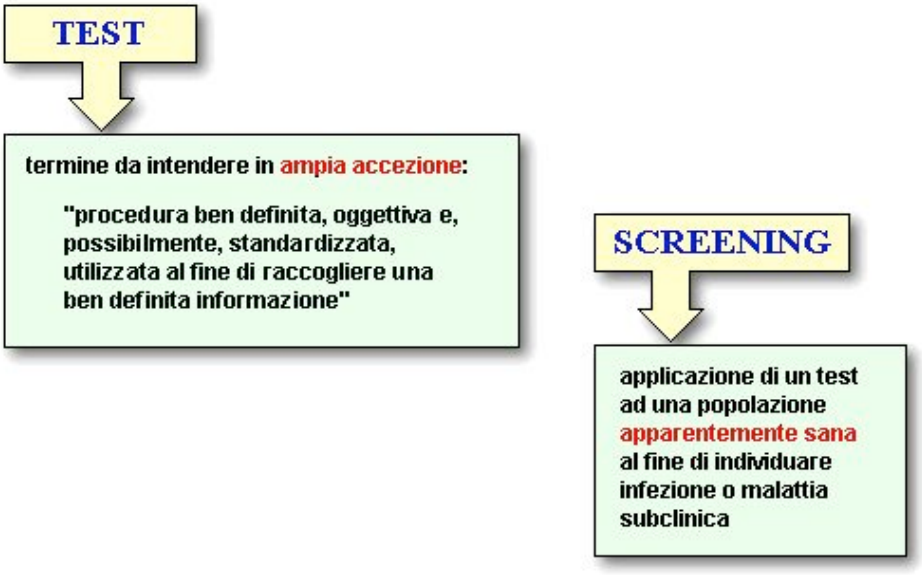
Test di screening e test diagnostici

Utilizzo di un test per lo screening di popolazioni

In questo capitolo tratteremo dei test in una accezione più ampia, intendendo per «test» qualsiasi ben definita procedura, oggettiva e possibilmente standardizzata, che viene utilizzata al fine di raccogliere una ben definita informazione (procedura diagnostica).

Un «test» di screening è un test che viene applicato alla popolazione apparentemente sana (o a stato sanitario ignoto) soggetta ad una probabilità («rischio») più o meno elevata di presentare una data malattia.



In medicina, lo screening viene indirizzato preferenzialmente a quelle condizioni morbose in cui una diagnosi precoce ed il conseguente intervento terapeutico siano in grado di ridurre incidenza, prevalenza o mortalità.

Anche se azioni di screening e procedimenti diagnostici possono essere effettuate impiegando lo stesso «test», tuttavia lo screening differisce dalla diagnosi.

Infatti, contrariamente allo screening, nel procedimento diagnostico l'eventuale impiego di un «test» viene attuato su soggetti ammalati, cioè che mostrano sintomi clinici che, in una qualche misura, fanno sospettare la presenza di quella malattia;

mentre le azioni di screening, anche se implementate in popolazioni nelle quali la malattia è presumibilmente presente, prevedono l'applicazione del test su tutti gli individui della popolazione, indipendentemente dal loro stato di salute;

quanto detto serve a capire che poiché il valore predittivo di un test dipende dalla prevalenza della malattia, ne consegue che la performance del test sarà meno soddisfacente in caso di screening rispetto al caso in cui lo stesso test venga utilizzato a scopo diagnostico. Infatti, è evidente che la prevalenza della malattia fra gli individui che mostrano segni clinici sarà superiore rispetto alla prevalenza considerata nella popolazione nel suo complesso. Torneremo su questo punto.

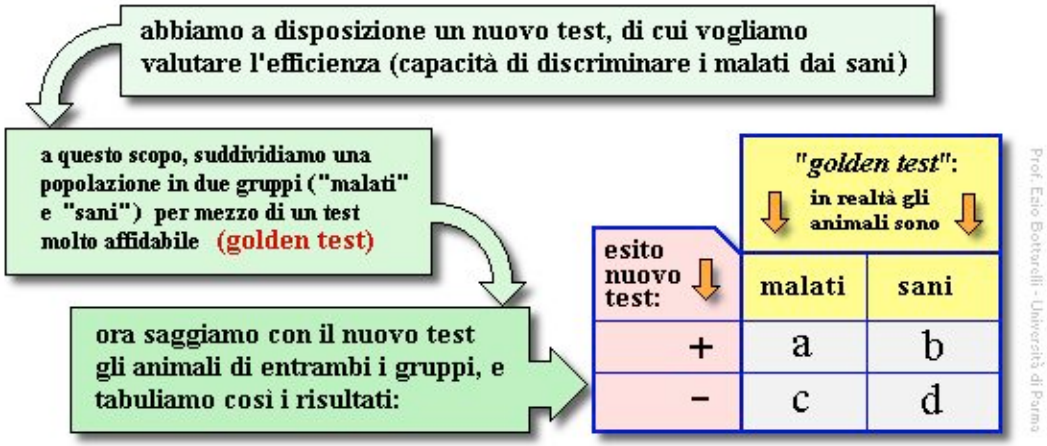
Per buono che sia, un test non fornisce mai risultati perfettamente rispondenti alla realtà. C'è sempre un rischio - anzi, una probabilità - che il test classifichi come "positivi" (cioè infetti o ammalati) soggetti che in realtà sono "negativi" (cioè non-infetti o sani).

D'altra parte, esiste anche il rischio inverso, cioè che il test restituisca un risultato "negativo" se applicato ad un soggetto che, in realtà, è ammalato. Quanto detto serve a capire l'utilità di un test diagnostico, ovvero la identificazione degli individui realmente positivi o negativi con la maggiore accuratezza.

Valutazione della performance di un test

Non esistono test capaci di individuare la presenza di una malattia e capaci di fornire risultati certi ed affidabili in tutte le situazioni e nel 100% dei casi. In altre parole: non esistono test «infallibili». L'esito (sia esso positivo, cioè deponga a favore dell'esistenza della malattia, o negativo) deve essere visto come una indicazione di «probabilità», tranne che nel caso in cui il test quando fornisce un esito positivo indica con certezza la presenza della malattia. Ma questa è una condizione assai rara a cui noi non ci riferiamo.

Possiamo inoltre supporre che la probabilità di ottenere risultati «veri» (cioè aderenti alla realtà) sia soprattutto legata al tipo di test, e che non tutti i test raggiungano la stessa probabilità, ma che possa essere invece stilata una sorta di 'classifica' della performance dei vari test.



Per valutare la performance di un test è necessario disegnare un esperimento allo scopo di ottenere i dati delle quattro celle (a, b, c, d) della tabella a doppia entrata. Vediamo di spiegare meglio.

Immagina di dover giudicare la performance di un test che, per semplicità, supponiamo fornire un risultato del tipo vero/falso o, per meglio dire, sano/malato.

È evidente che la performance dipende dalla quota di risultati falsi-positivi (cella b della Tabella) e falsi-negativi (cella c della Tabella) ottenuti applicando il test ad una popolazione che comprende sia soggetti malati che soggetti sani. Evidentemente, un test infallibile farà registrare valori nulli nelle celle «b» e «c». Tuttavia, i test infallibili purtroppo non esistono.

Il nuovo test non è infallibile; quando lo applichi al campione in studio, potrai suddividere i soggetti in due categorie: quelli test-positivi (a+b) e quelli test-negativi (c+d). Quindi, poiché il test non è infallibile, dovrai rispondere a due domande:

(1) «quanti dei soggetti test-negativi sono falsi-negativi?»

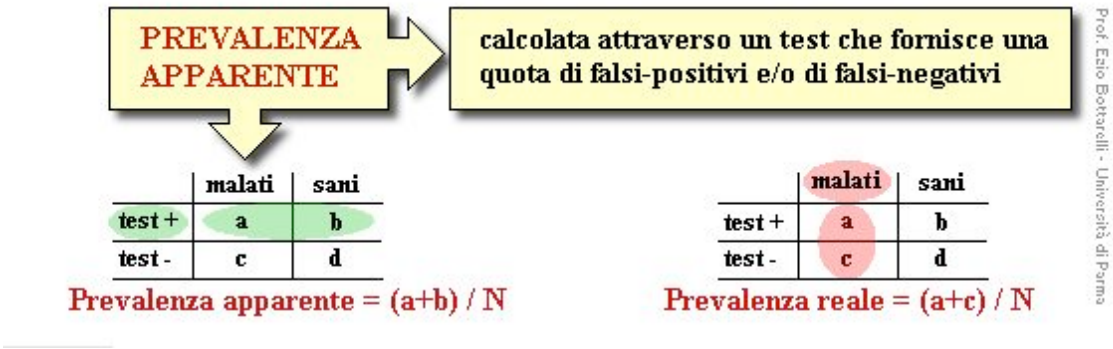
(2) «quanti dei soggetti test-positivi sono falsi-positivi?»

Per rispondere alle due domande dobbiamo conoscere qual è lo stato sanitario «vero» di ciascun soggetto saggiato con il nostro test. A questo scopo sottoponiamo la popolazione ad esame con un altro test differente dal primo. Questo secondo test lo chiamiamo “di riferimento”, e quindi dovrà essere il migliore disponibile (gold standard), cioè quello che fornisce i migliori risultati (idealmente: nessun falso negativo e nessun falso positivo).

In base ai risultati del gold test potremo individuare le frequenze delle quattro classi a, b, c, d (secondo lo schema) e quindi procedere nella valutazione del test in studio sulla base delle sue caratteristiche.

L'individuazione di un gold-test non è sempre facile. Si può utilizzare un test relativamente semplice e poco costoso. Ma potrebbe essere utilizzato anche un test molto complesso e costoso, ma che si caratterizza per la elevata precisione dei risultati. In alcuni casi la migliore proposta come test di riferimento consiste direttamente nella valutazione clinica a posteriori del paziente.

In base a quanto finora esposto, è chiaro che un test diagnostico fornisce quasi sempre una certa quota di risultati falsi-positivi e falsi-negativi. Di conseguenza, il calcolo della prevalenza di una malattia in una popolazione (campione) in base al risultato di un test non fornisce il valore della prevalenza reale, bensì la cosiddetta "prevalenza apparente".



Sensibilità e specificità di un test

La sensibilità e la specificità sono due misure che vengono impiegate per valutare la capacità di individuare, fra gli individui di una popolazione, quelli provvisti del «carattere» ricercato e quelli che invece ne sono privi. In pratica, per i nostri scopi, il «carattere» è rappresentato quasi sempre dalla malattia .

Ricordati che la disposizione delle variabili a, b, c, d nella tabella 2x2 è codificata, e che sebbene questa disposizione non è obbligatoria, è quella adottata più comunemente e rappresenta uno "standard".

Sensibilità

La sensibilità è la capacità di identificare correttamente i soggetti ammalati.

La sensibilità risponde alla domanda: "Fra i soggetti malati, quanti risultano test-positivi?"

In termini di probabilità, la sensibilità è la probabilità che un soggetto ammalato risulti positivo al test. Possiamo anche dire che la sensibilità è la proporzione di soggetti ammalati che risultano positivi al test.

Quest'ultima definizione è la più indicata per risalire al calcolo del valore di sensibilità: nella tabella i soggetti ammalati sono rappresentati da (a+c) e, fra questi, i positivi al test sono rappresentati da (a); quindi, la sensibilità si calcola con la proporzione a/(a+c).

La frazione a/(a+c) ha la particolarità di includere al denominatore il valore presente al numeratore; si tratta quindi di una proporzione che può assumere soltanto un valore compreso fra 0 e 1 (esprimibile anche come valore percentuale da 1 a 100).

Ad un esame superficiale, potrebbe sembrare che la sensibilità sia l'unica qualità desiderabile in un test: infatti, sembrerebbe un eccellente risultato il poter identificare correttamente tutti i soggetti con la malattia impiegando un test con una sensibilità del 100%.

Tuttavia, esaminando meglio la questione, si giunge alla conclusione che la suddetta qualità non è sufficiente. Infatti, è necessario anche un altro requisito: il test deve identificare come positivi soltanto i soggetti che hanno la malattia; cioè, è necessario che fra i positivi al test non siano inclusi anche soggetti sani. Da questa osservazione discende il concetto di specificità.

Specificità

La specificità è la capacità di identificare correttamente i soggetti sani.

La specificità risponde alla domanda: "Fra i soggetti sani, quanti risultano test-negativi?"

In termini di probabilità, la specificità è la probabilità che un soggetto sano risulti negativo al test. Possiamo anche dire che la specificità è la proporzione di soggetti sani che risultano negativi al test.

Quest'ultima definizione è la più indicata per risalire al calcolo del valore di specificità. Nella tabella i soggetti sani sono rappresentati da (b+d) e, fra questi, i negativi al test sono rappresentati da (d); quindi, la specificità si calcola con la proporzione d/(b+d).

Notare che anche la specificità, analogamente alla sensibilità, è definita attraverso una proporzione e quindi assume un valore compreso fra 0 e 1.

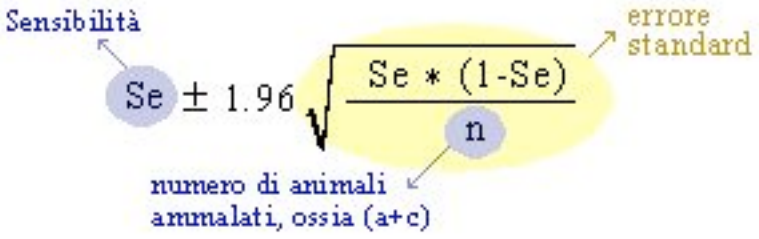
Nelle operazioni di screening su larga scala, che coinvolgono un elevato numero di individui, la specificità del test è di grande importanza. Ad esempio, nel 2002 in Italia sono stati effettuati 746.678 test per la BSE (encefalopatia spongiforme bovina); di essi, 34 sono risultati positivi. Questi dati indicano che il test utilizzato era dotato di specificità straordinariamente elevata. Se si fosse utilizzato un test con specificità pari a 0.99 (ossia 99%), l'1% dei bovini SANI saggiati sarebbe risultato positivo: ossia ben 7467 soggetti!

Stima della sensibilità e specificità

I valori di sensibilità e specificità di un test vengono di norma calcolati sperimentalmente su un campione. In questo caso, la variabilità dovuta al caso può essere ampia e, quindi, i valori ottenuti possono essere poco affidabili.

