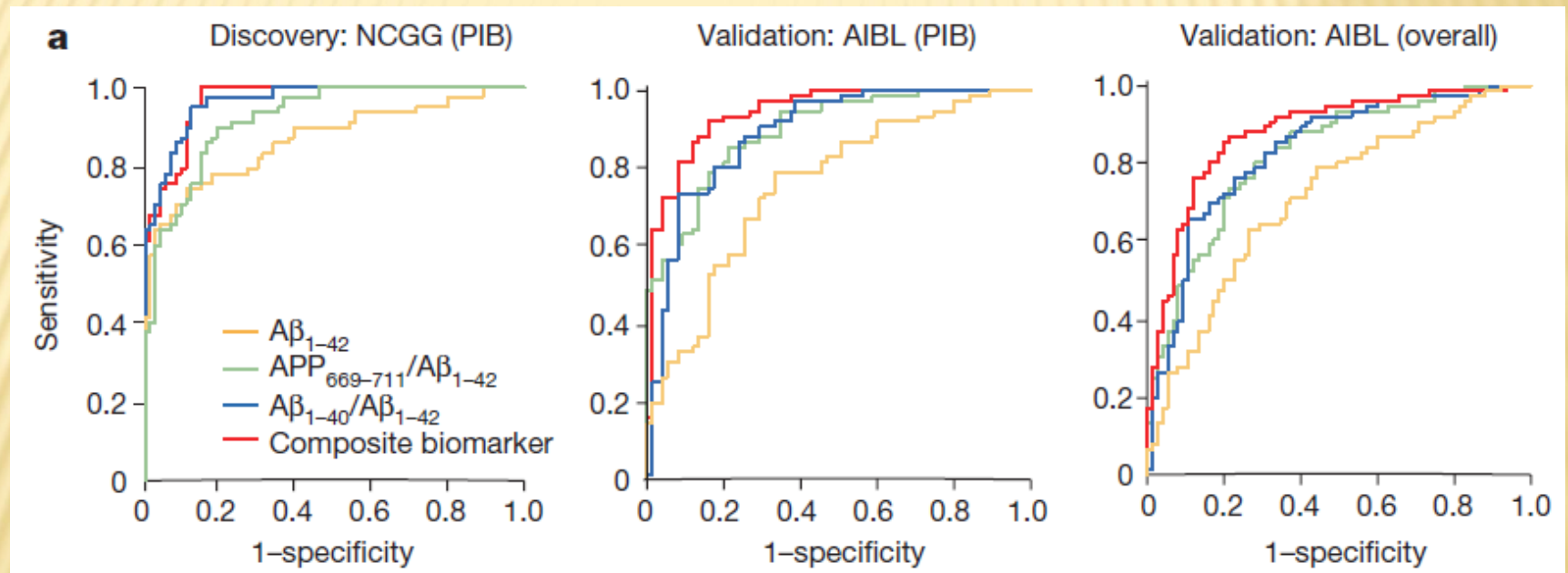
<sup>2</sup>Department of Urology, St. Elisabeth Hospital, Ravensburg, Germany

<sup>3</sup>Department of Urology, Muenster University Medical Center, Muenster, Germany

# High performance plasma amyloid- $\beta$ biomarkers for Alzheimer's disease

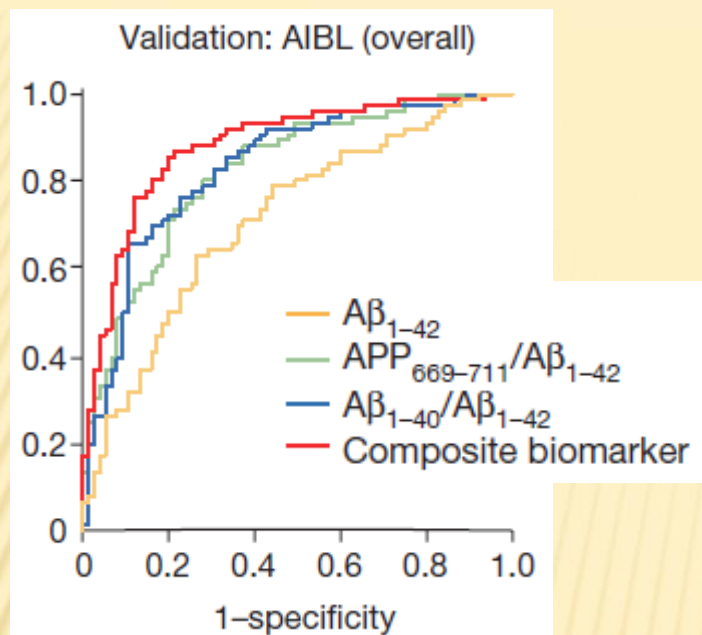
Akinori Nakamura<sup>1</sup>, Naoki Kaneko<sup>2</sup>, Victor L. Villemagne<sup>3,4</sup>, Takashi Kato<sup>1,5</sup>, James Doecke<sup>6</sup>, Vincent Doré<sup>3,6</sup>, Chris Fowler<sup>4</sup>, Qiao-Xin Li<sup>4</sup>, Ralph Martins<sup>7</sup>, Christopher Rowe<sup>3,4</sup>, Taisuke Tomita<sup>8</sup>, Katsumi Matsuzaki<sup>9</sup>, Kenji Ishii<sup>10</sup>, Kazunari Ishii<sup>11</sup>, Yutaka Arahata<sup>5</sup>, Shinichi Iwamoto<sup>2</sup>, Kengo Ito<sup>1,5</sup>, Koichi Tanaka<sup>2</sup>, Colin L. Masters<sup>4</sup> & Katsuhiko Yanagisawa<sup>1</sup>

8 FEBRUARY 2018 | VOL 554 | NATURE | 249



**Figure 2 | High performance of the plasma biomarkers. a**, ROC analyses for each biomarker when predicting individual A $\beta$  +/A $\beta$  - status for the discovery and validation data sets. Unadjusted analyses of the NCGG PIB discovery data (left), the AIBL PIB (middle) and AIBL overall (all tracers, right) validation data. See Extended Data Table 1a for detailed performance values. Data are from 121, 111 and 252 individuals for the NCGG PIB, AIBL and AIBL overall data, respectively.





## Use a column table

Search...

Data with Results

Fig2a AIBL overall Ab142

ROC

Fig2a AIBL overall APP669/Ab142

ROC

Fig2a AIBL overall AB140/Ab142

ROC

Fig2a AIBL overall Composite

ROC

Fig2a AIBL overall Grouped

ROC

New Data Table...

Info

Project info 1

New Info...

