

La validazione dei metodi diagnostici

S. Ricci, G. Lanza

SINV

Soc. Italiana Interdisciplinare

NeuroVascolare

Statistics is like a bikini:
what they show is intriguing,
but what they conceal is vital

Verificare, verificare.....

- Un nuovo tipo di esame: ci prenderà?
(Validità)
- Lo stesso problema, ma due opinioni diverse.... (Riproducibilità)

NESSUNA UMANA INVESTIGAZIONE
SI PUO' DOMANDARE VERA
SCIENZA SE NON PASSA PER LE
MATEMATICHE DIMOSTRAZIONI

Leonardo Da Vinci

Validità di un test diagnostico ovvero “perchè serva a qualcosa...”

- Perchè una manovra diagnostica sia utile, occorre che individui ciò che a noi serve individuare con la maggiore accuratezza possibile, quando confrontata con analoga e validata manovra, che però preferiamo non utilizzare per costi, rischi, complessità, etc (gold standard).

Ogni test diagnostico ha quattro possibili risultati...

- Può essere positivo, proprio come sarebbe stato il gold standard
- Può essere negativo, proprio come sarebbe stato il gold standard
- Può essere negativo, ma il gold standard sarebbe stato positivo
- Può essere positivo, ma il gold standard sarebbe stato negativo

La stessa cosa, ma in tabella!

		Gold standard	
		+	-
test	+	a Vero positivo	b Falso positivo
	-	c Falso negativo	d Vero negativo

Qualche definizione...

- $a/a+c$ = sensibilità, ovvero la percentuale di esami realmente positivi che il test riesce ad evidenziare

		GS	
		+	-
test	+	a	b
	-	c	d

Ancora definizioni...

- $d/b+d$ = specificità, ovvero la percentuale di esami realmente negativi che il test riesce ad evidenziare

		GS	
		+	-
test	+	a	b
	-	c	d

Attenzione alla pratica...

- Sensibilità e specificità sono percentuali, e quindi si possono calcolare per esse gli IC, che saranno tanto più ampi quanto più limitato è il numero degli esami considerati
- Nella pratica clinica, però, noi non conosciamo il risultato dell' esame "gold standard", ed il risultato del test è tutto ciò che abbiamo; perciò la domanda vera è: **qual' è la probabilità che il test ci dia una diagnosi corretta?**

Validità di un test diagnostico: un esempio (stenosi carotidea di interesse chirurgico)

		Angiografia	
		+	-
Doppler	+	90	50
	-	10	850

Sensibilità 90% (83%-95%)

Specificità: 94% (93%-96%)

Ma qual' è la probabilità che il doppler dica il vero? A questa domanda non danno risposta nè la sensibilità nè la specificità.....

Altre, necessarie, definizioni...

- $a/a+b =$ Valore Predittivo Positivo, ovvero la percentuale di esami positivi che sono realmente patologici

		+	GS	-
test	+	a	b	
	-	c	d	

Ancora ...

- $d/c+d = \underline{\text{Valore Predittivo Negativo}}$, ovvero la percentuale di esami negativi che sono realmente normali

		+	GS	-
test	+	a	b	
	-	c	d	

Dall' esempio precedente

		Angiografia	
		+	-
Doppler	+	90	50
	-	10	850

Sensibilità 90% (83%-95%)

Specificità: 94% (93%-96%)

VPP: 64% (56%-72%)

VPN: 99% (98%-99%)

Allora usiamo i valori predittivi?

- **Attenzione!!** I VP di un test nella pratica clinica dipendono in maniera sostanziale dalla prevalenza della patologia nella popolazione studiata, e questa può essere anche molto differente da quella osservata negli studi che hanno calcolato originariamente i VP!!
- **Perciò voi dovete cercarvi i vostri VP!**

Spiegazione con un esempio

- Se eseguiamo un ecodoppler carotideo su tutti i presenti in questa aula, la probabilità pre esame (o prevalenza) di stenosi di interesse chirurgico sarà certamente molto bassa
- Ma se facciamo la stessa cosa in una corsia dove sono ricoverati pazienti con recente ictus ischemico non lacunare, che non fibrillano, la prevalenza sarà molto alta
- Lo stesso esame, fatto dallo stesso operatore, avrà VP molto diversi in queste due circostanze

Con la tabella (sensibilità 90%, specificità 94%, ma prevalenza 5% o 75%)