Pertanto, soprattutto quando il campione studiato è piccolo, è opportuno calcolare l'intervallo di confidenza (es. intervallo di confidenza 95%), che serve a quantificare la precisione della stima dei valori di sensibilità e specificità.

Per il calcolo dell'intervallo di confidenza 95% di un dato valore di sensibilità, si utilizza la formula seguente:



Ovviamente la suddetta formula, con le opportune variazioni, può essere utilizzata anche per il calcolo dell'intervallo di confidenza della specificità: basta sostituire Se con Sp e n con il totale degli soggetti non-ammalati (ossia b+d).

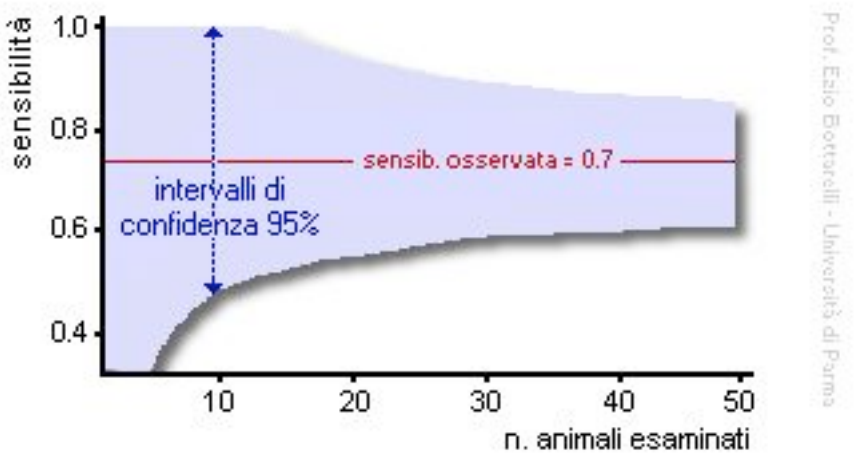
ESEMPIO. Hai applicato un test in un campione di 94 soggetti. Quaranta soggetti hanno la malattia, e 54 non hanno la malattia. Dei 40 soggetti che hanno la malattia, sono positivi al test 36. Quindi la sensibilità del test è di 36/40 = 0.9. Puoi calcolare l'intervallo di confidenza 95% applicando il metodo ora ricordato.

Prima è necessario calcolare la varianza: (0.9*0.1)/40 = 0.00225. Poi si calcola l'errore standard estraendo la radice quadrata della varianza: radq(0.00225)=0.0474. Infine, si calcolano i limiti fiduciale come segue:

Limite inferiore: 0.9-(1.96*0.0474)=0.807

Limite superiore: 0.9+(1.96*0.0474)=0.993

L'ampiezza dell'intervallo di confidenza dipende dal numero n di soggetti che hai esaminato: più grande è questo numero, più ristretto è l'intervallo di confidenza. Il grafico seguente illustra la precisione di una stima di sensibilità di un test in rapporto al numero di soggetti esaminati.



Nel grafico di esempio vengono riportati gli intervalli fiduciali ottenuti per un test la cui sensibilità è risultata pari a 0.75 (75%).

Per concludere, si può ricordare che...

Sensibilità e specificità: influenza del valore di soglia (cut-off)

Finora abbiamo illustrato le caratteristiche di un ipotetico test presumendo che esso fornisse risultati del tipo (positivo/negativo) oppure (sano/malato) oppure (si/no). Questo tipo di misurazione, in cui i dati vengono suddivisi in due categorie, viene detto «nominale dicotomico», ed i test vengono detti «qualitativi» in quanto misurano la «qualità» e non la «quantità» di un fenomeno.

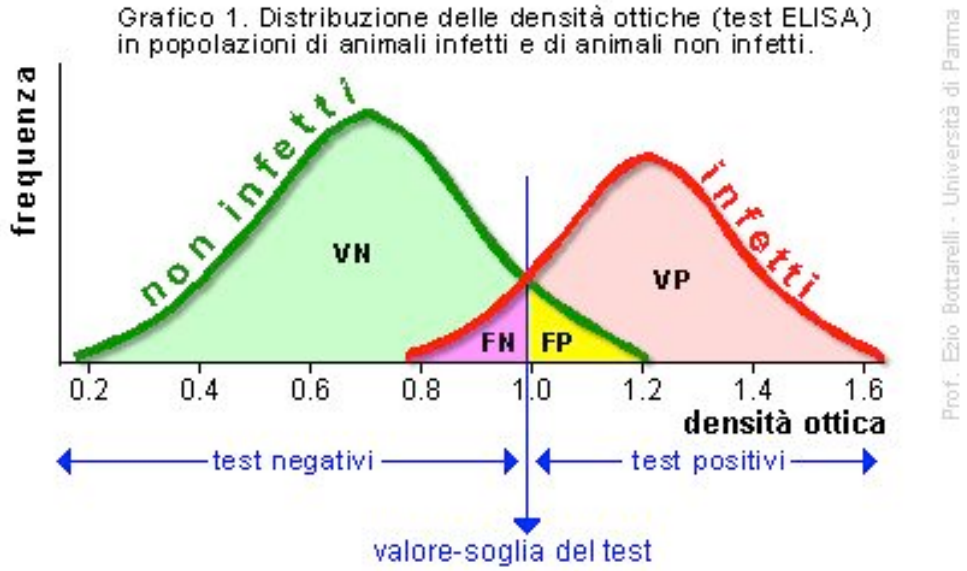
In altri casi, però, i test forniscono risultati classificabili in più di due categorie. Ad esempio, lo stato di un paziente potrebbe essere classificato come: peggiorato, stazionario, poco migliorato, migliorato, molto migliorato. Le variabili di questo tipo, in cui dati qualitativi vengono suddivisi in più categorie con una direzione chiaramente implicita dal migliore al peggiore, vengono dette «ordinali». Ancora, i test possono fornire risultati numerici (variabili «continue») che variano con un continuum, come ad esempio i valori di densità ottica (D.O.) di un test ELISA misurato con lo spettrofotometro. In tutti questi casi occorre stabilire un valore critico o soglia o cut-off, che rappresenta il limite di separazione tra «positività» e «negatività» del test, cosa che - in epidemiologia - corrisponde generalmente alla separazione fra ammalato e sano.

Potrà sorprendere l'affermazione secondo la quale, nel caso dei test non-dicotomici, la sensibilità e la specificità possono essere fatte variare a piacimento. Giustifichiamo questa affermazione, e discutiamone le implicazioni, per mezzo di un esempio.

Nel grafico sottostante sono riportate delle curve ottenute supponendo di saggiare una popolazione costituita da soggetti con una certa infezione e soggetti non infetti. Vediamo come è stato costruito il grafico.

Sull'asse delle ascisse (orizzontale) è stata riportata la densità ottica rilevata saggiando con un test ELISA i sieri dei suddetti soggetti. I valori sono compresi fra 0.2 a 1.6; essi sono proporzionali alla quantità di anticorpi presenti nel siero: più anticorpi ci sono nel siero e maggiore è la D.O. Ovviamente, più la D.O. è elevata , maggiori sono le probabilità che il soggetto abbia di recente subito la infezione.

Sull'asse delle ordinate (verticale) è riportata la frequenza di osservazioni, cioè il numero di soggetti che hanno presentato il titolo corrispondente in ascissa. Si nota che i soggetti non-infetti (la curva più a sinistra) hanno fatto registrare valori di D.O. mediamente più bassi rispetto agli soggetti infetti. Si nota anche che le due curve si sovrappongono parzialmente, ed è proprio questa area di sovrapposizione che verrà presa in considerazione nella discussione che segue.



Non devi stupirti che una certa quota di soggetti sani possa evocare una risposta positiva ad un test: questo fenomeno può essere dovuto ad una varietà di cause che non è possibile trattare in questa sede.

Ora, il problema è quello di stabilire un limite di separazione (soglia o cut-off) fra infetti e sani, ossia di stabilire un titolo (valore di O.D.) al di sopra del quale l'animale viene ritenuto infetto e al di sotto del quale viene ritenuto sano. Per essere più chiari, devi rispondere alla seguente domanda:

a partire da quale valore di D.O. giudichi infetto un animale?

Supponi di adottare come cut-off il valore di 1.0. Ciò significa che dichiarerai come "sano" ogni soggetto con D.O. <=1.0, e dichiarerai "infetto" ogni soggetto con D.O. >1.0.

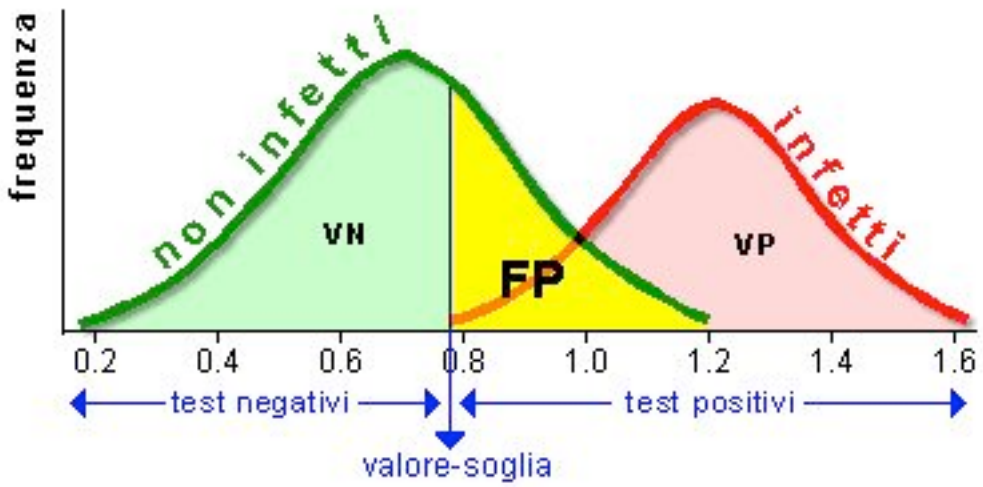
Esaminando ancora il grafico, puoi notare che con un cut-off di 1 i soggetti vengono suddivisi in quattro classi: (1) i veri negativi (VN); (2) i veri positivi (VP); (3) i falsi negativi (FN); (4) i falsi positivi (FP). Falsi negativi perchè in effetti si tratta di soggetti infetti con valori di OD inferiore al valore di cut-off prescelto. Falsi positivi perchè si tratta di soggetti non infetti con valori di OD superiore al valore di cut-off prescelto. In effetti queste quattro classi corrispondono a quelle già viste nell'unità precedente in cui sono state definite la sensibilità e la specificità di un test:

Quanto finora esposto è valido nell'ipotesi di adottare il valore di 1.0 come cut-off. Ma che cosa succede se adotti un cut-off diverso?

Esamina il Grafico successivo in cui il cut-off è stato abbassato a 0.8: sotto questa nuova ipotesi, il test riesce ad individuare TUTTI i soggetti infetti (cioè raggiunge una sensibilità del 100%, essendo il valore della cella C = zero). Quindi, se adotterai un criterio di interpretazione secondo cui tutti i soggetti con siero a D.O. >0.8 sono

dichiarati infetti", avrai il vantaggio di riuscire ad individuare tutti gli infetti, ma al prezzo di includere tra i soggetti positivi un numero considerevole di soggetti sani (quelli appartenenti alla «coda di destra» della distribuzione dei soggetti sani). Questo comporta un abbassamento della specificità.

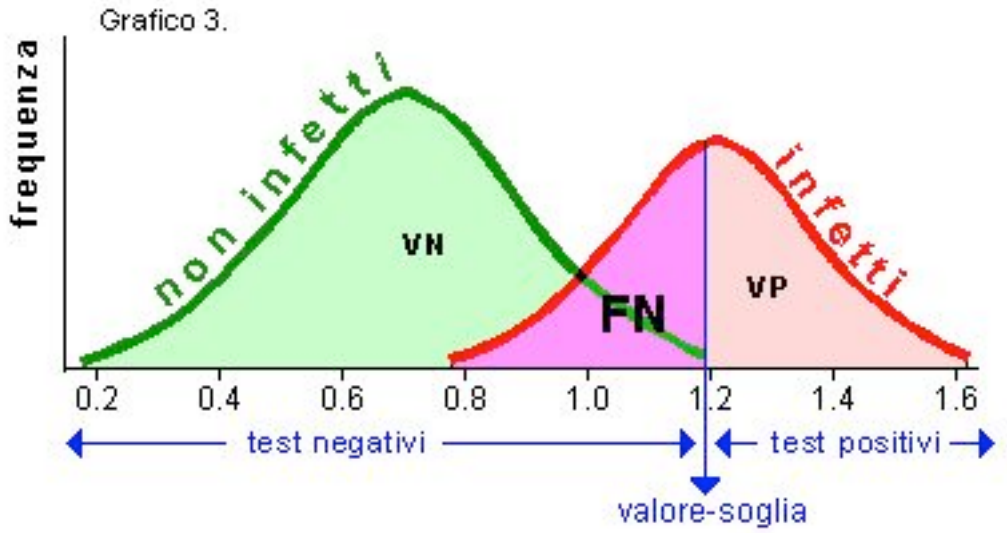
Grafico 2.



Supponi ora di voler adottare un criterio di interpretazione diverso rispetto ai precedenti, ossia un criterio che consenta di individuare con certezza tutti i soggetti sani. In altre parole, vuoi che tutti i soggetti non-infetti vengano classificati come negativi al test, ossia vuoi che il test abbia una specificità del 100%. Osserva il grafico successivo: dovrai scegliere un valore-soglia di D.O. >1.2 ma, come contropartita, classificherai erroneamente come sani molti soggetti infetti, e quindi otterrai un basso valore di sensibilità.

Ma allora, di quale valore di sensibilità e di specificità mi devo fidare ?

In genere è conveniente scegliere una situazione di compromesso, come ad esempio quella indicata nel primo Grafico, scegliendo un valore di cut-off prossimo a 1.0. Questo comporta che sia la sensibilità che la specificità siano inferiori al 100%, e che quindi osserverai una certa proporzione di risultati sia falsi positivi che falsi negativi.



Gli inconvenienti ora accennati derivano dalla inevitabile, parziale sovrapposizione dei valori della variabile misurata dal test nelle due curve di distribuzione (sani e infetti).