Graphs

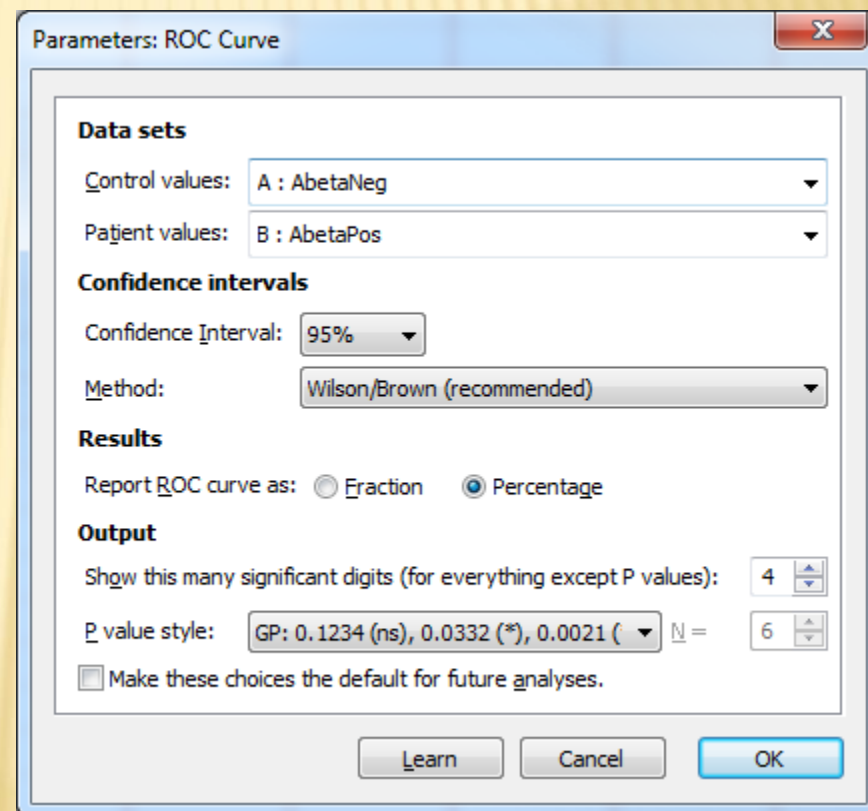
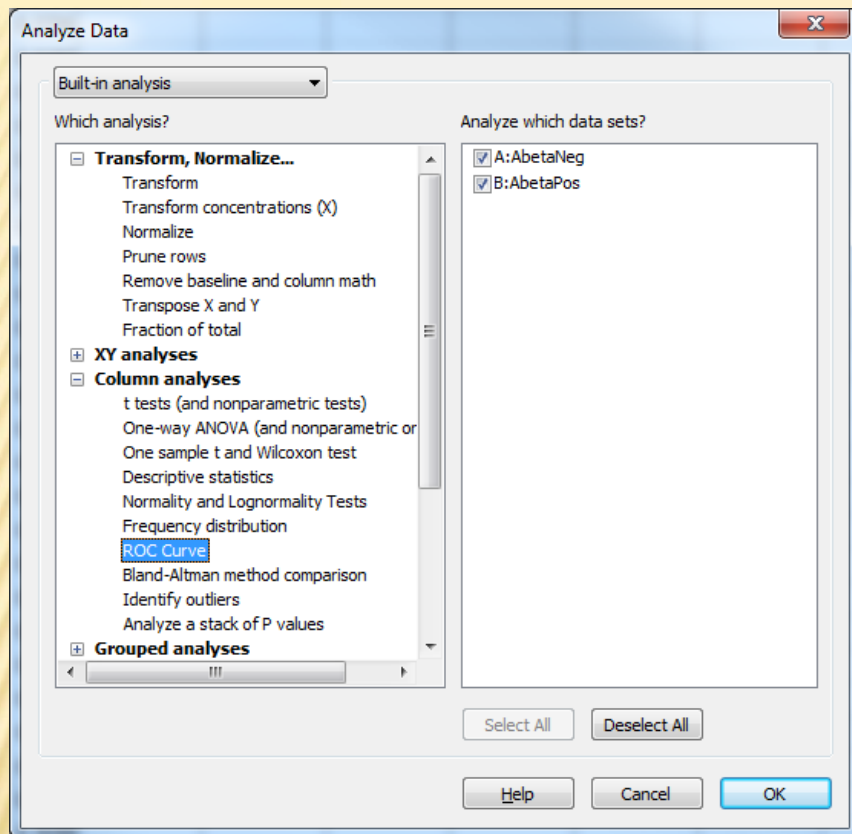
Fig2a AIBL overall Ab142

ROC curve: ROC of Fig2a AIBL overall A

Fig2a AIBL overall APP669/Ab142

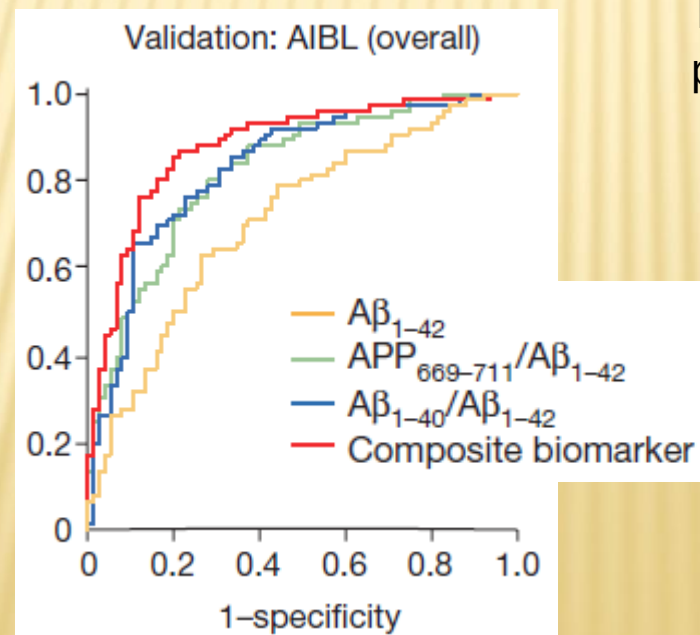
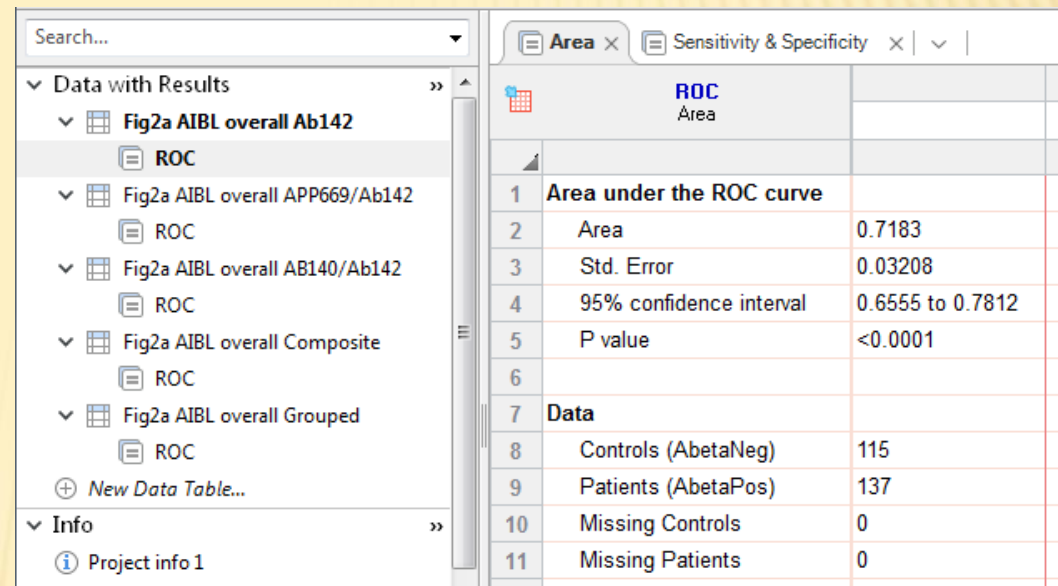
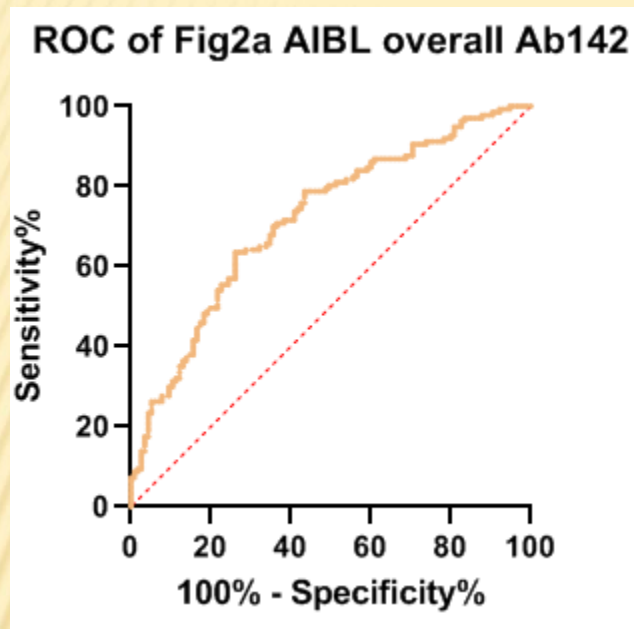
ROC curve: ROC of Fig2a AIBL overall AP