	+ Angiografia	-
+ Doppler	45	57
-	5	893

VPP 44%
VPN 99%

	+ Angiografia	-
+ Doppler	675	15
-	75	235

VPP 98%
VPN 76%

Una regola generale

- Più bassa è la prevalenza della patologia, più alto è il VPN e più basso è il VPP, e viceversa.
- Esiste una formula universale valida per ogni prevalenza:
 - $VPP = \frac{\text{sensibilità} \times \text{prevalenza}}{\text{sensibilità} \times \text{prevalenza} + (1 - \text{specificità}) \times (1 - \text{prevalenza})}$
 - $VPN = \frac{\text{specificità} \times (1 - \text{prevalenza})}{(1 - \text{sensibilità}) \times \text{prevalenza} + \text{specificità} \times (1 - \text{prevalenza})}$

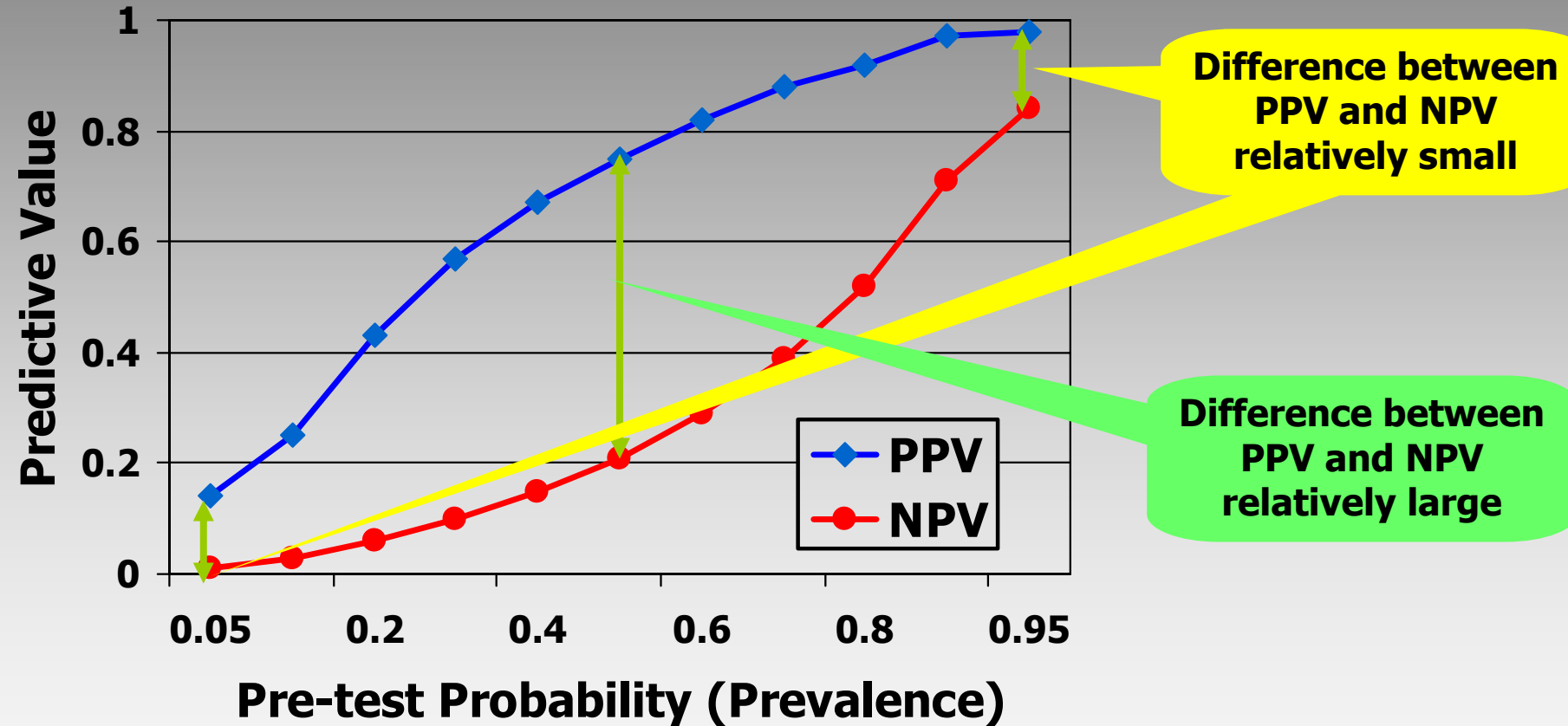
Probabilità prima, probabilità dopo

- La prevalenza può essere considerata come la probabilità di malattia prima del test
- I VP possono essere considerati come probabilità di malattia dopo il test
- La differenza tra probabilità pre test e probabilità post test ci dà un'idea della utilità del test stesso in quel contesto clinico

Attenti agli screening...