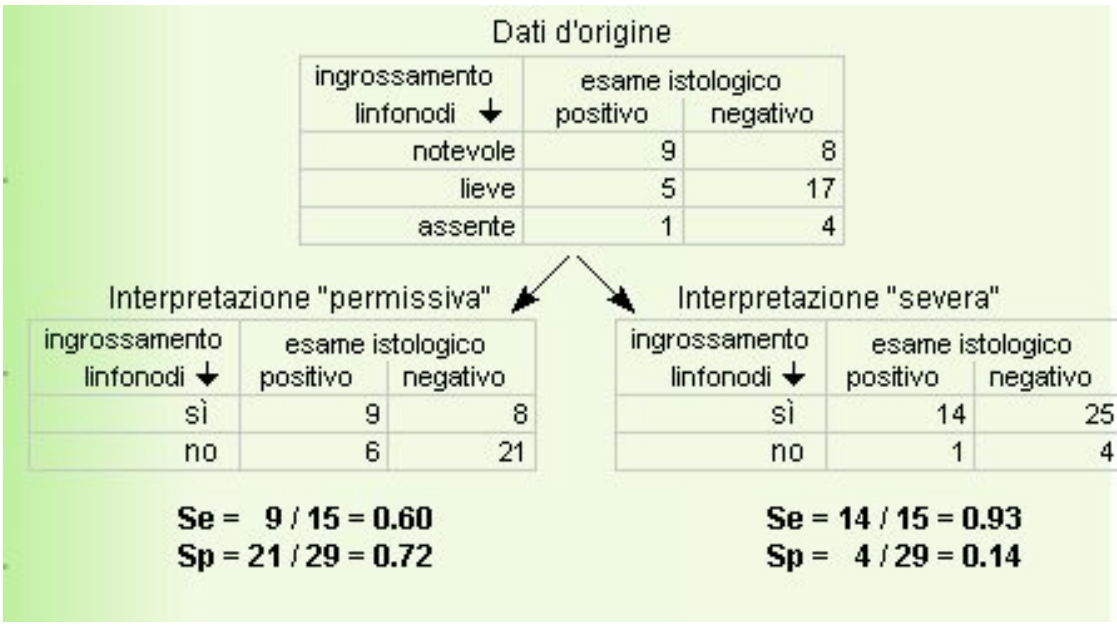
Pertanto, la sensibilità può essere aumentata, ma solo a spese della specificità, e viceversa.

ESEMPIO. Langenbach e coll. (2001) hanno calcolato sensibilità e specificità del test "esame clinico dei linfonodi" di cani e gatti al fine di diagnosticare metastasi di tumori solidi. L'esame clinico è stato posto a raffronto con un test di riferimento (golden test) rappresentato dall'esame istologico dei linfonodi.

È noto che, durante la malattia neoplastica, si possono verificare tumefazioni ed ingrossamenti dei linfonodi regionali, e l'ingrossamento è rilevabile mediante una semplice palpazione. Uno dei problemi connessi con questa tecnica è rappresentato dalla valutazione dell'entità dell'ingrossamento e dalla sua interpretazione. In particolare: è sufficiente un modico ingrossamento oppure l'aumento di volume del linfonodo deve essere notevole?

Gli Autori hanno classificato la modificazione del volume in 3 categorie: (1) no ingrossamento; (2) ingrossamento lieve; (3) ingrossamento notevole. Questo tipo di classificazione ha fatto sorgere il quesito se i soggetti appartenenti alla categoria (2) (ingrossamento moderato) fossero da assegnare alla categoria degli ammalati o dei non ammalati. Gli Autori hanno applicato due criteri di interpretazione: interpretazione permissiva e interpretazione severa. Nel primo caso i soggetti con linfonodi lievemente ingrossati venivano classificati come "sani", nel secondo come "malati". Ciò corrisponde proprio ad una variazione del cut-off.

In sintesi, sono stati ottenuti i seguenti risultati:



Come si vede, l'interpretazione permissiva (che corrisponde ad un innalzamento del cut-off) ha fatto registrare una Se inferiore ed una Sp superiore rispetto alla interpretazione severa (corrispondente ad un abbassamento del cut-off).

Privilegiare la sensibilità o la specificità ?

Purtroppo a questa domanda non può essere data una risposta univoca. Come abbiamo ora dimostrato, il valore di cut-off influenza sia la sensibilità che la specificità del test. Esso viene scelto in base ad una serie di considerazioni attinte da altre informazioni come ad esempio, la storia naturale della malattia, nonché le conseguenze sanitarie ed economiche dei falsi negativi e dei falsi positivi. Nel caso di alcune malattie infettive, talvolta anche un solo soggetto falso negativo può risultare particolarmente pericoloso, in quanto escretore dell'agente di malattia e quindi disseminatore del contagio.

ESEMPIO. Nello screening effettuato sulle persone donatrici di sangue è necessario adottare test provvisti della massima sensibilità. Infatti, é assolutamente indispensabile tutelare chi riceve la donazione e quindi non si può correre il rischio di trasfondere sangue infetto (risultato falsamente negativo ai test di sicurezza). Su questa base, diventa tollerabile la distruzione di una certa quota di campioni non infetti (risultati falsamente positivi ai test di sicurezza).

Anche nel caso di malattie rare conviene utilizzare un test ad alta sensibilità, altrimenti si rischia di non individuare i pochi casi presenti; al contrario, se la prevalenza della malattia è elevata, è generalmente più utile un test altamente specifico: infatti vanno assolutamente contenuti i falsi positivi al fine di non esaurire rapidamente le risorse per le richieste diagnostiche o terapeutiche del gran numero di soggetti positivi (veri e falsi).

Un test sensibile dovrebbe essere scelto quando le conseguenze di una mancata diagnosi sono particolarmente gravi (es. malattie ad esito solitamente mortale, ma che possono essere efficacemente curate).

I test sensibili sono utili anche durante il processo diagnostico iniziale, al fine di ridurre il ventaglio di possibilità (diagnosi differenziale) quando esso è ampio. In tal caso, il test sensibile viene applicato soprattutto allo scopo di escludere una o più malattie. Infatti, un test sensibile è di maggior aiuto al clinico quando fornisce un risultato negativo.

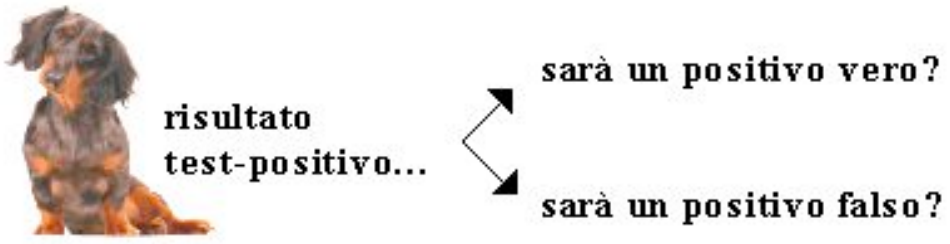
Un test specifico è particolarmente utile per confermare una diagnosi già effettuata con altri mezzi. Infatti, un test specifico raramente è positivo in assenza della malattia. I test altamente specifici sono particolarmente utili quando un risultato falso positivo risulta particolarmente dannoso (sotto l'aspetto organico, emotivo per il proprietario, finanziario ecc.).

Valore predittivo di un test

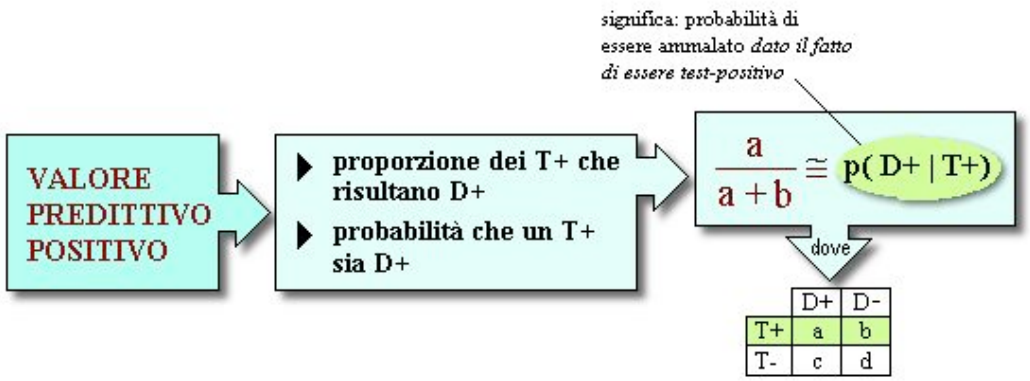
I valori predittivi (positivo e negativo) sono importanti per determinare lo stato di salute del soggetto, una volta effettuato il test. Mentre sensibilità e specificità sono impiegati per la valutazione di un test e quindi sono orientati sul test e danno per scontato la condizione del paziente, VPP e VPN sono impiegati per la valutazione del paziente e quindi sono orientati per il paziente.

Quando si deve interpretare il risultato di un test, occorre valutare quanti soggetti veri positivi (a) saranno presenti nel gruppo dei positivi al test (a+b).

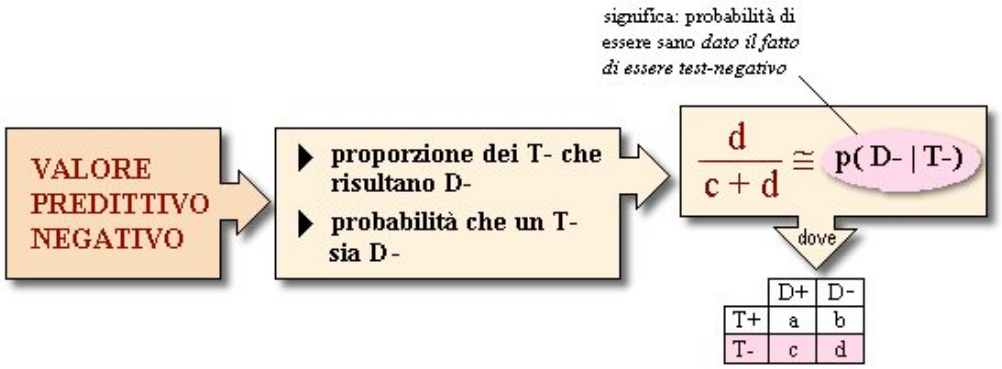
In altre parole, l'operatore sanitario viene chiamato a rispondere - in presenza di un soggetto positivo al test - alla seguente domanda:



Una risposta a questa domanda può venire, in termini di probabilità, dalla conoscenza del valore predittivo positivo (VPP) del test utilizzato. Questo valore indica, infatti, la probabilità che un soggetto test-positivo sia ammalato, e si calcola con la proporzione a/(a+b).



Si può calcolare anche il valore predittivo negativo (VPN) di un test; di fronte ad un soggetto test-negativo, attraverso il valore predittivo negativo (d/c+d) si può rispondere, sempre in termini di probabilità, alla seguente domanda: «si tratta di un negativo-vero o di un negativo-falso?»



Stima di un valore predittivo

Anche per i valori predittivi si possono calcolare gli intervalli di confidenza.

Per il calcolo degli intervalli di confidenza 95% di un dato valore predittivo positivo, si utilizza la formula seguente, in cui *n* corrisponde al totale dei soggetti risultati test-positivi:

$$IC95\% = VPP \pm \sqrt{\frac{VPP * (1 - VPP)}{n}} \quad \text{x } 1.96$$

Per il calcolo dell'intervallo di confidenza del valore predittivo negativo, basta sostituire VPP con VPN e *n* con il totale dei soggetti test-negativi.

	M+	M-
T+	36	6
T-	4	49

ESEMPIO. Supponiamo che tu abbia effettuato uno screening applicando un test ad un campione composto da 95 soggetti, ottenendo i risultati riassunti nella Tabella a lato. [in realtà, nella pratica, dopo lo screening, non otterrai i 4 valori della tabella, ma soltanto i seguenti due dati: test-positivi n=42; test-negativi n=53]. I calcoli sono come segue:

$$\mathbf{VPP} = 36/(36+6) = 0.857$$

$$\mathbf{IC95\%} = \pm 1.96\sqrt{\frac{0.857 * 0.143}{42}} \rightarrow 0.751 \text{ , } 0.963$$

Se vuoi ottenere l'IC99% (anzichè 95%), basta sostituire il coefficiente 1.96 con 2.54.

Da cosa dipendono i valori predittivi?

Il valore predittivo positivo dipende, come è lecito attendersi, dalla Se e dalla Sp del test; in particolare, esso aumenta con l'aumentare di questi due parametri. É però importante sottolineare un altro aspetto più sorprendente, e particolarmente importante nella pratica: *il valore predittivo positivo dipende anche da un fattore indipendente dal test*: la *prevalenza* della malattia nella popolazione.

Valore predittivo in epidemiologia clinica Una volta che il test è stato effettuato ed ha fornito un risultato (non importa se positivo o negativo), la Se e la Sp perdono importanza. Infatti, Se e Sp si riferiscono a individui il cui stato di salute/malattia è noto. Ma se si conoscesse lo stato del paziente, non sarebbe necessario effettuare alcun test ! Ecco quindi che, nell'attività del clinico, l'obiettivo è il seguente: determinare lo stato del paziente, dato il risultato di un test. In questa ottica, sono importanti i valori predittivi (negativo e positivo).

Relazione tra valori predittivi e prevalenza

Fra il valore predittivo (VPP) e negativo (VPN) di un test e la prevalenza della malattia nella popolazione che viene sottoposta a screening esiste una relazione molto importante.

Supponi di effettuare uno screening nei confronti della brucellosi bovina in 3 diverse aree geografiche (AreaA, AreaB, AreaC), usando un test con sensibilità e specificità note. In ciascuna area sono presenti 30000 animali da sottoporre a screening. Supponiamo di conoscere la prevalenza reale in ciascuna area:

AreaA: prevalenza = 0.1(10%);

AreaB: prevalenza = 0.01 (1%);

AreaC: prevalenza = 0.001 (0.1%).

Lo screening viene effettuato con test di «Agglutinazione rapida al Rosa bengala»*(Rose Bengale Test*, con Se=0.620 e Sp=0.995 [Gall D. & Nielsen K., Rev. sci. tech. Off. int Epiz., 2004, 23 (3), 989-1002].

Le 3 Tabelle sottostanti riassumono i risultati ottenuti nello scenario ora descritto (attenzione: c'è un errore nella prima tabella relativa a Area 1. 1860 è diviso e non addizionato a 1860 + 135. Il risultato è corretto).