	Group A	Group B
	AbetaNeg	AbetaPos
1	0.4448	0.2796
2	0.3313	0.3482
3	0.6844	0.3242
4	0.4370	0.3200
5	0.3325	0.3395
6	0.2868	0.2409
7	0.3562	0.2098
8	0.2412	0.2440
9	0.5087	0.2473
10	0.4255	0.2554
11	0.2806	0.2171
12	0.3520	0.2755
13	0.2383	0.2709
14	0.5182	0.2410
15	0.2565	0.3038
16	0.3351	0.2730
17	0.4809	0.2814
18	0.3873	0.3135
19	0.3784	0.2724







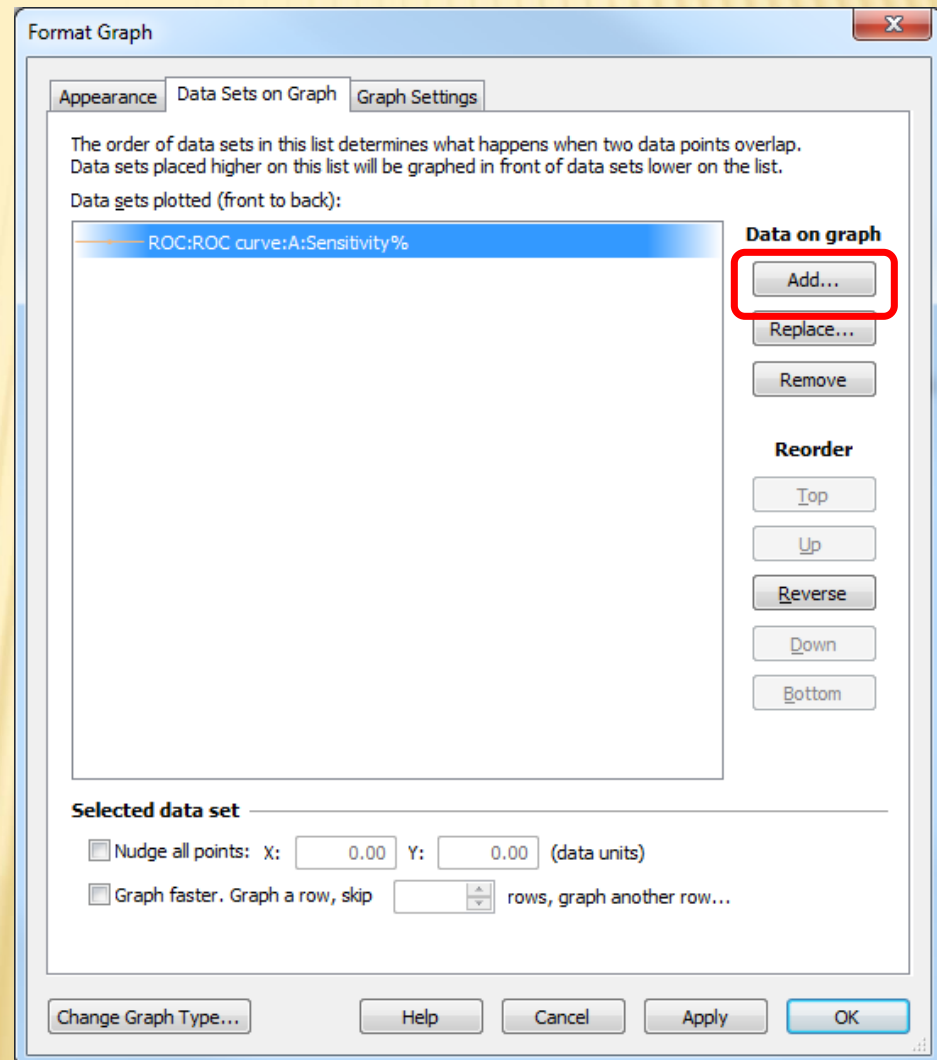
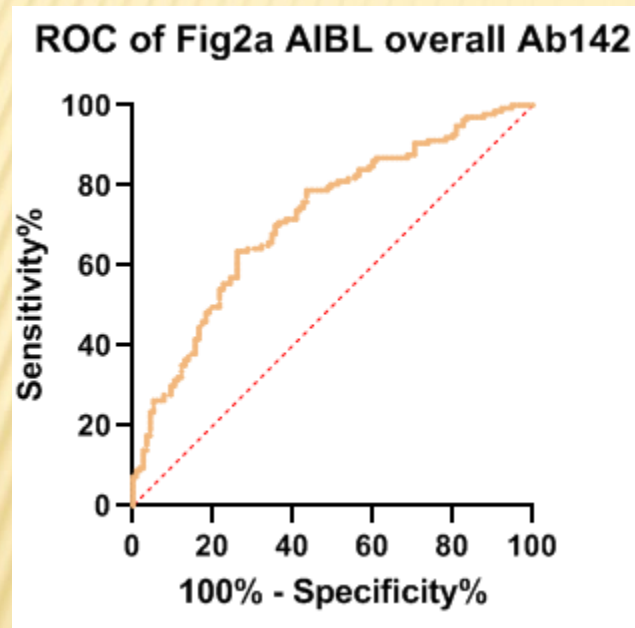