- Perfino test esternamente specifici, se utilizzati in condizioni di bassa prevalenza di patologia, daranno luogo ad un elevato numero di falsi positivi
- A causa di ciò, in condizioni di bassa prevalenza di patologia, il VPP di un test è comunque basso
- Tuttavia, il VPN ne risulta molto meno influenzato

RELATIONSHIP BETWEEN PREVALENCE AND PREDICTIVE VALUE



Based on a test with 90% sensitivity and 82% specificity

Definizioni di validità

- $a/a+c$ = sensibilità
- $d/b+d$ = specificità
- $a/a+b$ = VPP
- $d/c+d$ = VPN
- $a+d/a+b+c+d$ = Accuratezza
- $a+c/a+b+c+d$ = Probabilità pre-test +
- $b+d/a+b+c+d$ = Probabilità pre-test -

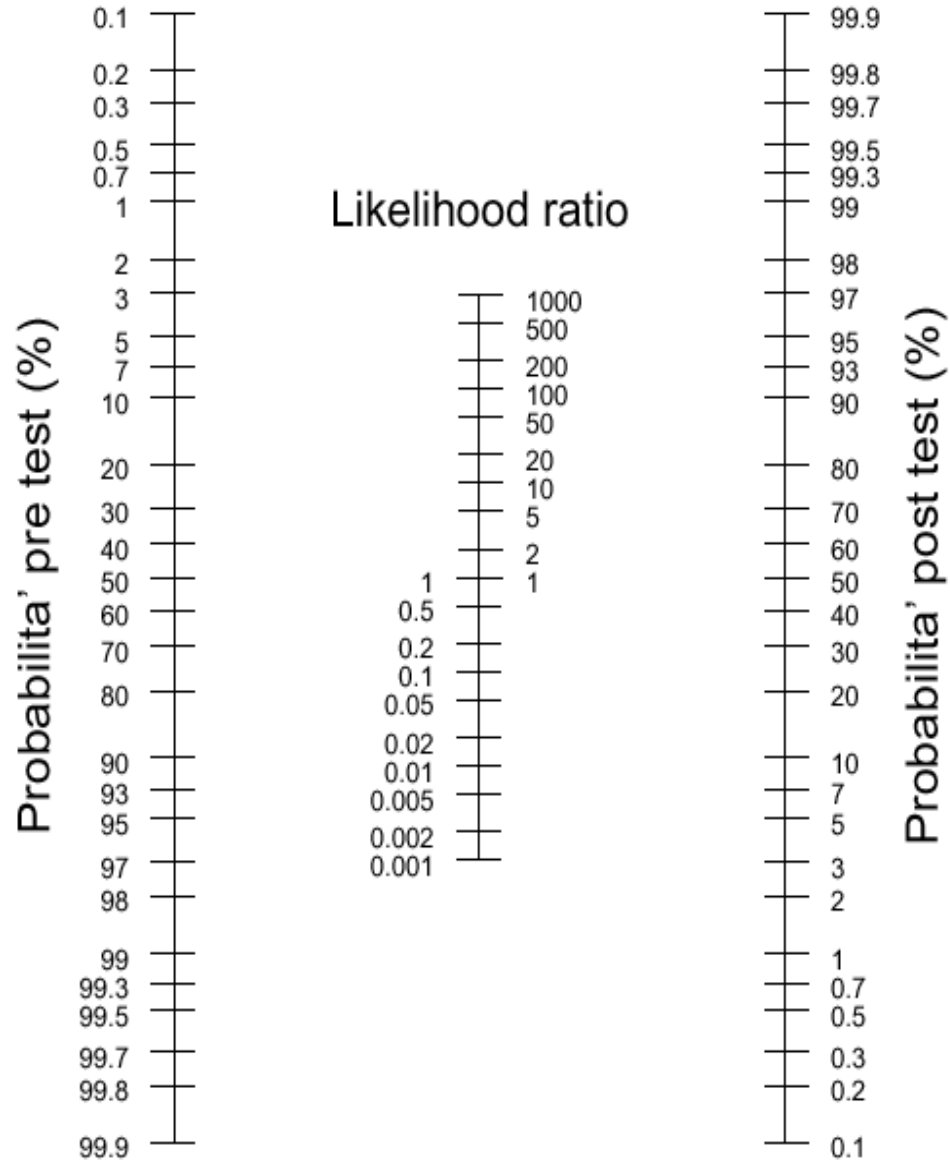
Un passo in più....

- $LR+$ = probabilità di un risultato positivo del test in un soggetto malato rispetto alla stessa in un soggetto sano ($SE/1-SP$)
- $LR-$ = probabilità di un risultato negativo del test in un soggetto malato rispetto alla stessa in un soggetto sano ($1-SE/SP$)

A che serve?