Area1

	M+	M-
T+	1860	135
T-	1140	26865

VPP=
1860+(1860+135)=
0.932

Area2

	M+	M-
T+	186	149
T-	114	29552

VPP=
186/(186+149)=
0.556

Area3

	M+	M-
T+	19	150
T-	11	2890

VPP=
19/(19+150)=
0.110

Il valore predittivo positivo

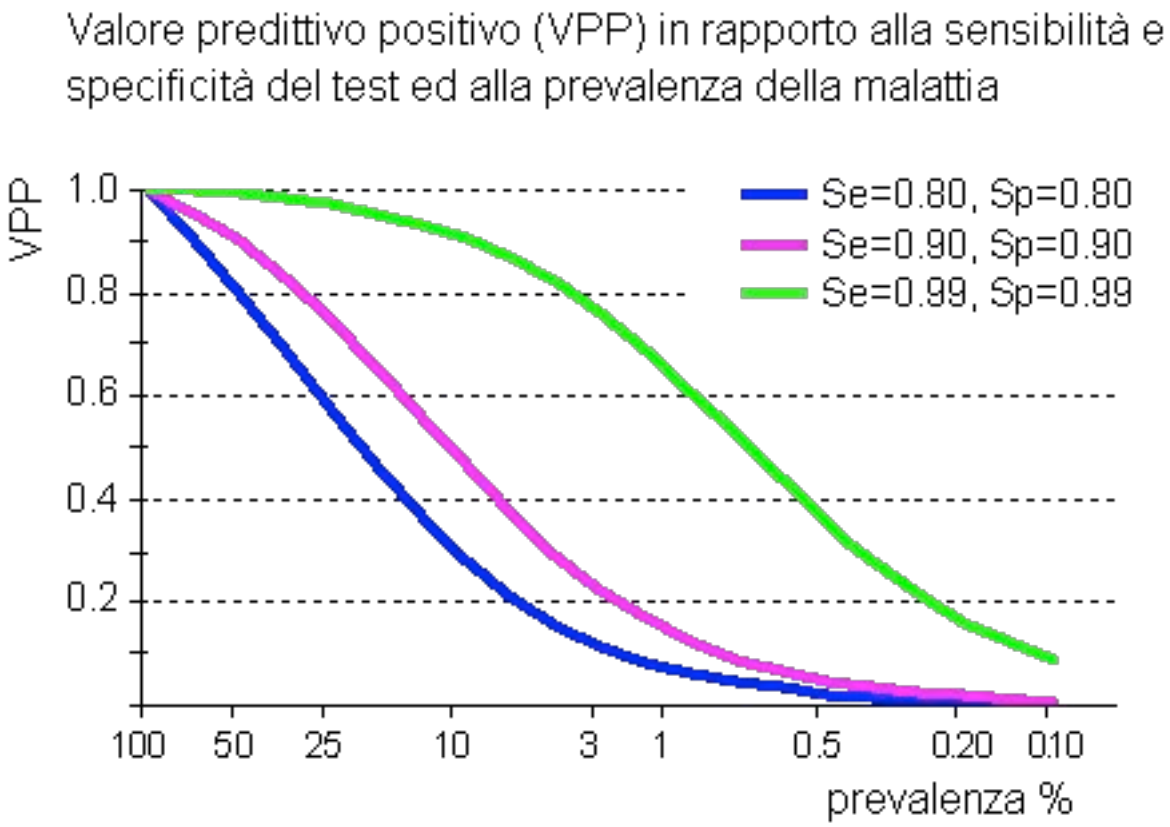
Il valore predittivo positivo (VPP) diminuisce con il diminuire della prevalenza della malattia. Questo accade perché la diminuzione della prevalenza comporta l'incremento degli animali sani; ciò, a sua volta, fa sí che aumenti il numero di esiti positivi falsi.

Nella **Area1**, con prevalenza elevata, il VPP è di 0.932, ciò significa che su 100 bovini positivi al test, 93 sono ammalati, mentre 7 sono positivi falsi. Considera che la profilassi della brucellosi avviene per *eradicazione*, ed i bovini che risultano infetti devono essere abbattuti; *se ci si basasse soltanto sull'esito di questo test di screening*, nello scenario in questione si pagherebbe una sorta di «tassa a fondo perduto» del 7%, rappresentata dai bovini sani da abbattere erroneamente in quanto considerati infetti. L'entità (7%) di questo effetto collaterale sembra accettabile.

Consideriamo l'**Area2** e l'**Area3**, nelle quali il VPP è rispettivamente 0.556 e 0.110. Ciò comporta che il 44.4% e l'89.0% degli abbattimenti sarebbe ingiustificato, riguardando animali sani ma test-positivi. Questa situazione risulterebbe inaccettabile nella pratica per una serie di motivi, che possono essere riassunti in uno solo fondamentale: un eccessivo rapporto costi / benefici (ossia il rapporto fra il costo delle azioni di profilassi ed i benefici indotti da tali azioni).

Pertanto, in questi casi dovranno essere adottate misure correttive per *migliorare il VPP*. Ad esempio, si potranno utilizzare due test invece di uno solo.

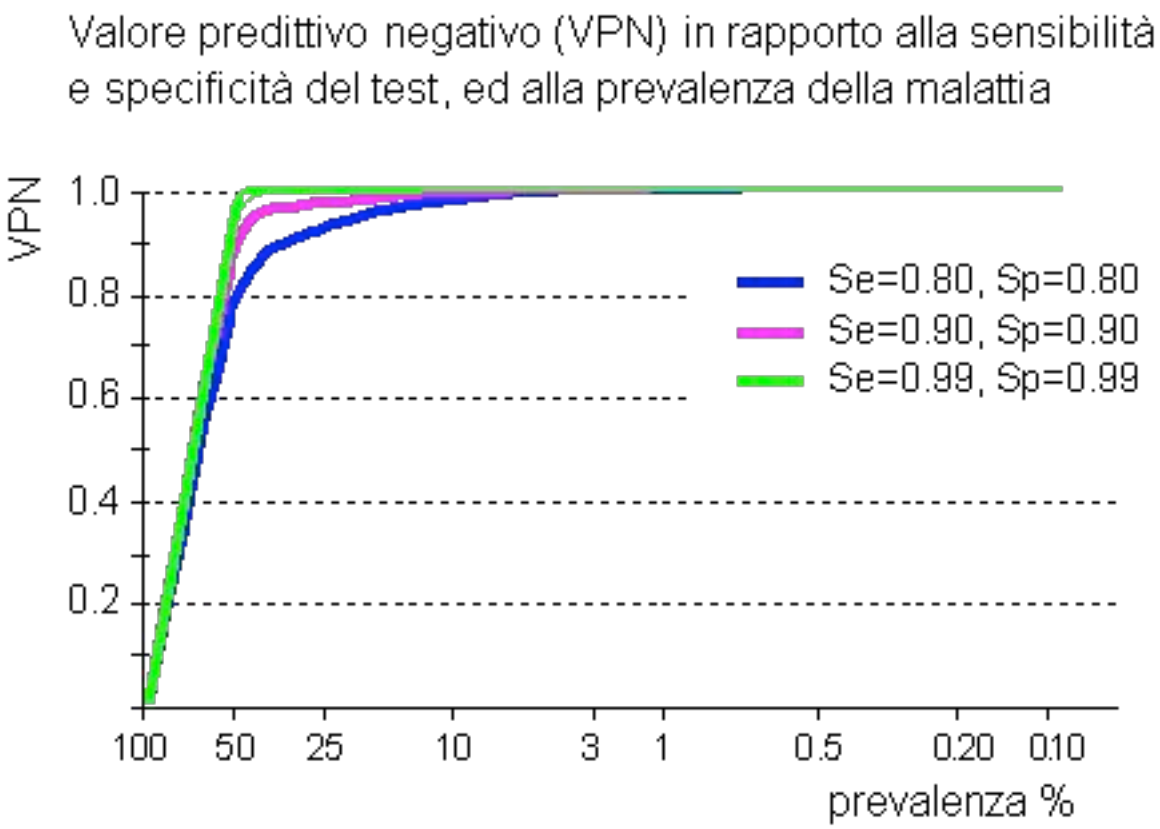
Il **grafico** sottostante illustra l'andamento del valore predittivo positivo in rapporto alla prevalenza, per tre test di esempio a diversa sensibilità e specificità. È evidente che, quando la prevalenza della malattia nella popolazione è elevata, la performance di tutti i test è buona. Invece, per valori di prevalenza molto bassi il valore predittivo di tutti i test si avvicina a zero; in queste condizioni, qualsiasi test diagnostico diventa virtualmente inutile a scopo diagnostico. Puoi notare, confrontando l'andamento delle 3 curve, che l'effetto della prevalenza sul valore predittivo è proporzionale al decrescere della sensibilità e specificità del test.



Il fatto che il valore predittivo positivo dipenda dalla prevalenza sconsiglia azioni di screening per malattie rare. Infatti, uno screening per una malattia rara presenta i seguenti inconvenienti: (1) pochi individui ne trarranno beneficio (proprio in quanto malattia rara); (2) molti individui (i falsi-positivi) ne trarranno un danno, in quanto verranno ingiustamente considerati ammalati.

Il valore predittivo negativo

Analogamente al VPP, anche il VPN dipende dalla prevalenza della malattia nella popolazione. La relazione va in senso opposto rispetto a quanto hai visto per il VPP. Infatti, il VPN aumenta con il diminuire della prevalenza, come schematizzato nel grafico sottostante.



VPN: il rischio di importare un animale ammalato

Quando si acquista un animale (o un prodotto di origine animale), è buona norma accertarsi che esso non sia affetto da malattie trasmissibili o contaminazioni che, in tal modo, potrebbero essere introdotte in un allevamento indenne (ossia nel quale l'agente della malattia non è presente). Di solito prima dell'acquisto l'animale viene sottoposto ad un test. Tuttavia, hai imparato che i test non sono infallibili. Perciò sorge legittima la domanda dell'allevatore:

"se l'animale da acquistare è *test-negativo*, che probabilità ci sono che esso sia *ammalato*?"

Si può rispondere semplicemente che la probabilità è pari a (1-VPN). Infatti, ti ricordo che la probabilità si esprime con un numero compreso fra 0 (l'evento non si verifica mai) e 1 (l'evento si verifica sempre). Come ben ricordi, il VPN rappresenta la probabilità dell'evento "l'animale test-negativo è sano" di conseguenza (1-VPN) rappresenta la probabilità dell'evento alternativo, ossia

"l'animale test-negativo è ammalato". Il punto importante è che, come hai appena visto, il VPN (ed anche il VPP) sono correlati alla sensibilità ed alla specificità del test, ma dipendono anche dalla prevalenza della malattia nella popolazione. Per questo motivo, anche conoscendo la sensibilità e la specificità del test, non è possibile rispondere direttamente alla domanda dell'allevatore, a meno di non conoscere (o stimare) la prevalenza.

In tal caso, si può applicare il teorema di Bayes.

L' utilizzo del Teorema di Bayes è abbastanza semplice: basta applicare la formula appropriata, ossia quella che consente di ottenere, conoscendo Se e Sp del test, la prevalenza reale a partire dalla prevalenza apparente.

Una volta nota la prevalenza reale, sarà facile risalire al valore predittivo positivo e negativo (vedi Bayes.docx).

Teorema di Bayes

Descrive matematicamente la relazione tra due eventi.

Specificamente, il teorema calcola la probabilità che si presenti l'evento B quando avviene l'evento A.
Per esempio l'evento A è il risultato positivo di un test e l'evento B è che si abbia la malattia.
Quindi la domanda è: quale è la probabilità che si abbia la malattia quando il risultato del test è positivo ?

La p di avere la malattia quando il risultato del test è positivo =
il VALORE PREDITTIVO di un test (positivo)

VP (+) = TP / TP + FP

VP (-) = TN / TN + FN

Formula bayesiana in cui viene messo in risalto la influenza della prevalenza della malattia:

$$1\text{-}VPN = \frac{P * (1-Se)}{P * (1-Se) + [(1-P) * Sp]}$$

probabilità che un animale test-negativo sia ammalato

dove: P = prevalenza
Se = sensibilità del test
Sp = specificità del test

ESEMPIO. Un bovino appartenente ad un gruppo in cui si stima che la prevalenza della brucellosi sia pari a 0.20 viene sottoposto con esito negativo al test di agglutinazione rapida al Rosa bengala (Se=0.620 e Sp=0.995). Ci si domanda qual è: la probabilità quell'animale sia ammalato.

Applicando la formula soprastante ottieni:

1 - VPN = ((0.2*0.380) / [(0.2*.380)+ (0.8*0.995)] = 0.0872 (8.72%)

Applicazione del Teorema di Bayes

Il Teorema di Bayes è uno strumento per investigare, in termini di probabilità, le interazioni tra due o più variabili.

In epidemiologia, viene usato con una certa frequenza nell'interpretazione dei risultati dei test di screening, allo scopo di ottenere maggiori informazioni di quelle fornite dai semplici dati grezzi.

Sotto questo aspetto, il teorema di Bayes può essere visto come uno strumento che "crea conoscenza"

LA SIMULAZIONE

Sei il responsabile del piano di eradicazione della brucellosi bovina in un territorio

Insieme al tuo staff, hai sottoposto al test di "Agglutinazione rapida al Rosa bengala" (*Rose Bengale Test*, RBT) tutti i bovini adulti del territorio

Il test ha le seguenti caratteristiche:

Sensibilità (Se) = 0.812

Specificità (Sp) = 0.862

(rif. bibliografico: Gall D. & Nielsen K., Rev. sci. tech. Off. int Epiz., 2004, 23 (3), 989-1002)

Hai ottenuto i seguenti risultati:

Totale bovini saggiati n=4820

di cui

Bovini test-positivi n= 932

Bovini test-negativi n=3888

In base a questi risultati, puoi calcolare la prevalenza dell'infezione come segue:

Pr = 932 / 4820 = 0.1934 (19.34%)

RISPOSTA

Si tratta di una prevalenza **apparente**, in quanto calcolata in base all'esito del test RBP, che non è infallibile. Ciò significa che:

- fra i 932 animali test-positivi è compreso un certo numero di animali sani (falsi positivi)
- fra i 3888 animali test-negativi è compreso un certo numero di animali malati (falsi negativi)

Ovviamente non ti accontenti delle... ***apparenze***, ma vuoi accertare quanto segue, *a partire dai soli dati che hai a disposizione*:

- la prevalenza reale
- il numero di animali falsi positivi
- il numero di animali falsi negativi
- il valore predittivo positivo (VPP)
- il valore predittivo negativo (VPN)

Spero che ti renda conto dell'importanza del raggiungimento di questi obiettivi....