The null hypothesis is that the population curve AUC = 0.50

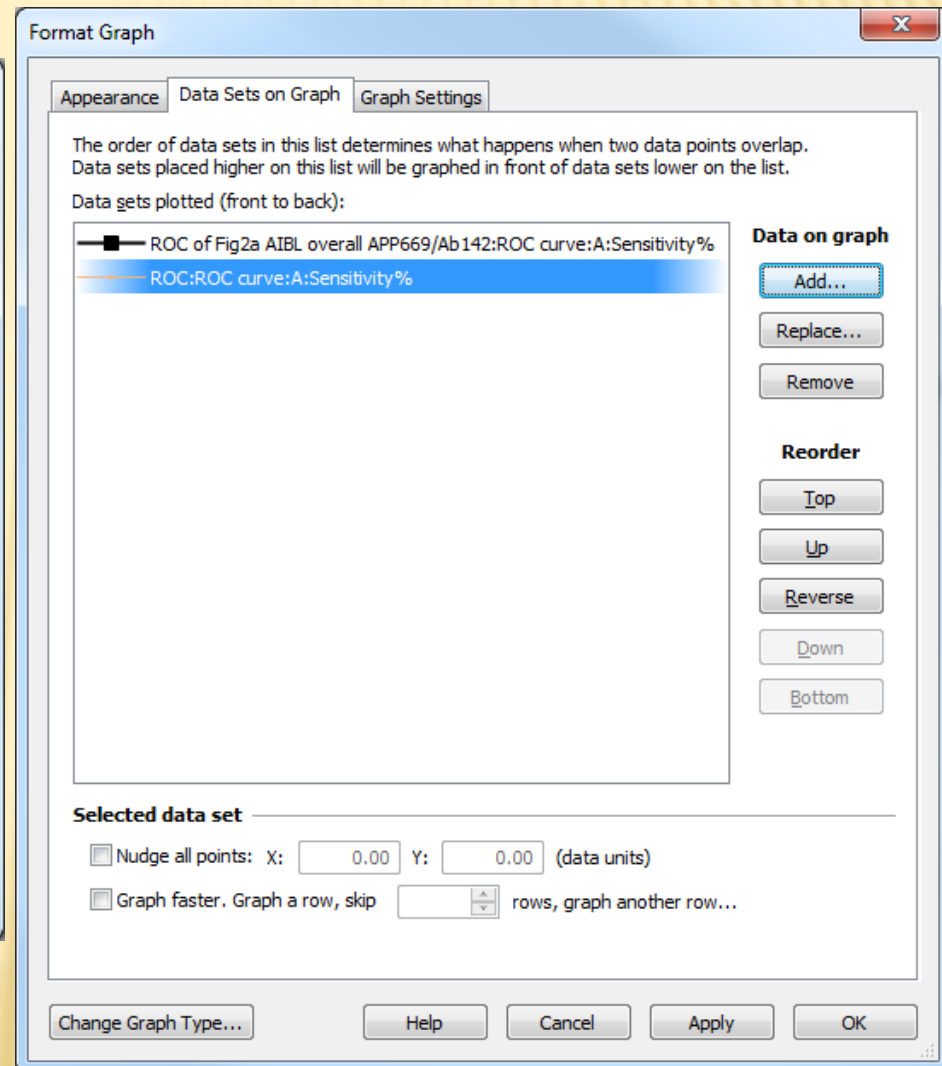
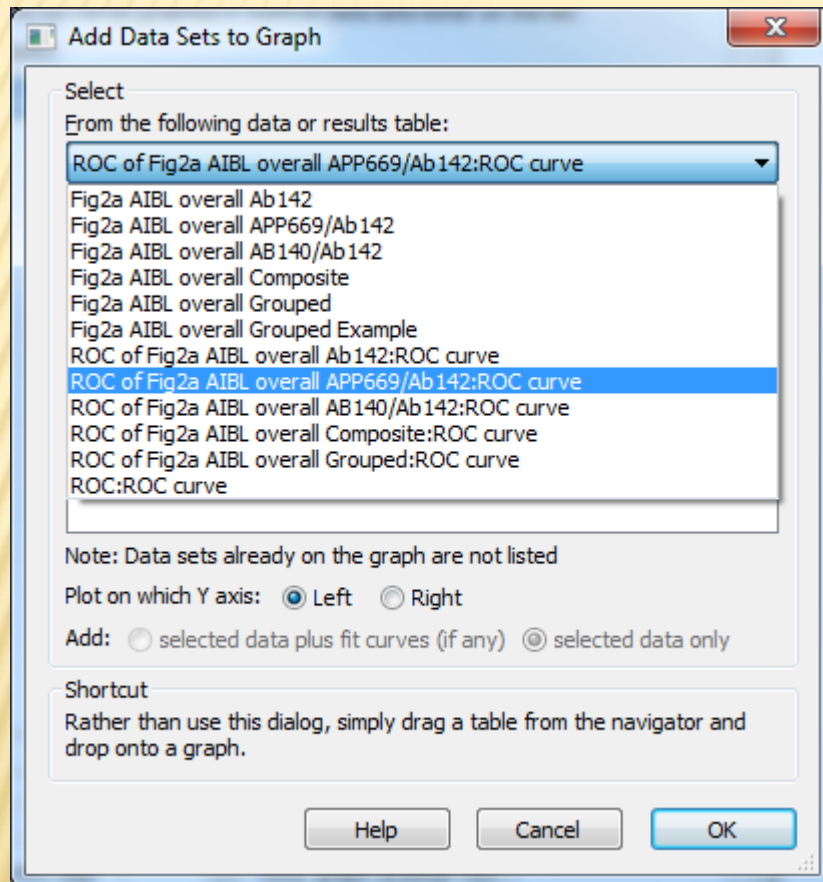
# Adding Other ROC Curves to the Graph