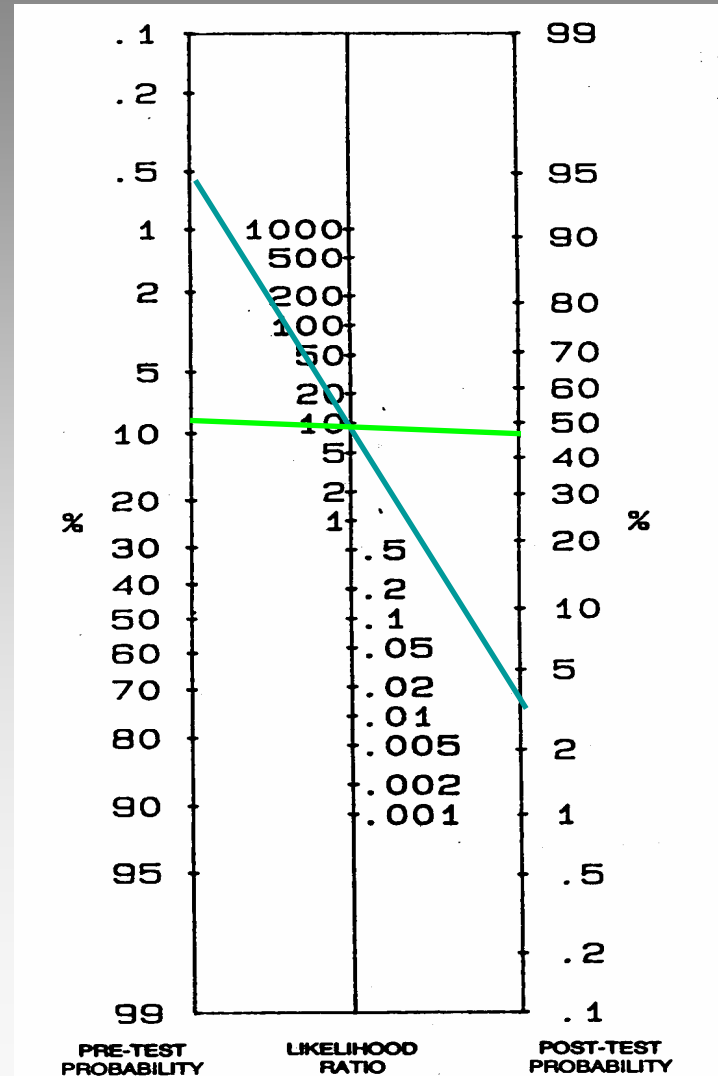
- Mediante un semplice nomogramma, utilizzando le LR, è possibile stabilire il guadagno diagnostico di ogni procedura! (Nell' esempio precedente, si passa da 10% a 62%, con un guadagno del 52%)
- In linea di principio, se la probabilità pre-test è alta, il guadagno si avrà solo in senso negativo, e viceversa.
- Il valore da dare al rischio di falsi positivi e di falsi negativi dipende da cosa faremo dopo!

Nomogramma di Fagan



LIKELIHOOD RATIO AND PRE- AND POST-TEST PROBABILITIES

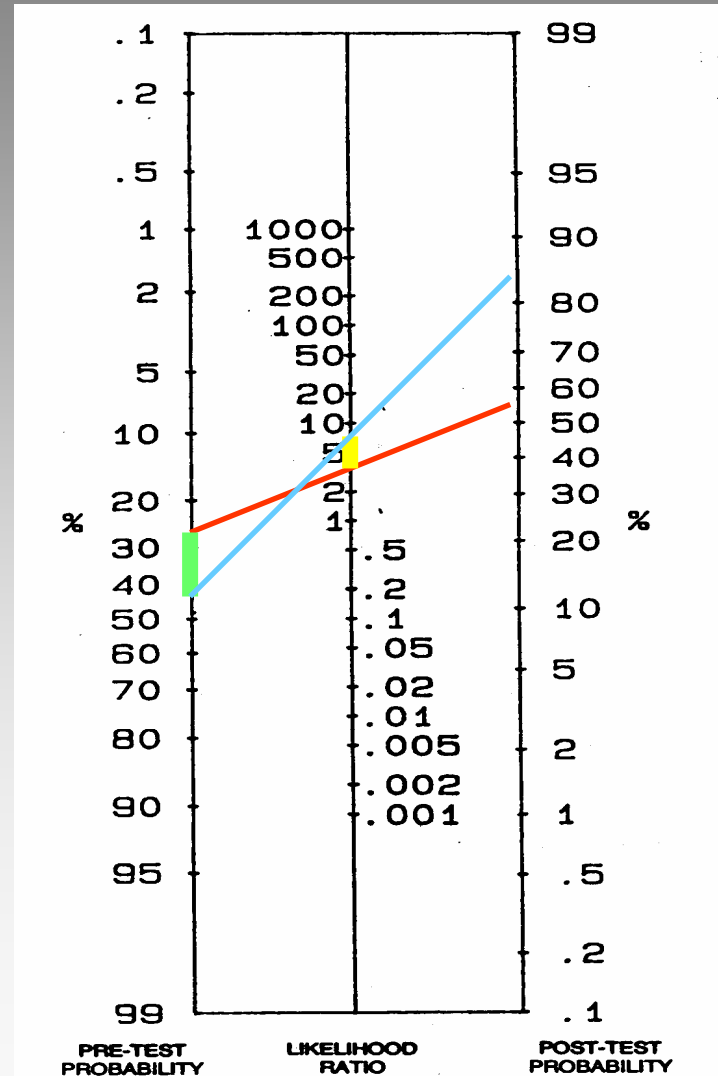
For a given test with a given likelihood ratio, the post-test probability will depend on the pre-test probability (that is, the prevalence of the condition in the sample being assessed)