Riassumiamo i dati di cui disponi:

DATI RACCOLTI NELLO SCREENING:

- **Totale bovini saggiati** **n=4820**
- **Bovini test-positivi** **n=932**
- **Bovini test-negativi** **n=3888**

DATI NOTI IN PARTENZA:

- **Sensibilità del test (Se)** **0.812**
- **Specificità del test (Sp)** **0.862**

Ti conviene distribuire i dati raccolti nello screening nella "solita" tabella di contingenza, come segue:

	Malati M+	Sani M-	TOT
Test positivi T+	a?	b?	932
Test negativi T-	c?	d?	3888
TOT	a+c?	b+d?	4820

Tieni sempre alla mano la tabella di contingenza, che trovi duplicata qui:

	Malati M+	Sani M-	TOT
Test positivi T+	a	b	932
Test negativi T-	c	d	3888
TOT	a+c	b+d	4820

Con i dati della Tabella, si tratta di trovare il valore delle 4 celle **a**, **b**, **c**, **d**, sapendo che Se=0.812 e Sp=0.862

Una volta noto il valore di a, b, c, d sarà facile calcolare la prevalenza reale ed i valori predittivi

Ti sembra un compito impossibile?

Chiedi aiuto al Teorema di Bayes!

Per... tua fortuna, qui non ti spiego il teorema di Bayes: puoi considerarlo una sorta di "scatola nera" che ti permette di raggiungere l'obiettivo che ti interessa.

Però, poi ti illustrerò un sistema con cui puoi raggiungere *da solo* lo stesso obiettivo, utilizzando le tue conoscenze di algebra!

Ecco qui, fra le svariate formule che possono essere derivate dal Teorema di Bayes, quella utile nel tuo caso:

$$P_{reale} = \frac{P_{app} + Sp - 1}{Se + Sp - 1}$$

Ecco i dati che ti servono:

Se =	**0.812**
Sp =	**0.862**
Prevalenza apparente =	**0.1934**

$$P_{reale} = \frac{P_{app} + Sp - 1}{Se + Sp - 1}$$

$$P_{reale} = \frac{0.1934 + 0.862 - 1}{0.812 + 0.862 - 1} = 0.082$$

Attraverso l'applicazione del teorema di Bayes hai ottenuto un risultato molto importante. Infatti:

la prevalenza dell'infezione nella tua popolazione bovina, che stimavi essere pari a 0.1984 (19.84%), in realtà è invece molto più bassa: 0.082 (8.2%).

Procedi a calcolare il valore predittivo positivo ed il valore predittivo negativo raggiunti dal test nelle tue condizioni. Ora che conosci la prevalenza reale, questo calcolo è molto facile...

La prevalenza reale è 0.0821. Guarda la Tabella: la prevalenza reale si calcola come: (a+c)/(a+b+c+d).
Quindi
a+c = 4280 * 0.0821 = 396

	Malati M+	Sani M-	TOT
Test positivi T+	a?	b?	932
Test negativi T-	c?	d?	3888
TOT	396	b+d?	4820

E, per differenza b+d=4424:

	Malati M+	Sani M-	TOT
Test positivi T+	a?	b?	932
Test negativi T-	c?	d?	3888
TOT	396	4424	4820

	Malati M+	Sani M-	TOT
Test positivi T+	a?	b?	932
Test negativi T-	c?	d?	3888
TOT	396	4424	4820

Poichè Se = a/(a+c) = 0.812,
a = 396 * 0.812 = 322
e, per differenza,
c = 74

Poichè Sp = d/(b+d) = 0.862,
d = 4424 * 0.862 = 3813
e, per differenza,
b = 611

HAI COMPLETATO LA TABELLA!

	Malati M+	Sani M-	TOT
Test positivi T+	322	611	932
Test negativi T-	74	3813	3888
TOT	396	4424	4820

	Malati M+	Sani M-	TOT
Test positivi T+	322	611	932
Test negativi T-	74	3813	3888
TOT	396	4424	4820

Infine, calcola i valori predittivi:

VPP = 322 / 932 = 0.345

VPN = 3813/3888 = 0.981

Un ultimo commento: il VPP dello screening è molto basso: soltanto il 34.5% degli animali positivi è effettivamente ammalato. L'efficienza dello screening è compromessa.

Ed ora, risolverai lo stesso problema con un po' di algebra, e senza l'ausilio del teorema di Bayes. Parti di nuovo dalla solita Tabella:

	Malati M+	Sani M-	TOT
Test positivi T+	a	b	932
Test negativi T-	c	d	3888
TOT	a+c	b+d	4820

Si tratta di trovare il valore delle 4 incognite a, b, c, d sapendo che:

a+b= 932
c+d= 388
a/(a+c)= 0.812 (questa è la Se)
d/(b+d)= 0.862

	Malati M+	Sani M-	TOT
Test positivi T+	a?	b?	932
Test negativi T-	c?	d?	3888
TOT	a+c?	b+d?	4820

I tuoi dati di partenza sono:

a+b=932 (n. test-positivi)
c+d=3888 (n. test-negativi)
a/(a+c)=0.812 (sensibilità)
d/(b+d)=0.862 (specificità)

	Malati M+	Sani M-	TOT
Test positivi T+	a?	b?	932
Test negativi T-	c?	d?	3888
TOT	a+c?	b+d?	4820

I tuoi dati di partenza sono:

{ a+b=932 (n. test-positivi)
c+d=3888 (n. test-negativi)
a/(a+c)=0.812 (sensibilità)
d/(b+d)=0.862 (specificità)

E' un sistema
a quattro incognite!
Risolviamolo insieme...

$$\begin{cases} a+b=932 \\ c+d=3888 \\ a/(a+c)=0.812 \\ d/(b+d)=0.862 \end{cases}$$

$$\begin{cases} a+b=932 \\ c+d=3888 \\ a=0.812(a+c) \\ d=0.862(b+d) \end{cases}$$

$$\begin{cases} a+b=932 \\ c+d=3888 \\ a-0.812a+0.812c \\ d=0.826b+0.826d \end{cases}$$

$$\begin{cases} a+b=932 \\ c+d=3888 \\ a-0.812a=0.812c \\ d-0.826d=0.826b \end{cases}$$

$$\begin{cases} a+b=932 \\ c+d=3888 \\ 0.188a=0.812c \\ 0.138d=0.862b \end{cases}$$

$$\begin{cases} a+b=932 \\ c+d=3888 \\ a=4.319c \\ d=6.246b \end{cases}$$

$$\begin{cases} 4.319c+b=932 \\ c+6.246b=3888 \\ a=4.319c \\ d=6.246b \end{cases}$$

$$\begin{cases} b=932-419c \\ c+6.246(932-4.319c)=3888 \\ a=4.319c \\ d=6.246b \end{cases}$$

b=932-4.319c
c=5821.272-26.976c=3888
a=4.319c
d=6.246b

b=932-4.319c
c-26.976=3888-5821.272
a=4.319c
d=6.246b

b=932-4.319c
-25.976c=1933.272
a=4.319c
d=6.246b

b=932-4.319c
c=-1933.272/-25.976=74
a=4.319c
d=6.246b

{
b=932-4.319*74.425=611
c=74
a=4.319*74.425=321
d=6.246*610.558=3813

... et voila! :_)))

Metodi per migliorare il valore predittivo di un test di *screening*

Abbiamo visto che il valore predittivo positivo (VPP) è un elemento di fondamentale importanza per la riuscita delle operazioni di screening. Abbiamo visto anche che il VPP è correlato alla prevalenza: esso decresce al diminuire della prevalenza, potendo raggiungere livelli inaccettabilmente bassi e tali da compromettere l'efficienza dell'azione di screening.

Come migliorare il VPP?

Il primo metodo per ottenere un accettabile VPP è quello di operare su popolazioni ad alto rischio, nelle quali la prevalenza si presume assuma valori elevati. Se questa opzione non è praticabile, allora si può cercare di individuare, nella popolazione, sottogruppi ad alto rischio sui quali concentrare il programma di screening.

Fra gli interventi di diagnosi precoce dei tumori previsti dal Sistema Sanitario Nazionale, si annovera la mammografia. Il test viene effettuato allo scopo di individuare tempestivamente i tumori della mammella e prevede l'esame biennale delle donne di età compresa fra 50 e 69 anni. Questa fascia di età è più a rischio rispetto alle altre ed in essa l'incidenza è maggiore. Perciò il valore predittivo del test è più elevato rispetto a quanto si avrebbe esaminando indiscriminatamente tutte le donne.

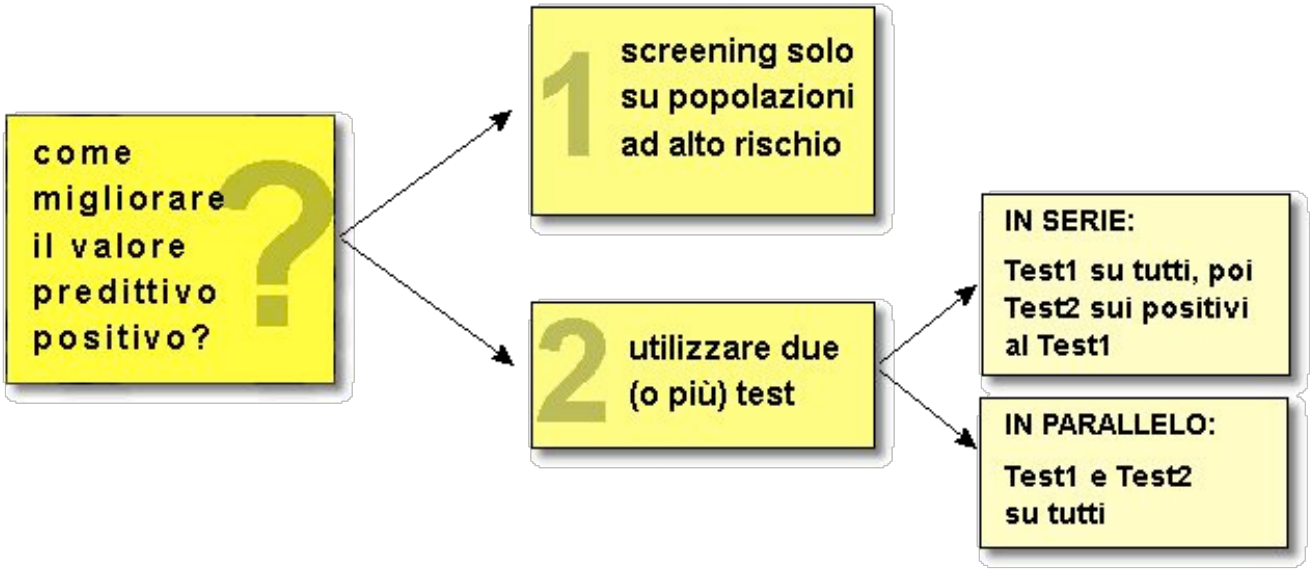
Il secondo metodo per migliorare il VPP è quello di utilizzare due (o, raramente, più di due) test. Operativamente, ciò può avvenire con due diverse modalità:

(1) *in serie*, cioè prima un test e poi, *su quelli risultati positivi*, l'altro;

(2) *in parallelo* su tutti gli animali.

L'interpretazione dei risultati di test in serie è ovvia: si considerano ammalati gli animali risultati positivi al primo ed al secondo test. Più complicata è l'interpretazione dei test in parallelo.

Riassumendo:



Test multipli: utilizzo di 2 test in serie

Per capire meglio il funzionamento di questa strategia di screening, ci serviremo di un esempio in cui i due test in serie vengono utilizzati su una popolazione di 8000 animali. Supponiamo anche, ma solo allo scopo di rendere l'esempio comprensibile, di conoscere lo stato reale (malato/sano) di ciascuno di questi 8000 animali. In particolare, sappiamo che 111 di essi sono ammalati e 7889 sono sani.

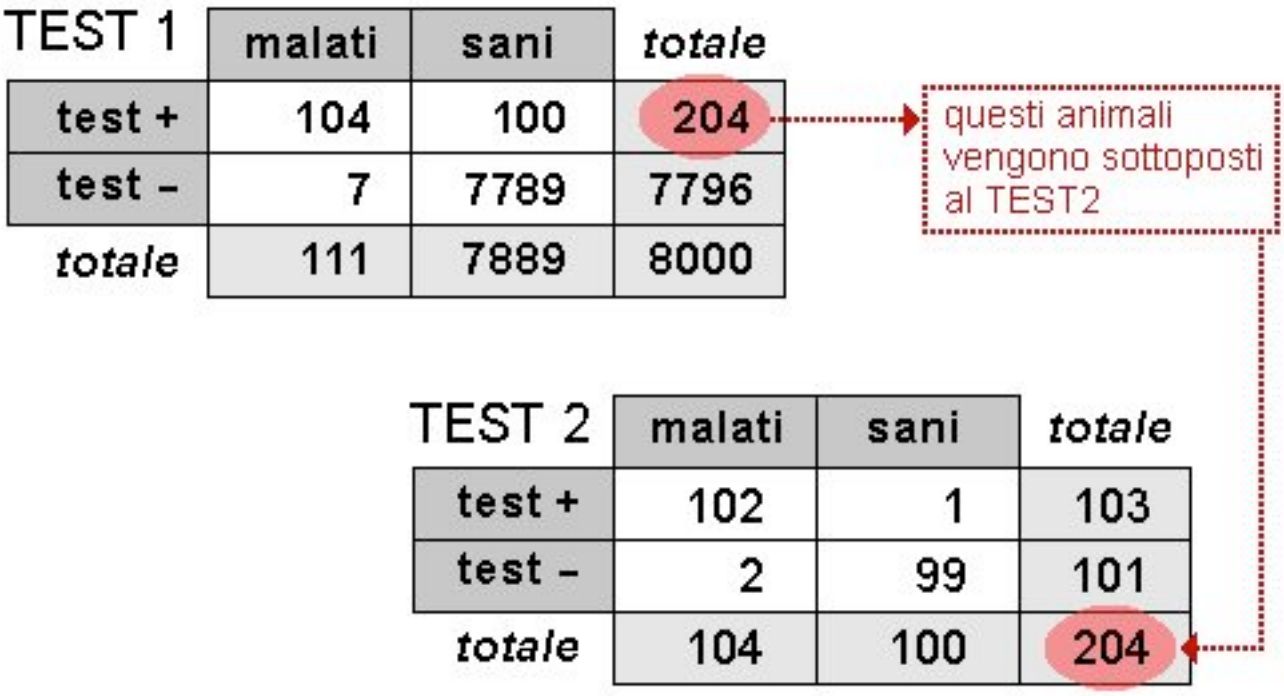
I due test hanno le seguenti caratteristiche:

- TEST1: Se=0.937, Sp 0.987
- TEST2: Se=0.981, Sp=0.990

TEST2 è più costoso e più sensibile del TEST1.