Do this after you have created separate graphs for each ROC curve

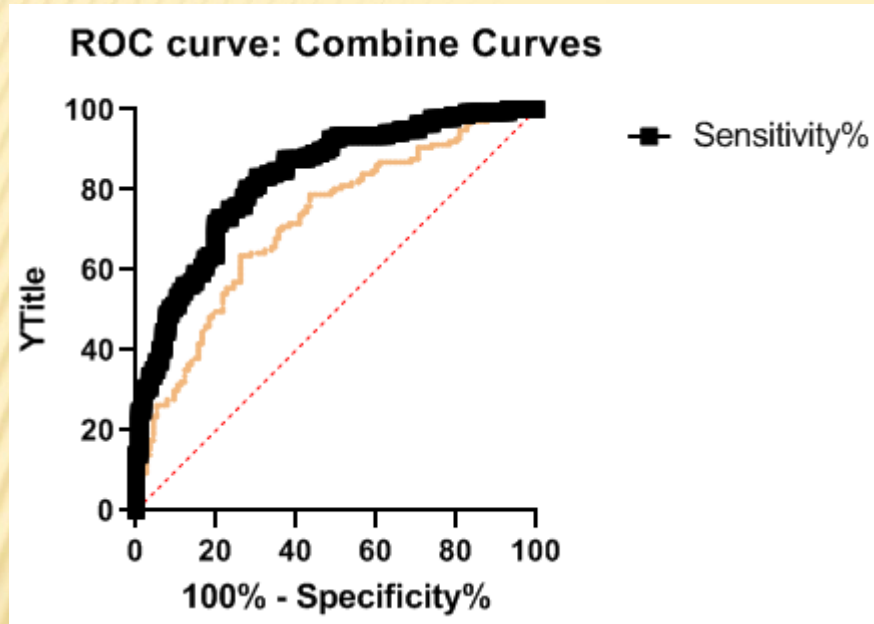
Right click on the graph



From the drop-down menu choose the ROC for the next curve



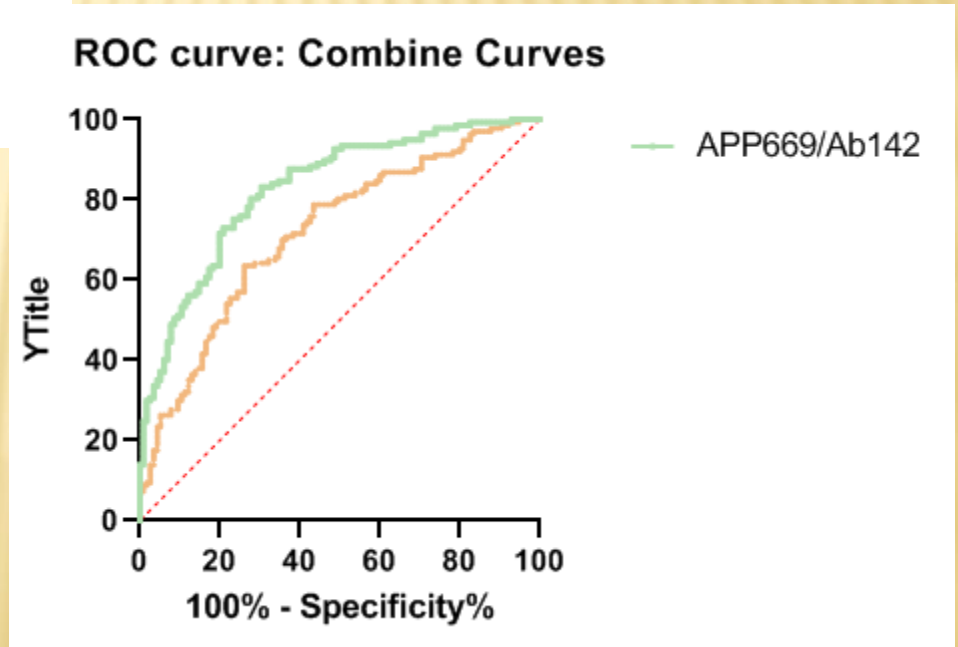


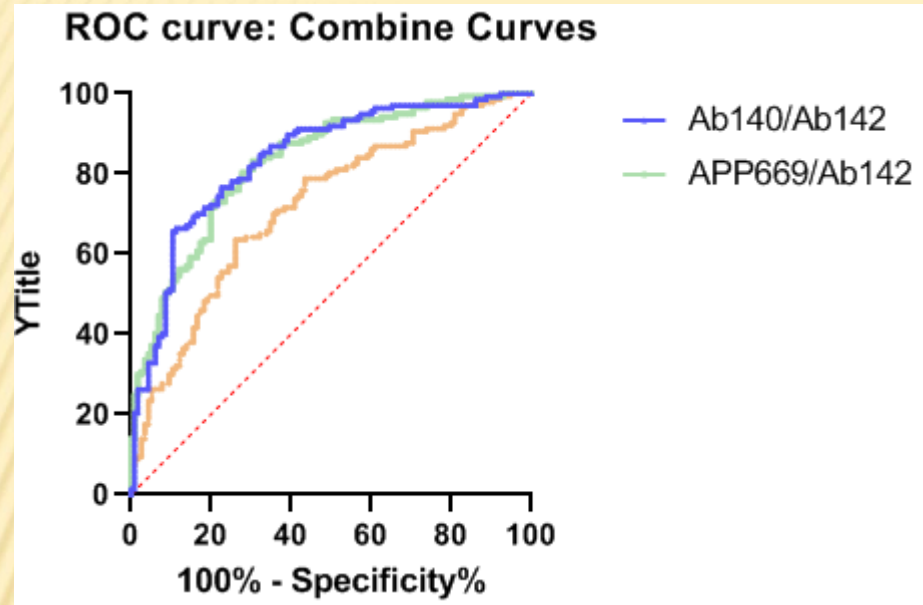


Double-click on the black line and change to the color you want

I changed the legend manually

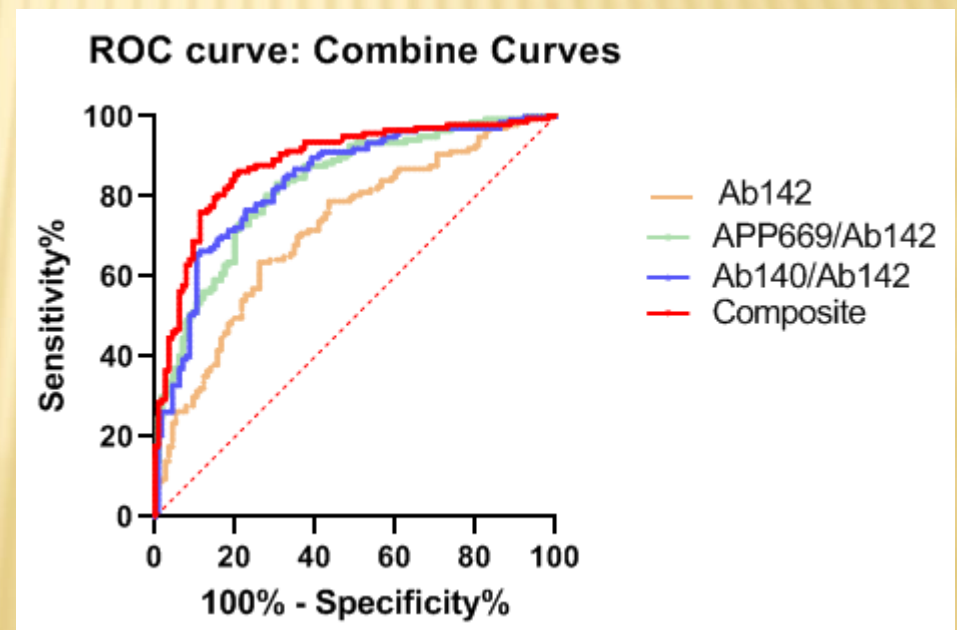
Double-click on the graph to add another ROC curve



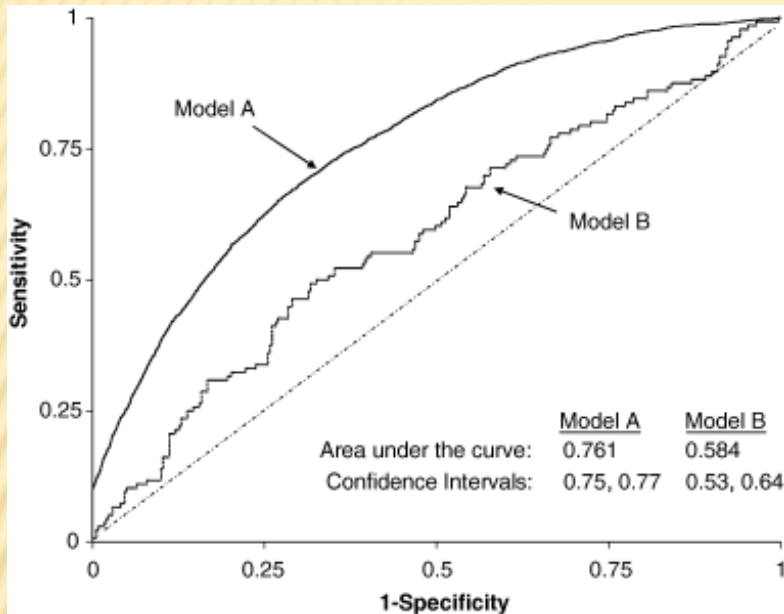


And so on until you get all the curves you want on the graph

I manipulated the graph to create the legend. I had to add the Ab142 part of the legend.