SENSITIVITY ANALYSIS OF A DIAGNOSTIC TEST

	Value	95% CI
Pre-test probability	35%	26% to 44%
Likelihood ratio	5.0	3.0 to 8.5

Applying the 95% confidence intervals above to the nomogram, the post-test probability is likely to lie in the range 55-85%



Calcolo on line della sensibilità e specificità

<http://www.medcalc.com/bayes.html>

Normale o anormale: qual è il cut off ?

- Molti test non danno immediatamente un risultato “si/no”, ma sono di tipo quantitativo
- Per scegliere il miglior cut off occorre in questi casi calcolare sensibilità e specificità per ogni livello, e quindi trasferirli su un grafico con sensibilità e 1-specificità sugli assi
- Il valore più vicino al vertice superiore sn è il migliore (cirve ROC)

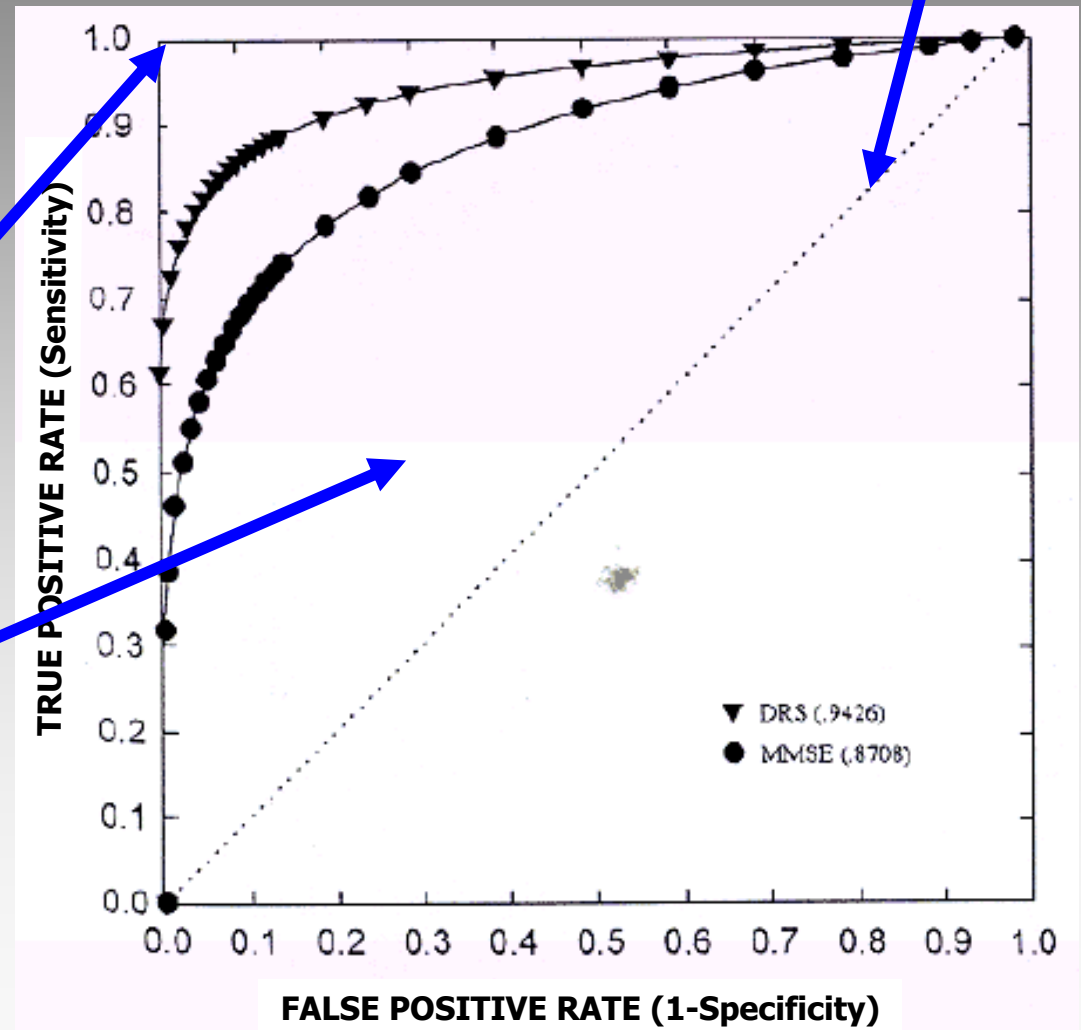
The diagonal line (representing Sensitivity=0.5 and Specificity=0.5) represents performance no better than chance