Sottoponi al TEST1 tutti gli 8000 soggetti della popolazione, e poi applichi il TEST2 *solo su quelli risultati positivi al TEST1*.

Otterrai questi risultati:



Se vuoi, puoi fare il seguente breve "ripasso" verificando che i risultati tabulati soddisfano i dati forniti: TEST1: Se = 104/111 = 0.937; Sp=7789/7889 = 0.987; TEST2: Se = 102/104 = 0.981; Sp = 99/100 = 0.990

Come vedi, alla fine del procedimento di applicazione in serie del TEST1 e del TEST2, gli 8000 animali sono stati classificati come segue: - 103 positivi (di cui 102 positivi veri e 1 positivo falso) - 7897 negativi (7796 negativi al TEST1 + 101 negativi al TEST2). Fra questi 7987 negativi, 7888 sono negativi veri e 7+2 negativi falsi.

Ora puoi calcolare Se, Sp e VPP complessive, ossia quelle ottenute utilizzando i due test in serie:

$$\text{SENSIBILITA'} = \frac{\text{positivi veri TEST2}}{\text{ammalati}} = \frac{102}{111} = 0.919$$

TEST 1

	malati	sani	totale
test +	104	100	204
test –	7	7789	7796
totale	111	7889	8000

TEST 2

	malati	sani	totale
test +	102	1	103
test –	2	99	101
totale	104	100	204

$$\text{SPECIFICITA'} = \frac{\text{neg.veri TEST1} + \text{neg.veri TEST2}}{\text{sani}} = \frac{7789 + 99}{7889} = 0.999$$

TEST 1

	malati	sani	totale
test +	104	100	204
test –	7	7789	7796
totale	111	7889	8000

TEST 2

	malati	sani	totale
test +	102	1	103
test –	2	99	101
totale	104	100	204

$$\text{VALORE PREDITTIVO POSITIVO} = \frac{\text{positivi veri TEST2}}{\text{totale positivi TEST2}} = \frac{102}{103} = 0.990$$

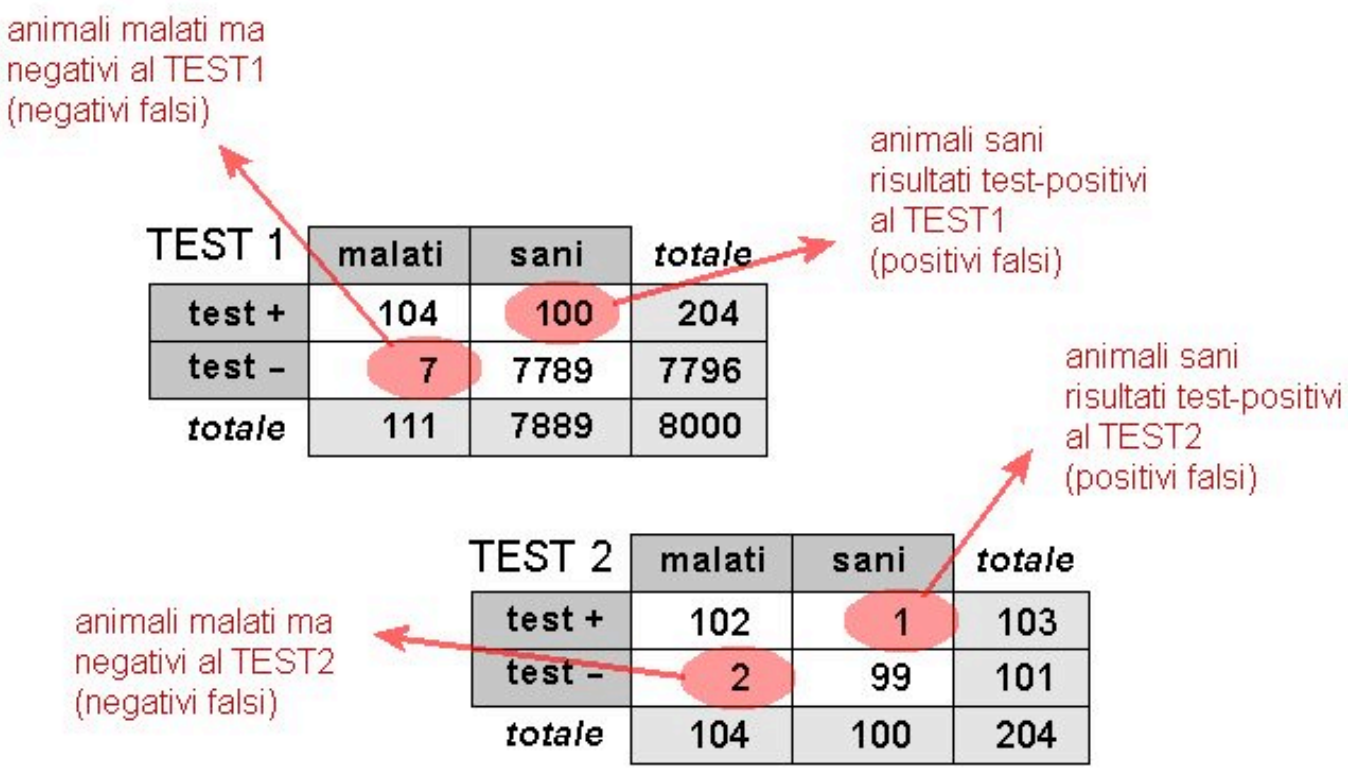
TEST 1

	malati	sani	totale
test +	104	100	204
test –	7	7789	7796
totale	111	7889	8000

TEST 2

	malati	sani	totale
test +	102	1	103
test –	2	99	101
totale	104	100	204

Nella Tabella che segue sono evidenziati in colore gli animali classificati incorrettamente (mis-classificati). Puoi vedere che: - al TEST1 hai classificato come sani 7 animali che in realtà sono malati; questi 7 animali non verranno più testati; - al TEST1 hai classificato come malati 100 animali in realtà sani, che però verranno poi saggiati con il TEST2; - al TEST2 hai classificato come malato 1 animale che in realtà è sano; - al TEST1 hai classificato come sani 2 animali che in realtà sono malati.



Rispetto agli esiti del TEST1, al TEST2 sono stati dichiarati sani 2 animali che in realtà sono malati. Questo effetto negativo derivante dall'applicazione del TEST2 è largamente compensato dal fatto che con il secondo test si sono quasi annullati i positivi falsi, che passano da 100 a 1.

Proviamo a fare qualche conto sui benefici della strategia «Due test in serie» utilizzando i dati dello scenario che hai appena visto, ed ipotizzando che un esame di laboratorio con il TEST1 costi 1€ e con TEST2 costi 10€: (a) se utilizzi soltanto TEST1, spendi 8000€ ed ottieni 107 mis-classificazioni: 100 positivi falsi e 7 negativi falsi); (b) se utilizzi soltanto TEST2, spendi 80000€ ed ottieni 81 mis-classificazioni: 79 positivi falsi e 2 negativi falsi (dati non mostrati nelle tabelle); (c) se utilizzi i due test in serie, spendi 10040€ ed ottieni 10 mis-classificazioni: 1 positivo falso e 9 negativi falsi.

Ricordati che ti puoi attendere risultati simili a quelli dell'esempio soltanto se i due test sono **biologicamente indipendenti** l'uno dall'altro. Per «biologicamente indipendenti» si intende che i test sono basati su meccanismi diversi o, meglio, che misurano grandezze diverse (es. differenti classi di anticorpi).

L'utilizzo di due test in serie è previsto, ad esempio, nell'ambito del piano di profilassi di Stato della brucellosi bovina, con l'impiego di due test sierologici: il test al rosa-bengala in prima istanza e quindi, sui positivi, la fissazione del complemento. Questi due test possono essere ritenuti «biologicamente indipendenti», in quanto misurano classi diverse di anticorpi, e quindi il loro impiego in serie risulta efficace.

Se i due test sono biologicamente dipendenti, allora si otterranno probabilmente risultati meno brillanti di quelli dell'esempio. Infatti, in test biologicamente simili i risultati tendono ad essere correlati, nel senso che aumenta la probabilità che essi forniscano lo stesso risultato quando applicati allo stesso animale.

Attenzione a non commettere l'errore di confrontare il valore predittivo positivo del primo test con quello del secondo. Infatti nel nostro esempio:

VPP TEST1 = 104/204 = 0.509

VPP TEST2 = 102/103 = 0.990

Apparentemente il TEST2 è di gran lunga superiore rispetto al TEST1; tuttavia, il confronto è viziato. Infatti occorre ricordare che il VPP dipende (oltre che dalla specificità e sensibilità intrinseche del test), anche dalla prevalenza; in questo caso il TEST2 è stato applicato su un gruppo di animali già positivi al TEST1 test e nei quali, perciò, la prevalenza era molto elevata, come risulta dal seguente calcolo.

Popolazione sottoposta al TEST1:

- Prevalenza reale = 111/8000 = 0.013

- Prevalenza apparente (in base al test) = 204/8000 = 0.025

Popolazione sottoposta al TEST2:

- Prevalenza reale = 104/204 = 0.509

- Prevalenza apparente (in base al test) = 103/204 = 0.505

La scelta dell'ordine di serie dei due test (prima TEST1 poi TEST2, oppure viceversa prima TEST2 poi TEST1) viene effettuata tenendo conto soprattutto dei costi e della praticità di esecuzione dei test. Infatti, è preferibile che il primo test (quello applicato su un numero maggiore di individui) sia il meno costoso e/o quello di più facile esecuzione oppure meno invasivo per il paziente.

Test multipli: utilizzo di 2 test in parallelo

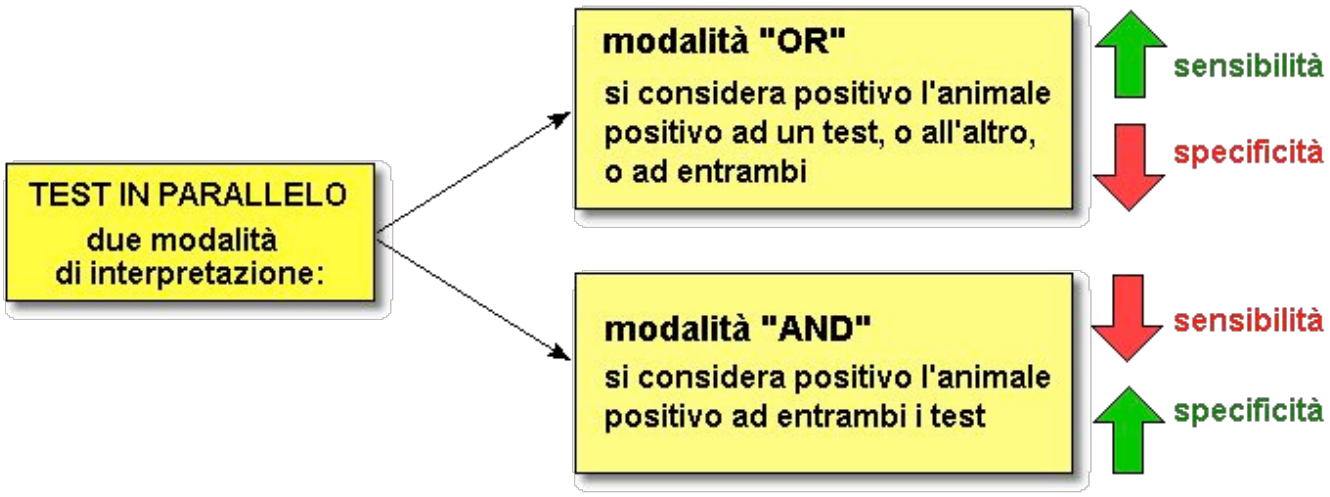
Oltre all'utilizzo di test multipli *in serie*, un altra modalità di impiego di test multipli è quella di applicare 2 (o più) test **contemporaneamente** agli animali della popolazione da saggiare. La sensibilità e la specificità della combinazione di test dipendono dalla modalità di interpretazione dei risultati. Infatti, potranno ottenere le seguenti combinazioni di risultati:

- TEST1 positivo e TEST2 negativo (T1+/T2-)
- TEST1 negativo e TEST2 positivo (T1-/T2+)

- TEST1 positivo e TEST2 positivo (T1+/T2+)
- TEST1 negativo e TEST2 negativo (T1-/T2-)

Applicando due test contemporaneamente sorge un problema di interpretazione dei risultati. Infatti, è pacifico che gli animali T1+/T2+ siano considerati ammalati. Analogamente, è pacifico che gli animali T1-/T2- siano considerati sani. Ma come interpretare gli animali T1+/T2- e quelli T1-/T2+?

Vi sono due possibilità: l'interpretazione «OR» e l'interpretazione «AND».



L'interpretazione ***in modalità OR*** considera infetto (o ammalato) un animale che è risultato positivo ad un test o all'altro o ad ambedue. Vengono quindi classificati come ammalati i seguenti animali: T+/T+, T+/T-, T-/T+. Questa interpretazione aumenta la sensibilità ma diminuisce la specificità. Ciò è intuitivo, in quanto si fornisce a ciascun animale una maggiore opportunità (=probabilità) di reagire positivamente.

Nota che, ai fini della sensibilità e della specificità globali, l'utilizzo di due test «in parallelo modalità AND» é sovrapponibile all'utilizzo degli stessi due test «in serie». Però, nella pratica, fra le due strategie (serie/parallelo) vi è una differenza importante: nel caso dei test in serie il numero complessivo di test da effettuare é inferiore rispetto ai test in parallelo. Infatti, nel caso della strategia «in serie» si effettua il primo test su tutti gli animali, ed il secondo test solo su quelli risultati positivi. Invece, con la strategia «in parallelo» si saggiano tutti gli animali sia con il primo che con il secondo test. Per questo motivo la strategia «test in parallelo modalità AND» normalmente non viene utilizzata.