# Area Under the Curve: Compare Curves



$$Z = \frac{|Area_1 - Area_2|}{\sqrt{SE_{Area1}^2 + SE_{Area2}^2}}$$

With many repeated tests, the distribution of  $z$  is centered at zero with a standard deviation of 1.0. To calculate a two-tail  $p$ -value, therefore, use the Microsoft Excel function:

$$p = 2 * (1 - \text{NORMSDIST}(z))$$

This method is appropriate when the two ROC curves are from different subjects. A different method is needed to compare ROC curves when both laboratory tests were evaluated in the same group of patients and controls.



# A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases<sup>1</sup>

Radiology 148: 839-843, September 1983

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$$

r = the correlation between the 2 areas

TABLE I: Correlation Coefficients\*

Average Correlation between Ratings <sup>†</sup>	Average Area <sup>‡</sup>											
	.700	.725	.750	.775	.800	.825	.850	.875	.900	.925	.950	.975
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
0.04	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02
0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.02
0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.04	0.03
0.10	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.06	0.06	0.04
0.12	0.11	0.11	0.11	0.10	0.10	0.10	0.09	0.09	0.08	0.08	0.07	0.05
0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.10	0.09	0.08	0.06
0.16	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.12	0.11	0.11	0.09	0.07
0.18	0.16	0.16	0.16	0.16	0.15	0.15	0.14	0.14	0.13	0.12	0.11	0.09
0.20	0.18	0.18	0.18	0.17	0.17	0.17	0.16	0.15	0.15	0.14	0.12	0.10
0.22	0.20	0.20	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.15	0.14	0.11
0.24	0.22	0.22	0.21	0.21	0.21	0.20	0.19	0.19	0.18	0.17	0.15	0.12
0.26	0.24	0.23	0.23	0.23	0.22	0.22	0.21	0.20	0.19	0.18	0.16	0.13
0.28	0.26	0.25	0.25	0.25	0.24	0.24	0.23	0.22	0.21	0.20	0.18	0.15
0.30	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.24	0.23	0.21	0.19	0.16
0.32	0.29	0.29	0.29	0.28	0.28	0.27	0.26	0.26	0.24	0.23	0.21	0.18
0.34	0.31	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.25	0.23	0.19
0.36	0.33	0.33	0.32	0.32	0.31	0.31	0.30	0.29	0.28	0.26	0.24	0.21
0.38	0.35	0.35	0.34	0.34	0.33	0.33	0.32	0.31	0.30	0.28	0.26	0.22
0.40	0.37	0.37	0.36	0.36	0.35	0.35	0.34	0.33	0.32	0.30	0.28	0.24
0.42	0.39	0.39	0.38	0.38	0.37	0.36	0.36	0.35	0.33	0.32	0.29	0.25
0.44	0.41	0.40	0.40	0.40	0.39	0.38	0.38	0.37	0.35	0.34	0.31	0.27
0.46	0.43	0.42	0.42	0.42	0.41	0.40	0.39	0.38	0.37	0.35	0.33	0.29
0.48	0.45	0.44	0.44	0.43	0.43	0.42	0.41	0.40	0.39	0.37	0.35	0.30
0.50	0.47	0.46	0.46	0.45	0.45	0.44	0.43	0.42	0.41	0.39	0.37	0.32
0.52	0.49	0.48	0.48	0.47	0.47	0.46	0.45	0.44	0.43	0.41	0.39	0.34
0.54	0.51	0.50	0.50	0.49	0.49	0.48	0.47	0.46	0.45	0.43	0.41	0.36
0.56	0.53	0.52	0.52	0.51	0.51	0.50	0.49	0.48	0.47	0.45	0.43	0.38
0.58	0.55	0.54	0.54	0.53	0.53	0.52	0.51	0.50	0.49	0.47	0.45	0.40
0.60	0.57	0.56	0.56	0.55	0.55	0.54	0.53	0.52	0.51	0.49	0.47	0.42
0.62	0.59	0.58	0.58	0.57	0.57	0.56	0.55	0.54	0.53	0.51	0.49	0.45
0.64	0.61	0.60	0.60	0.59	0.59	0.58	0.58	0.57	0.55	0.54	0.51	0.47
0.66	0.63	0.62	0.62	0.62	0.61	0.60	0.60	0.59	0.57	0.56	0.53	0.49
0.68	0.65	0.64	0.64	0.64	0.63	0.62	0.62	0.61	0.60	0.58	0.56	0.51
0.70	0.67	0.66	0.66	0.66	0.65	0.65	0.64	0.63	0.62	0.60	0.58	0.54
0.72	0.69	0.69	0.68	0.68	0.67	0.67	0.66	0.65	0.64	0.63	0.60	0.56
0.74	0.71	0.71	0.70	0.70	0.69	0.69	0.68	0.67	0.66	0.65	0.63	0.59
0.76	0.73	0.73	0.72	0.72	0.72	0.71	0.71	0.70	0.69	0.67	0.65	0.61
0.78	0.75	0.75	0.74	0.74	0.74	0.73	0.73	0.72	0.71	0.70	0.68	0.64
0.80	0.77	0.77	0.77	0.76	0.76	0.76	0.75	0.74	0.73	0.72	0.70	0.67
0.82	0.79	0.79	0.79	0.79	0.78	0.78	0.77	0.77	0.76	0.75	0.73	0.70
0.84	0.82	0.81	0.81	0.81	0.81	0.80	0.80	0.79	0.78	0.77	0.76	0.73
0.86	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.81	0.80	0.78	0.75
0.88	0.86	0.86	0.86	0.85	0.85	0.85	0.84	0.84	0.83	0.82	0.81	0.79
0.90	0.88	0.88	0.88	0.88	0.87	0.87	0.86	0.86	0.86	0.85	0.84	0.82

\* Correlation coefficient r between two ROC areas A<sub>1</sub> and A<sub>2</sub> as a function of average correlation between ratings (rows) and average area (columns).

<sup>†</sup> (r<sub>N</sub> + r<sub>A</sub>)/2.

<sup>‡</sup> (A<sub>1</sub> + A<sub>2</sub>)/2.