RECEIVER OPERATING CHARACTERISTIC CURVE

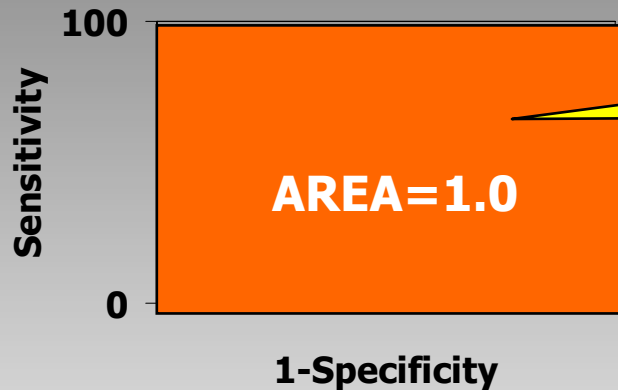
Overall shape is predicted by the reciprocal relationship between sensitivity and specificity

The closer the curve gets to Sensitivity=1 and Specificity=1, the better the overall performance of the test

Hence the area under the curve gives a measure of the test's performance

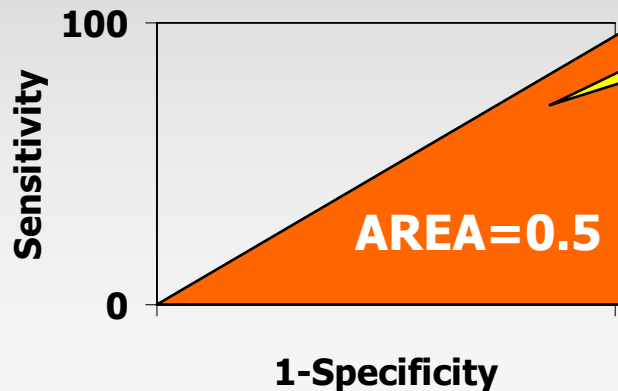


AREA UNDER ROC CURVES



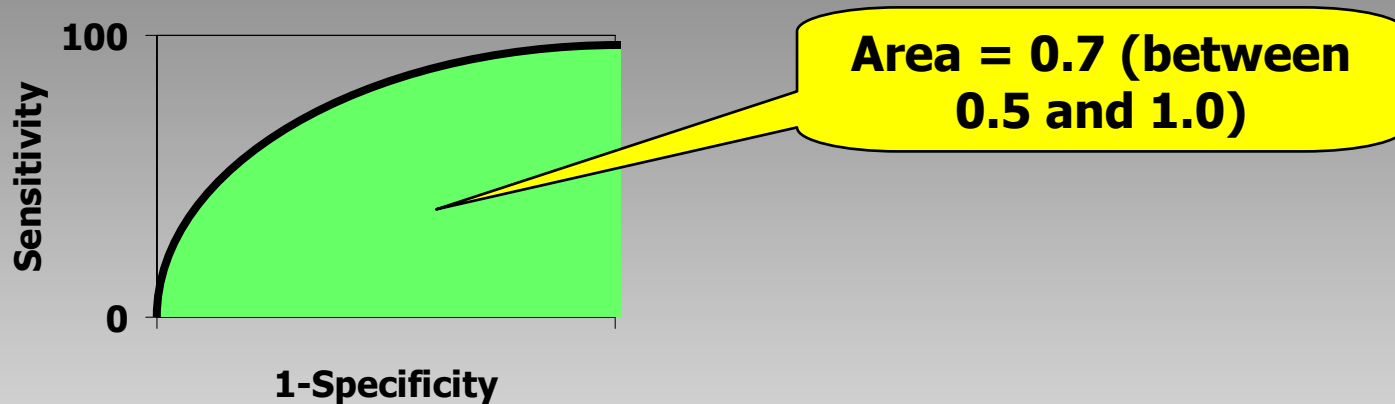
**Sensitivity and specificity
both 100% - TEST PERFECT**