L'interpretazione ***in modalità AND*** considera infetto (o ammalato) un animale che è risultato positivo ad entrambi i test. Vengono quindi classificati come ammalati i seguenti animali: T+/T+. Questa modalità consente di ottenere una maggiore specificità; ciò è facilmente intuibile se si pensa che, per ciascun animale saggiato, la *probabilità* di risultare positivo a entrambi i test è inferiore rispetto a quanto avviene interpretando i risultati con modalità OR.

Test in parallelo: interpretazione con modalità OR o AND

Esaminiamo una simulazione-esempio di applicazione di due test in parallelo.

In questa simulazione, esamini una popolazione di 6000 bovini, applicando su ciascuno di essi due test (TEST1 e TEST2). Per rendere efficace l'esempio, supponiamo che tu conoscsa già lo stato reale di ciascuno dei 6000 animali. In particolare, sai che 300 di essi sono ammalati, ed i restanti 5700 sono sani.

Una volta effettuati i test, ordini i dati ottenuti in un foglio di calcolo. Nel database ogni riga rappresenta un animale, mentre ogni colonna contiene i dati codificati utilizzando il codice «0» (che significa «0» oppure «negativo», o più in generale «assenza del fenomeno») oppure viceversa il codice «1».

ESTRATTO DAL DATABASE

	A	B	C	D	E	F
1	id. animale	esito TEST1	esito TEST2	**stato reale**	interpr. OR	interpr. AND
2	OK741	1	0	**1**	1	0
3	LB419	0	1	**1**	1	0
4	PT630	1	1	**1**	1	1
5	QO489	1	1	**1**	1	1
6	IX680	0	0	**1**	0	0
7	JD724	1	0	**0**	1	0
8	OM127	1	0	**0**	1	0
9	EG416	0	1	**0**	1	0
10	LS552	0	1	**0**	1	0
11	DW345	1	1	**0**	1	1
12	RE872	0	0	**0**	0	0
13	DQ879	0	0	**0**	0	0
14	EI129	0	0	**0**	0	0
15	DH749	0	0	**0**	0	0

GUIDA ALLA LETTURA DEI DATI DEL DATABASE.

Esempio: l'animale OK741 è risultato positivo al TEST1 (codice "1" in colonna B) e negativo al TEST2 (codice "0" in colonna C). In realtà esso è ammalato (codice "1" in colonna D).

In base all'esito dei due test, esso è da considerare "ammalato" se si adotta l'interpretazione "OR" (colonna E), ma è da considerare "sano" se si adotta l'interpretazione "AND" (colonna F).

Ora, sempre utilizzando i dati del database, puoi allestire le tabelle di contingenza separatamente per ciascuno dei due test, e calcolarne la sensibilità e la specificità, come segue:

TEST1	malati	sani	totale
test +	230	225	445
test –	70	5475	5545
totale	300	5700	6000

Se = 230 / (230 + 70) = 0.767

Sp = 5475 / (5475 + 225) = 0.961

TEST2	malati	sani	totale
test +	190	85	275
test –	110	5615	5725
totale	300	5700	6000

Se = 190 / (190 + 110) = 0.633

Sp = 5616 / (5615 + 85) = 0.985

A questo punto puoi tabulare, sempre a partire dal database, i risultati combinati dei due test in parallelo, come nella Tabella sottostante. Per chiarire la lettura della Tabella, aggiungo che: - il valore 80 indica che 80 animali *ammalati* sono risultati «+/-», ossia positivi al TEST1 e negativi al TEST2; - il valore 200 indica che 200 animali *sani* sono risultati «+/-», ossia positivi al TEST1 e negativi al TEST2; - il valore 40 indica che 40 animali *ammalati* sono risultati «-/+», ossia negativi al TEST1 e positivi al TEST2; ... e così via.

TEST1 / TEST2	malati	sani
+ / –	80	200
– / +	40	60
+ / +	150	25
– / –	30	5415

Infine, puoi calcolare la sensibilità e la specificità dei due test in parallelo, sia con intepretazione in modalità «AND» che in modalità «OR». Il calcolo è illustrato graficamente, con l'aiuto dei colori, nella Tabella che segue.

INTERPRETAZIONE "OR" (Test1 OR Test2)

$$\text{Sensibilità} = \frac{80+40+150}{300} = 0.900$$

TEST1/TEST2	malati	sani
+ / -	80	200
- / +	40	60
+ / +	150	25
- / -	30	5415
totale	300	5700

$$\text{Specificità} = \frac{5415}{5700} = 0.950$$

TEST1/TEST2	malati	sani
+ / -	80	200
- / +	40	60
+ / +	150	25
- / -	30	5415
totale	300	5700

INTERPRETAZIONE "AND" (Test1 AND Test2)

$$\text{Sensibilità} = \frac{150}{300} = 0.500$$

TEST1/TEST2	malati	sani
+ / -	80	200
- / +	40	60
+ / +	150	25
- / -	30	5415
totale	300	5700

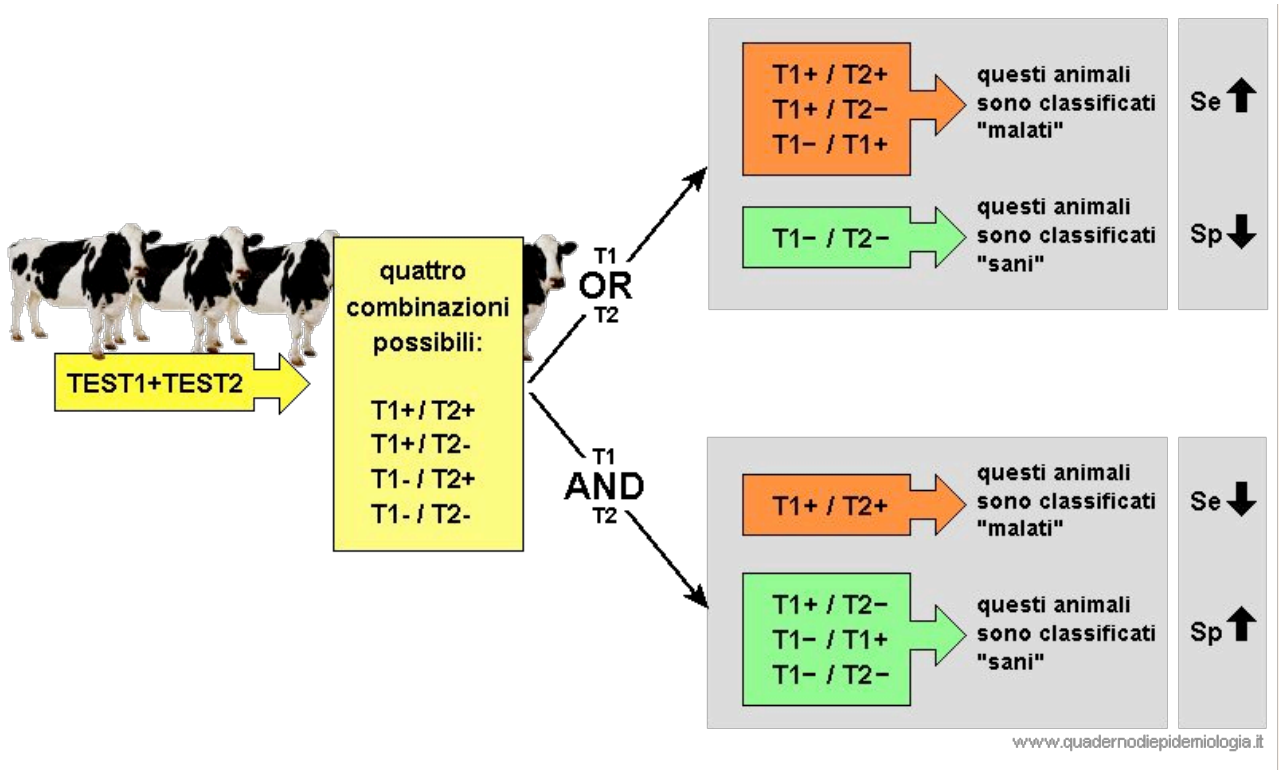
$$\text{Specificità} = \frac{200+60+5415}{5700} = 0.996$$

TEST1/TEST2	malati	sani
+ / -	80	200
- / +	40	60
+ / +	150	25
- / -	30	5415
totale	300	5700

In sintesi:

	Sensibilità	Specificità
Test1	0.767	0.960
Test2	0.633	0.985
Test1 OR Test2	0.900	0.950
Test1 AND Test2	0.500	0.996

Infine, per riassumere il tutto, nella figura seguente vengono schematizzate le differenze fra l'interpretazione OR e quella AND.



Si può vedere come, nel caso dell'**interpretazione «OR»** è meno probabile - rispetto all'utilizzo di un singolo test - che un animale infetto sfugga alla diagnosi; tuttavia, si avrà un incremento dei falsi-positivi, ossia dei soggetti sani che vengono classificati come ammalati. Si ottiene un innalzamento della sensibilità, a scapito della specificità.

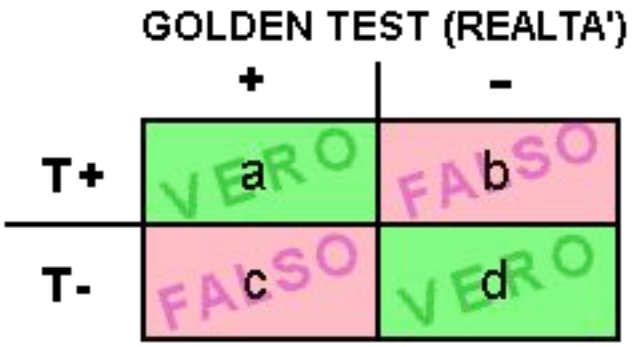
Nel caso dell'**interpretazione «AND»**, è più facile che sfuggano alla diagnosi animali infetti, però diminuisce la probabilità che un animale sano sia classificato come ammalato. La specificità aumenta, a scapito della sensibilità.

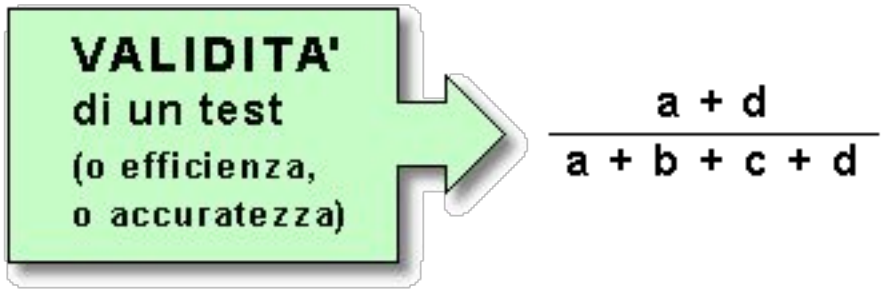
Validità di un test

Dopo aver applicato un test su una popolazione, assumono importanza due indici: il valore predittivo positivo ed il valore predittivo negativo, che misurano quanto i risultati ottenuti si avvicinano alla realtà. Ciò, però, può essere riassunto in *un solo* indice, detto validità (o efficienza, o accuratezza).

La validità di un test è la sua capacità di classificare correttamente sia gli individui malati che quelli sani. La validità è tanto più alta quanto più il test classifica come positivi gli individui realmente malati e come negativi quelli realmente sani. In altri termini, la validità è la capacità di generare risultati rispondenti al vero sia negli individui ammalati che in quelli sani.

La validità può essere calcolata facilmente qualora si conosca il vero stato degli individui che sono stati sottoposti al test. In tal caso, utilizzando la ben nota Tabella di contingenza, la validità si esprime con la proporzione: (a+d)/(a+b+c+d).





ESEMPIO. Abbiamo sottoposto 300 bovini al test della tubercolina per la diagnosi di tubercolosi; successivamente, i bovini sono stati macellati, e su di essi è stata effettuato un minuzioso esame anatomo-patologico di visceri e linfonodi per evidenziare le lesioni tipiche della tubercolosi. Hai ottenuto i risultati riportati nella

tabella.

		Presenza lesioni	
		sì	no
Tubercolina	+	25	16
	–	6	253

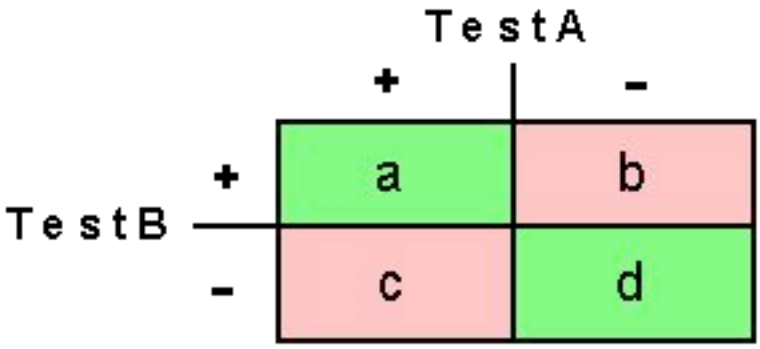
L'esame anatomo-patologico rappresenta il golden standard; infatti, puoi essere ragionevolmente certo che un animale privo di lesioni specifiche sia esente dall'infezione, e viceversa. La prova della tubercolina ha identificato correttamente 25 animali infetti e 253 animali sani (v. Tabella a lato). La validità della prova della tubercolina è: (25+253) / 300 = 0.928. Ciò significa che il test della tubercolina, nelle tue condizioni, ha identificato correttamente lo stato di un animale (non importa se malato o sano) nel 92.8% dei casi.

Concordanza fra due test

Quando si tratta di valutare la *performance* di un test, talvolta può essere necessario confrontarlo non con la realtà o con l'esito di golden test, bensì con un altro test, magari non eccellente ma di comune impiego nella pratica. In questo caso, non si parla più di validità, ma di «concordanza».

La concordanza può riguardare non solo il grado di accordo che si osserva fra due test, ma anche quello fra due (o più) operatori che interpretano l'esito di uno stesso test (es. radiografie, elettrocardiogramma, auscultazione cardiaca ecc.), oppure fra due letture effettuate da uno stesso operatore in tempi diversi. Non si vuole stabilire quale classificazione sia più corretta, bensí stabilire se i criteri utilizzati per l'interpretazione del test siano efficienti, e se classificazione sia riproducibile.