## Ab142

Area under the ROC curve	
Area	0.7183
Std. Error	0.03208
95% confidence interval	0.6555 to 0.7812
P value	<0.0001

## APP669/Ab142

Area under the ROC curve	
Area	0.8281
Std. Error	0.02559
95% confidence interval	0.7779 to 0.8782
P value	<0.0001

## Ab140/Ab142

Area under the ROC curve	
Area	0.8373
Std. Error	0.02550
95% confidence interval	0.7873 to 0.8872
P value	<0.0001

## Composite

Area under the ROC curve	
Area	0.8829
Std. Error	0.02173
95% confidence interval	0.8403 to 0.9255
P value	<0.0001

Let's assume the data are from different mice...

$$Z = \frac{|Area_1 - Area_2|}{\sqrt{SE_{Area1}^2 + SE_{Area2}^2}}$$

Let's compare Ab142 to Composite

Null hypothesis: the two population curves have the same AUC

$$Z = \frac{|0.72 - 0.88|}{\sqrt{(0.032)^2 + (0.021)^2}} = 4.21$$

$$p = 2 * (1 - \text{NORMSDIST}(4.21)) = 0.00003$$

$$p < 0.0001$$

# Calculating p-value to compare two ROC curves.xlsx

A	B
	<i>Enter values from ROC analysis in shaded areas</i>
AUC largest	0.88
AUC smallest	0.72
difference (numerator)	0.16
SE for largest	0.021
SE for smallest	0.032
square largest	0.00044
square smallest	0.00102
SUM	0.00147
square root of SUM (denominator)	0.038
<b>z</b>	<b>4.18</b>
<b>p-value</b>	<b>0.000029</b>



## Ab142

Area under the ROC curve	
Area	0.7183
Std. Error	0.03208
95% confidence interval	0.6555 to 0.7812
P value	<0.0001

Let's assume the data are from different mice...

$$Z = \frac{|Area_1 - Area_2|}{\sqrt{SE_{Area1}^2 + SE_{Area2}^2}}$$

## APP669/Ab142

Area under the ROC curve	
Area	0.8281
Std. Error	0.02559
95% confidence interval	0.7779 to 0.8782
P value	<0.0001

Let's compare APP669/Ab142 to Ab140/Ab142

## Ab140/Ab142

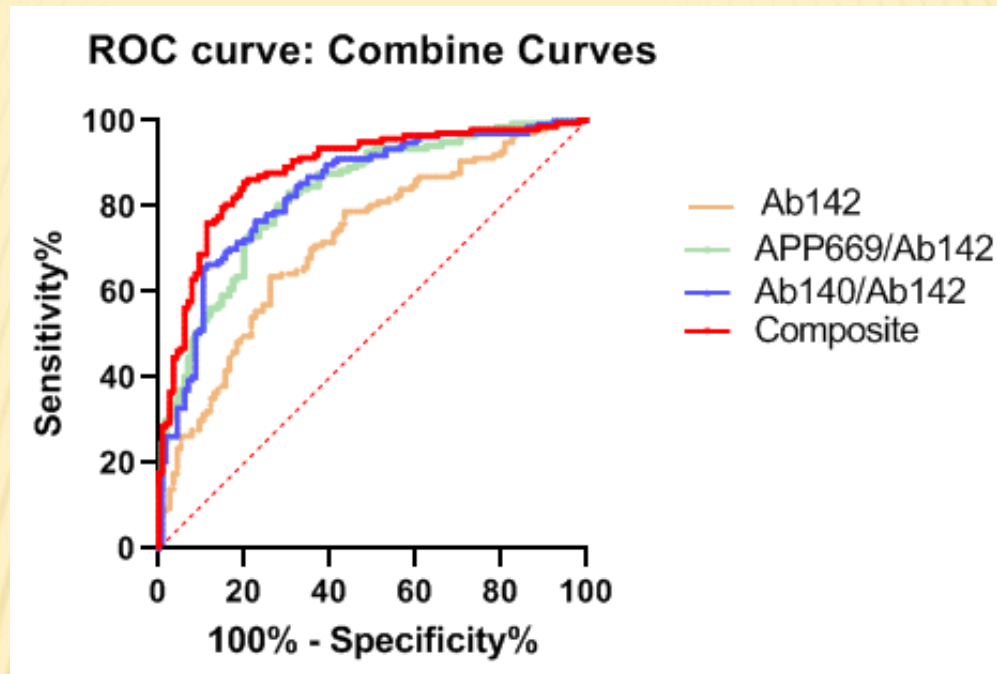
Area under the ROC curve	
Area	0.8373
Std. Error	0.02550
95% confidence interval	0.7873 to 0.8872
P value	<0.0001

$$Z = \frac{|0.83 - 0.84|}{\sqrt{(0.026)^2 + (0.026)^2}}$$

## Composite

Area under the ROC curve	
Area	0.8829
Std. Error	0.02173
95% confidence interval	0.8403 to 0.9255
P value	<0.0001

A	B
	<i>Enter values from ROC analysis in shaded areas</i>
AUC largest	0.84
AUC smallest	0.83
difference (numerator)	0.01
SE for largest	0.026
SE for smallest	0.026
square largest	0.00068
square smallest	0.00068
SUM	0.00135
Square root of SUM (denominator)	0.037
<b>z</b>	<b>0.27</b>
<b>p-value</b>	<b>0.785650</b>



Run ROC curve analyses for each pair of curves. Use a multiple testing correction (i.e., Bonferroni) to determine the differences between the different curves



# ROC Analyses Summary

The results of many clinical tests are on a continuous scale.

To help decide the presence or absence of disease, a cut-off point for 'normal' or 'abnormal' is chosen.

The sensitivity and specificity of a test vary according to the level that is chosen as the cut-off point.

The ROC curve, a graphical technique for describing and comparing the accuracy of diagnostic tests, is obtained by plotting the sensitivity of a test on the y axis against 1-specificity on the x axis.

The area under the ROC curve provides a measure of the overall performance of a diagnostic test.

# Power and sample size analyses: Why bother?

Sample size too big; too much power

wastes money and resources on extra subjects without improving statistical results

Sample size too small; having too little power to detect meaningful differences

exposure (treatment) discarded as not important when in fact it is useful

# Some definitions

## **Significance level ( $\alpha$ )**

Cut-off point for the p-value, below which the null hypothesis will be rejected and it will be concluded that there is evidence of an effect. Typically set at 5%. Probability of a Type I (false positive) error

## **$\beta$**

The probability of a Type II (false negative) error

## **Power ( $1-\beta$ )**

Power is the probability that the null hypothesis will be correctly rejected i.e. rejected when there is indeed a real difference or association. The higher the power, the lower the chance of missing a real effect. Power is typically set at 80% or 90% but not below 80%.