**Sensitivity and specificity
both 50% - TEST USELESS**

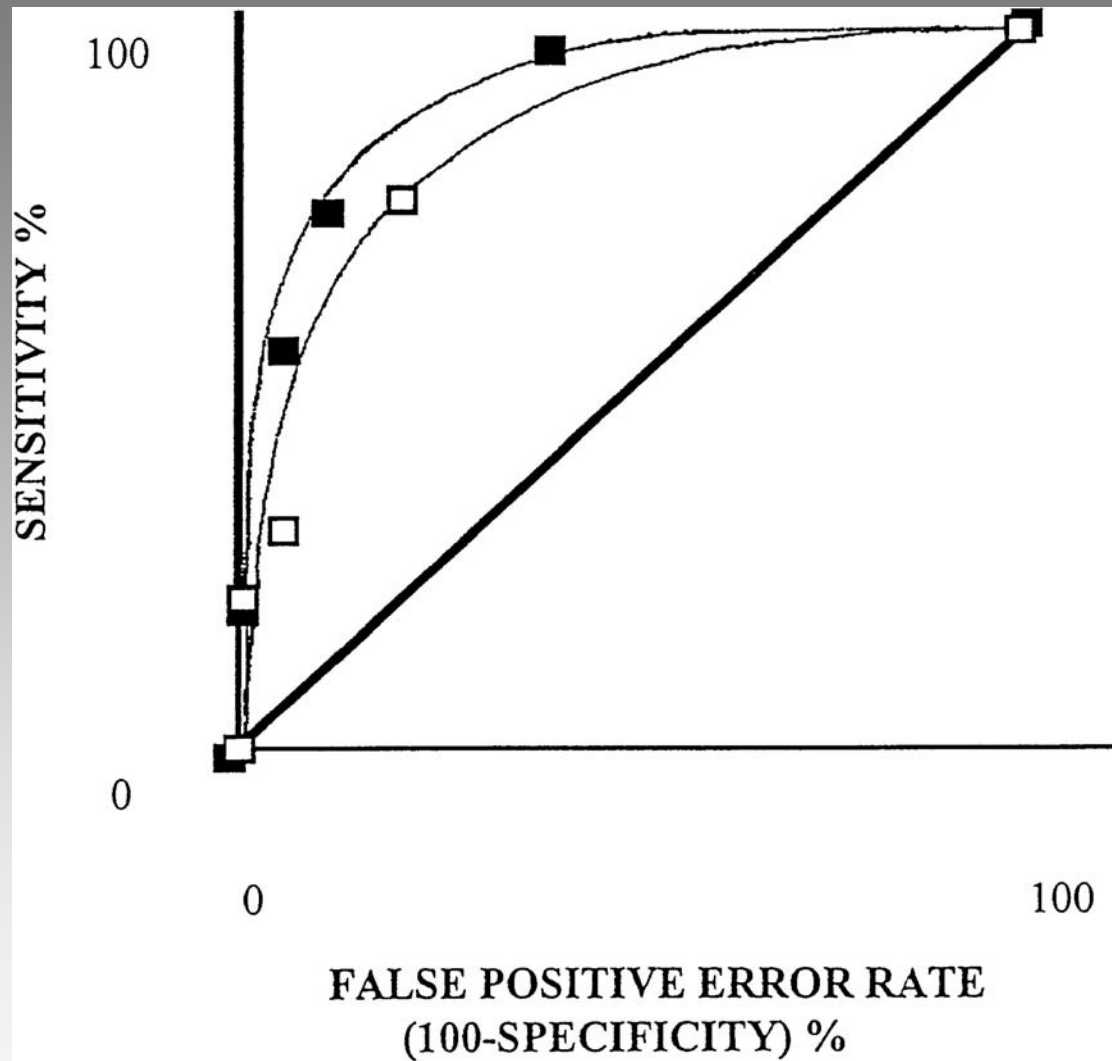


**The area under a ROC
curve will be between
0.5 and 1.0**

AREA UNDER ROC CURVES



- Consider (hypothetically) two patients drawn randomly from the DISEASE+ and DISEASE- groups respectively
- If the test is used to guess which patient is from the DISEASE+ group, it will be right 70% of the time



Riproducibilità ovvero accordo oltre il caso

- Il problema: due operatori danno un giudizio su un problema clinico (sintomo, segno, dato strumentale, etc); non sempre si troveranno d' accordo!
- Si potrebbe semplicemente calcolare la percentuale di casi in cui i due si trovano d' accordo, e -se questa è elevata- sostenere che l' interpretazione del dato ha una buona riproducibilità

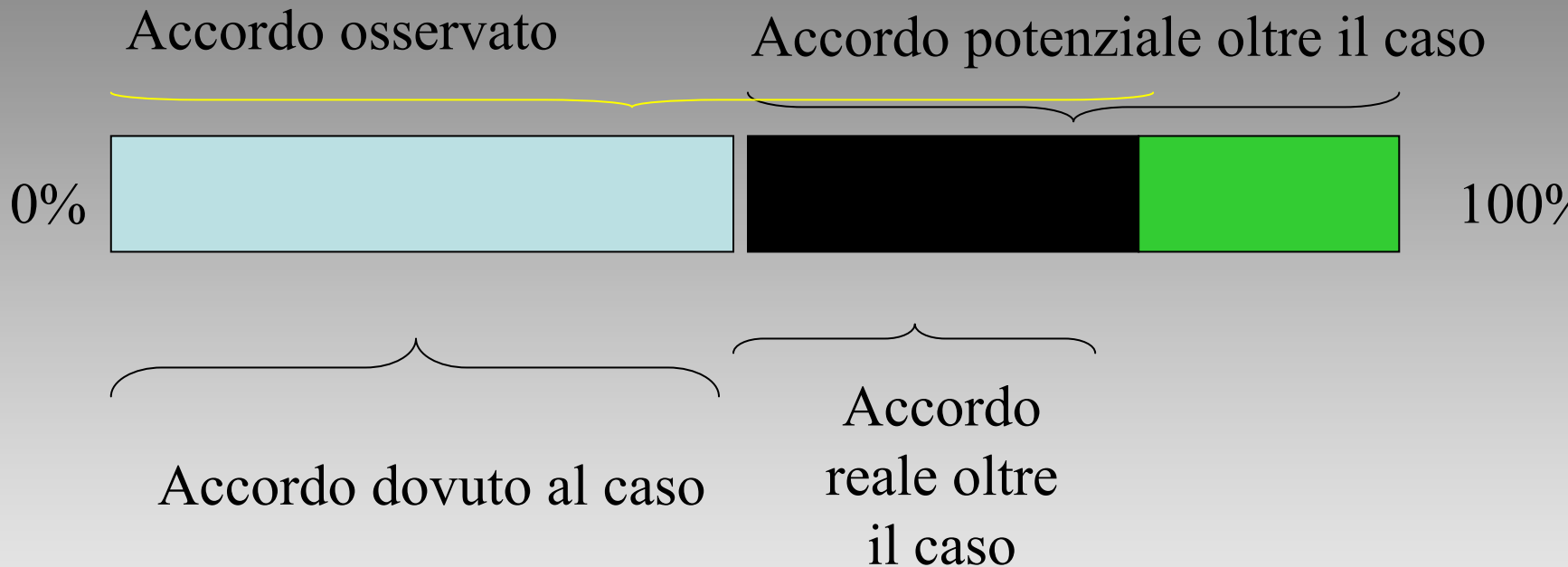
Riproducibilità ovvero accordo oltre il caso

2

Attenzione! Se uno dei due si limitasse a tirare in aria una moneta, e a dare la sua opinione in base al risultato, senza neanche valutare il problema clinico, comunque ci sarebbe una certa percentuale di accordo.

Se si vuole definire con certezza la riproducibilità di un segno, sintomo o dato di laboratorio, occorre eliminare la percentuale di accordo dovuta al caso.

Indice Kappa



Kappa: Accordo reale oltre il caso
accordo potenziale oltre il caso

Indice Kappa: un esempio

- Due osservatori leggono cento radiografie, e le classificano come patologica o normale
- I risultati sono:

	N	P
N	40	5
P	6	49

Il Kappa è 0,78

Indice Kappa: valori

- $< 0,20$ = Accordo scarso
- $> 0,20 < 0,40$ = Accordo buono
- $> 0,40 < 0,60$ = Accordo discreto
- $> 0,60 < 0,80$ = Accordo considerevole
- $> 0,80$ = Accordo perfetto

Variazioni sul...Kappa

- L' indice Kappa può essere calcolato anche per livelli diversi di risposta (p.es. Normale, lievemente alterato, francamente alterato)
- In questo caso si può ulteriormente “pesare” il disaccordo: se un operatore dice “normale” e l' altro dice “francamente alterato”, il disaccordo è più grave rispetto a quello tra “francamente alterato” e “lievemente alterato”.