# Alpha and Beta

Convention  $\alpha=0.05$  and  $\beta=0.20$

Use lower alpha levels to reduce false positives

Use lower beta levels to reduce false negatives

Increasing sample size will reduce type I and type II errors

# Power

Definition: the probability that you reject the null hypothesis given that the alternative hypothesis is actually the truth

This is a correct conclusion.

$$\text{Power} = P(\text{reject } H_0 \mid H_A \text{ is true}) = 1 - \beta$$



Power describes the test's ability to minimize type-II errors (false negatives)



Since this is a good thing, we want power to be high

## Factors Affecting Power

1. Sample size
2. Size of the effect
3. Standard deviation of the characteristic
4. Significance level



	Group A	Group B	
	Cancer	No Cancer	
			
1	54	43	
2	50	38	
3	45	58	
4	43	47	
5	45	46	
6	44	51	
7	39	47	
8	40	50	
9	47	51	
10	42	56	
11	34	54	
12	48	41	
13	51	44	
14	37	39	
15	41	46	

		A	B
		Cancer	No Cancer
			
1	Number of values	15	15
2			
3	Mean	44.00	47.40
4	Std. Deviation	5.451	5.962
5	Std. Error of Mean	1.407	1.539
6			

### Unpaired t test

P value	0.1143
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.630, df=28

# Larger Sample Size

N=15

		A	B
		Cancer	No Cancer
1	Number of values	15	15
2			
3	Mean	44.00	47.40
4	Std. Deviation	5.451	5.962
5	Std. Error of Mean	1.407	1.539

N=30

		A	B
		Cancer	No Cancer
1	Number of values	30	30
2			
3	Mean	44.00	47.40
4	Std. Deviation	5.356	5.858
5	Std. Error of Mean	0.9779	1.070

Unpaired t test	
P value	0.1143
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.630, df=28

Unpaired t test	
P value	0.0224
P value summary	*
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.346, df=58

## Bigger Difference in Effect

Difference = 3.4 (n=15)

		A	B
		Cancer	No Cancer
1	Number of values	15	15
2			
3	Mean	44.00	47.40
4	Std. Deviation	5.451	5.962
5	Std. Error of Mean	1.407	1.539

Difference = 5.4 (n=15)

		A	B
		Cancer	No Cancer
1	Number of values	15	15
2			
3	Mean	44.00	49.40
4	Std. Deviation	5.451	5.962
5	Std. Error of Mean	1.407	1.539

### Unpaired t test

P value	0.1143
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.630, df=28

### Unpaired t test

P value	0.0151
P value summary	*
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.589, df=28



# Reduced Variability

Difference = 3.4 (n=15)

		A	B
		Cancer	No Cancer
1	Number of values	15	15
2			
3	Mean	44.00	47.40
4	Std. Deviation	5.451	5.962
5	Std. Error of Mean	1.407	1.539

Difference = 3.4 (n=15)

		A	B
		Cancer	No Cancer
1	Number of values	15	15
2			
3	Mean	44.00	47.40
4	Std. Deviation	3.817	3.247
5	Std. Error of Mean	0.9856	0.8384

## Unpaired t test

P value	0.1143
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.630, df=28

## Unpaired t test

P value	0.0138
P value summary	*
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.628, df=28

# Larger alpha: everything else the same

Parameters: t tests (and Nonparametric Tests) ×

Experimental Design Residuals Options

**Calculations**

P value: ☐ One-tailed ☒ Two-tailed (recommended)

Report differences as: No Cancer - Cancer

Confidence level: Other...  %

Definition of statistical significance:  $P < 0.2$

**Graphing options**

☐ Graph differences (paired)

☐ Graph ranks (nonparametric)

☐ Graph correlation (paired)

☐ Graph CI of difference between means

**Additional results**

☐ Descriptive statistics for each data set

☐ t test: Also compare models using AICc

☐ Mann-Whitney: Also compute the CI of difference between medians  
Assumes both distributions have the same shape.

☐ Wilcoxon: When both values on a row are identical, use method of Pratt  
If this option is unchecked, those rows are ignored and the results will match prior version of Prism

**Output**

Show this many significant digits (for everything except P values):

P value style: GP: 0.1234 (ns), 0.0332 (\*), 0.0021 (\*\*), 0.0001 (\*\*\*) N =

☐ Make options on this tab be the default for future tests.

Learn Cancel OK

		A	B
		Cancer	No Cancer
1	Number of values	15	15
2			
3	Mean	44.00	47.40
4	Std. Deviation	5.451	5.962
5	Std. Error of Mean	1.407	1.539

Alpha = 0.05

<b>Unpaired t test</b>	
P value	0.1143
P value summary	ns
Significantly different ( $P < 0.05$ )?	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.630, df=28

Alpha = 0.20

<b>Unpaired t test</b>	
P value	0.1143
P value summary	ns
Significantly different ( $P < 0.2$ )?	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=1.630, df=28

## Preparing to Calculate Sample Size

What statistical tests will be used?

(t-test, ANOVA, chi-square, regression etc)

What  $\alpha$  level will you use?

$\alpha=0.05$

The hard one: How small an effect size (or difference) is important to detect?

What difference would you not want to miss?

With what degree of certainty (power) do you want to detect the effect? (80-95%)

## Types of needs

What sample size do I need to have adequate power to detect a specific effect size (or difference)?

I only have  $N$  subjects available. What power will I have to detect a particular effect size (or difference) with that sample size?

I have no preliminary data, now can I determine the appropriate sample size?



<https://clincalc.com/stats/samplesize.aspx>

clincalc.com/stats/samplesize.aspx



# Sample Size Calculator

**Determines the minimum number of subjects for adequate study power**

 [ClinCalc.com](#) » [Statistics](#) » Sample Size Calculator

## Study Group Design



 vs. 

Two independent  
study groups

 vs. 

One study group  
vs. population

Two study groups will each receive different treatments.

## Primary Endpoint



Dichotomous  
(yes/no)



Continuous  
(means)

The primary endpoint is an **average**.  
*Eg, blood pressure reduction (mmHg), weight loss (kg)*

## Statistical Parameters

### Anticipated Means

Group 1 ?

 ± 

Group 2 ?

Mean ▼

Enrollment ratio ?

### Type I/II Error Rate

Alpha ?

0.05

Power ?

80%

Reset

Calculate

Press 'Calculate' to view calculation results.

## Anticipated Means

Group 1 ?

10

±

5

Group 2 ?

20

Mean ▼

Enrollment ratio ?

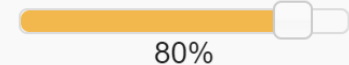
1

## Type I/II Error Rate

Alpha ?

0.05

Power ?



80%

Reset

Calculate

## RESULTS

### Continuous Endpoint, Two Independent Sample Study

#### Sample Size

Group 1	4
Group 2	4
<b>Total</b>	<b>8</b>

#### Study Parameters

Mean, group 1	10
Mean, group 2	20
Alpha	0.05
Beta	0.2
Power	0.8

## Smaller difference

### RESULTS

#### Continuous Endpoint, Two Independent Sample Study

Sample Size	
Group 1	16
Group 2	16
<b>Total</b>	<b>32</b>

Study Parameters	
Mean, group 1	10
Mean, group 2	15
Alpha	0.05
Beta	0.2
Power	0.8

## Double the variability

### RESULTS

#### Continuous Endpoint, Two Independent Sample Study

Sample Size	
Group 1	16
Group 2	16
<b>Total</b>	<b>32</b>

Study Parameters	
Mean, group 1	10
Mean, group 2	20
Alpha	0.05
Beta	0.2
Power	0.8



Change alpha to 0.01

## RESULTS

### Continuous Endpoint, Two Independent Sample Study

Sample Size	
Group 1	6
Group 2	6
<b>Total</b>	<b>12</b>

Study Parameters	
Mean, group 1	10
Mean, group 2	20
Alpha	0.01
Beta	0.2
Power	0.8

Be prepared for your meeting with a statistician to  
discuss power and sample size

Be prepared to discuss research questions and study design

Bring preliminary data that show type and variability of data  
Can use published data

Think about your acceptable measure of effect  
How small a difference is clinically/biologically significant