Il calcolo della concordanza è analogo a quello della validità. Supponendo quindi di confrontare due test (TestA e TestB), si ha quanto segue:



ESEMPIO. Hai saggiato 134 sieri suini con due test (TestA e TestB) allo scopo di verificare la presenza di paratubercolosi nel bovino, ottenendo i seguenti risultati: 18 positivi a entrambi i test; 102 negativi ad entrambi i test; 8 positivi a TestA e negativi a TestB; 6 negativi a TestA e positivi a TestB. La concordanza fra i due test è: (18+102)/(18+102+8+6) = 0.896.

L'indice «Kappa» di Cohen

La concordanza calcolata come sopra descritto è criticabile in quanto non tiene conto della quota di concordanza dovuta al caso.

		Studente A	
		+	–
Studente B	+	25	25
	–	25	25

Esempio. Due studenti decidono di valutare, ognuno per proprio conto, una serie di 100 radiografie dell'addome di altrettanti cani con sospetto di calcolosi epato-biliare. Gli studenti classificano le immagini radiologiche attraverso... il lancio di una moneta. Verosimilmente, essi otterranno risultati simili a quelli della Tabella, raggiungendo una concordanza del 50% in base al calcolo seguente: (25+25)/100=0.5.

Come vedi, una classificazione puramente casuale, come quella ottenuta attraverso il lancio di una moneta, restituisce valori di concordanza prossimi a 50%, che sono ovviamente ingannevoli. Per calcolare la quota di concordanza «vera» occorre stabilire quanta parte della concordanza totale osservata è dovuta al caso, e quanta è invece dovuta al reale accordo tra gli osservatori o i test utilizzati. Ciò si ottiene attraverso un metodo statistico che, a partire dai dati della tabella di contingenza, consente di calcolare il Kappa di Cohen.

L'interpretazione dei valori Kappa si esegue secondo le seguenti linee-guida: k<0.2= concordanza scarsa; k compreso fra 0.2 e 0.4 = concordanza modesta; fra 0.41 e 0.61 = moderata; fra 0.61 e 0.80 = buona; >0.80 = eccellente.

Per le modalità di calcolo, consulta la seguente presentazione.

K di Cohen

Il Kappa di Cohen è un indice che consente di calcolare il grado di accordo

- tra due test, oppure
- tra due valutatori che usano lo stesso test oppure
- tra due letture di uno stesso test, effettuate in tempi diversi da parte di uno stesso valutatore.

L'indice Kappa esprime la concordanza *reale*, cioè escludendo quella dovuta al caso

COS'E' LA CONCORDANZA DOVUTA AL CASO?

di Cohen

Uno studente maldestro ed un radiologo esperto esaminano separatamente 100 ecografie per la diagnosi di gravidanza, ottenendo i seguenti risultati:

		STUD. +	STUD. -	STUD. tot
RADIOL.	+	0	4	4
	-	0	96	96
	tot	0	100	100

Come vedi, per lo studente inesperto le ecografie sono tutte negative. Il radiologo invece ne identifica 4 positive.

Calcola la concordanza tra i due operatori..

		STUD. +	STUD. -	STUD. tot
RADIOL.	+	0	4	4
	-	0	96	96
	tot	0	100	100

di Cohen

Concordanza =
(0+96)/100 = 0.96 = 96%

Ti sembra appropriata questa analisi?

Ti sembra appropriato ritenere che lo studente sia molto bravo, concordando al 96% con il radiologo esperto?

Evidentemente no!

		STUD. +	STUD. -	STUD. tot
RADIOL.	+	0	4	4
	-	0	96	96
	tot	0	100	100

K

di Cohen

Concordanza =
(0+96)/100 = 0.96 = 96%

Ti sembra appropriata questa analisi?

Ti sembra appropriato ritenere che lo studente sia molto bravo, concordando al 96% con il radiologo esperto?

Evidentemente no!

Il fatto è che la concordanza del 96% è quella **OSSERVATA**, ma **NON** è quella **REALE**: infatti **una quota della concordanza osservata è DOVUTA AL CASO**

K

di Cohen

Il Kappa di Cohen consente di misurare la concordanza REALE, al netto di quella dovuta al caso.

Nelle prossime slide, attraverso un esempio ragionato, ti mostrerò passo-passo come si calcola il Kappa di Cohen.

Nella tabella che segue (che per tua comodità viene replicata nella slide successiva) sono riassunti i risultati di 2 diversi osservatori (OssA e OssB) che hanno classificato 200 animali usando, ciascuno per proprio conto, lo stesso test.

K di Cohen

		OssA		
		+	-	tot
OssB	+	97	21	118
	-	12	70	82
	tot	109	91	200

Accordo osservato = (97+70)/200 = 0.835 = 83.5%

		OssA		
		+	-	tot
OssB	+	97	21	118
	-	12	70	82
	tot	109	91	200

K di Cohen

Complessivamente:

OssA ha eseguito

109/200=**0.545** classificazioni "positive"

91/200=**0.455** classificazioni "negative"

OssB ha eseguito

118/200=**0.590** classificazioni "positive"

82/200=**0.410** classificazioni "negative"

In base alla "Regola della moltiplicazione"

(⇨ Cap. 5, Unità2) si ha che:

		OssA		
		+	-	tot
OssB	+	97	21	118
	-	12	70	82
	tot	109	91	200

di Cohen

Complessivamente:

OssA ha eseguito
109/200=**0.545** classificazioni "positive"
91/200=**0.455** classificazioni "negative"

OssB ha eseguito
118/200=**0.590** classificazioni "positive"
82/200=**0.410** classificazioni "negative"

In base alla "Regola della moltiplicazione" (⇨ Cap. 5, Unità2) si ha che:

probabilità *dovuta al caso* che *entrambi* gli osservatori abbiano espresso una

- **classificazione positiva=**
  0.545*0.590 = 0.322
- **classificazione negativa=**
  0.455*0.410 = 0.187

di Cohen

Perciò, la probabilità dovuta al caso che i due osservatori abbiano espresso un giudizio (non importa se positivo o negativo) concorde è:

0.322 + 0.187 = 0.509 = 50.9%

Questo è l'ACCORDO DOVUTO AL CASO

K

di Cohen

Poiché la probabilità dovuta al caso che i due osservatori abbiano espresso un giudizio concorde è **0.509**, allora la probabilità NON dovuta al caso che essi abbiano espresso un giudizio concorde è

1 - 0.509 = 0.491 = 49.1%

Questo è l'ACCORDO NON DOVUTO AL CASO

Ora sei in possesso dei dati che servono
per calcolare il Kappa di Cohen, e cioè:

Accordo osservato =	0.835
Accordo dovuto al caso=	0.509
Accordo non dovuto al caso=	0.491

La formula è la seguente:

$$k = \frac{\text{accordo osservato - accordo dovuto al caso}}{\text{accordo non dovuto al caso}}$$

k = (0.835-0.509) / 0.491 = 0.664

Interpretazione dei valori di Kappa
(Landis & Koch, Biometrics, 33.159, 1977)

Kappa	Concordanza
<0.01	nulla
0.01-0.20	scarsa
0.21-0.40	modesta
0.41-0.60	moderata
0.61-0.80	sostanziale
0.81-1.00	eccellente

Riprendendo quanto già osservato circa la relatività degli indici di Se e Sp in riferimento al variare del valore di cut-off, per cui:

1. la Se e la Sp non rappresentano descrittori esaurienti della *performance* del test;

2. i valori predittivi, in quanto dipendenti dalla prevalenza della malattia nella popolazione studiata, non sono caratteristiche intrinseche del test e quindi non possono essere utilizzati come descrittori esaurienti della *performance* dei test.

Cosa vuol dire tutto ciò ? Vuol dire che quando leggete un paper che tratta di studi su test diagnostici, dovete essere in grado di trovare informazioni circa tutte e quattro le misure trattate insieme agli intervalli di confidenza. Se la misura impiegata dal test è di tipo ordinale o metrica, dovreste trovare una qualche evidenza che siano stati valutati più di un cut-off. Dovete anche acquisire una qualche informazione sulla prevalenza della condizione nella popolazione o nelle popolazioni sulle quali è applicato il test, sopratutto se poi il test deve essere impiegato su popolazioni differenti o più grandi.

Per trovare un punto di equilibrio ottimale tra sensibilità e specificità in relazione al punto di cut-off da impiegare, un metodo molto popolare è quello che si avvale della curva ROC (vedi più avanti).

Il rapporto di verosimiglianza (Likelihood Ratio)

Il rapporto di verosimiglianza è la probabilità che un risultato (positivo o negativo) sia atteso in un paziente con una certa malattia rispetto alla probabilità che lo stesso risultato sia atteso in un paziente senza quella malattia.

LR+ corrisponde al rapporto tra la probabilità che un individuo che abbia la malattia risulti positivo al test e la probabilità che un individuo che non abbia la malattia risulti positivo al test

LR- corrisponde al rapporto tra la probabilità che un individuo che abbia la malattia risulti negativo al test e la probabilità che un individuo che non abbia la malattia risulti negativo al test.

Ad es. LR+ uguale a 24 significa che un risultato positivo del test è 24 volte più probabile che provenga da un soggetto affetto dalla malattia piuttosto che da un soggetto non affetto. Per un valore di LR- uguale a 0.02 significa che con un test negativo un soggetto ha la probabilità del 98% di non avere la malattia e del 2% di averla.

I rapporti di verosimiglianza positivi vanno presi in considerazione quando sono nel range di >= 2 e sono utili dal punto di vista clinico quando sono maggiori di 5. Per i rapporti di verosimiglianza negativi i valori da tenere in considerazione sono quelli < 0.1 .

LR viene usato per stabilire quanto un test diagnostico è valido e quindi per selezionare il test diagnostico appropriato.

LR ha il vantaggio di essere meno influenzato dalla prevalenza della malattia.

In associazione al dato sulla prevalenza della malattia LR può essere usato per calcolare la probabilità post-test per una data malattia.

Il rapporto di verosimiglianza impiega per il suo calcolo gli indici di sensibilità e di specificità del test diagnostico.

LR+ = sensibilità / 1 - specificità

LR- = 1 - sensibilità / specificità

La probabilità pre-test è quella di avere la malattia prima di effettuare l'esame e corrisponde alla prevalenza della malattia nel campione di soggetti rappresentato dal paziente in esame.

La probabilità post-test è la probabilità di avere la malattia dopo aver effettuato il test. LR contribuisce quindi a modificare la probabilità che un individuo sia affetto dalla condizione in esame. Un buon test diagnostico fa aumentare di numerose volte

la probabilità post-test rispetto a quella pre-test.

Ciò permette al clinico una migliore interpretazione del risultato e aiuta a predire la probabilità di un risultato vero positivo.

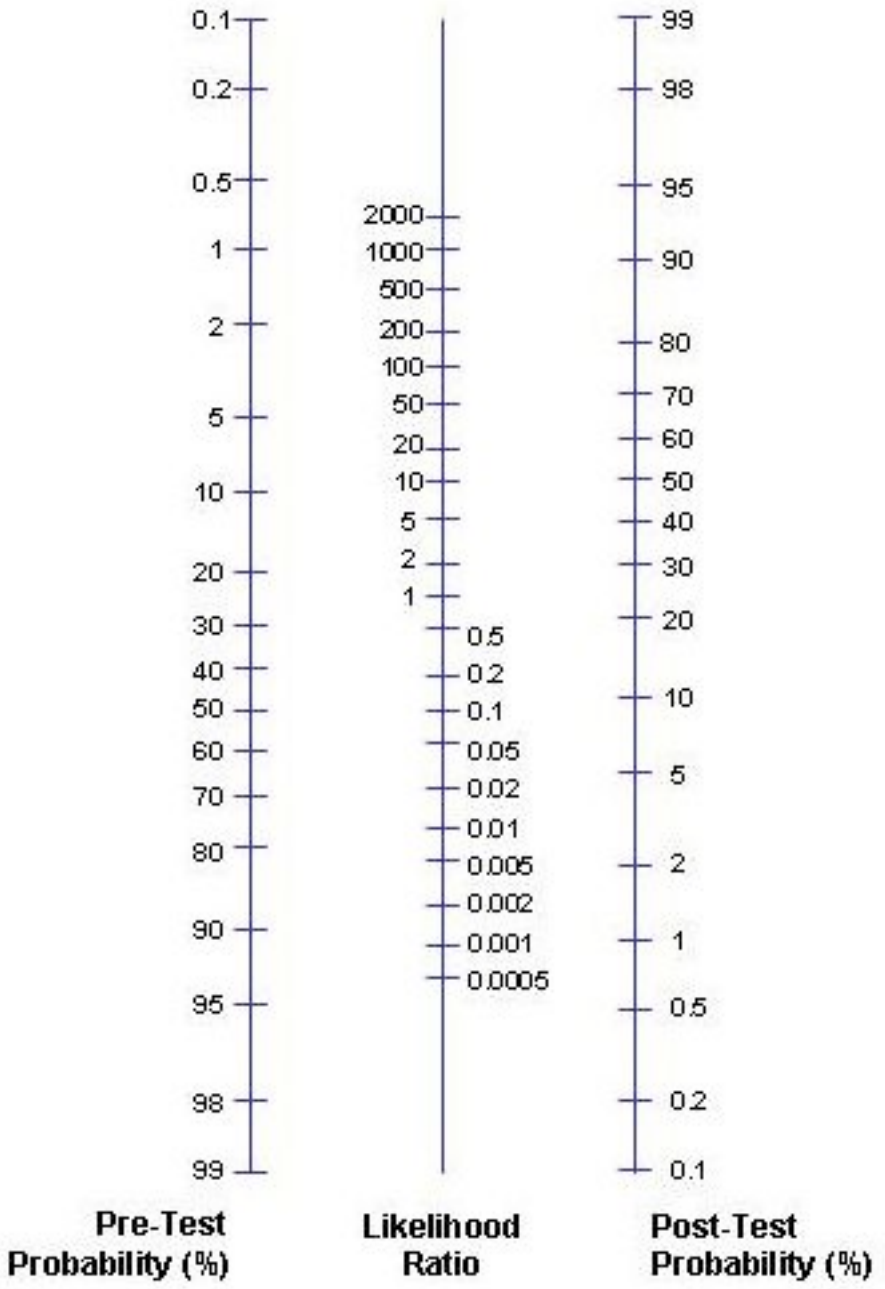
A titolo di esempio, un test diagnostico con una sensibilità del 67% e una specificità del 91% viene impiegato su un campione di 2030 soggetti per investigare una malattia dalla prevalenza di 1.48% .

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Likelihood ratio positive = sensitivity / (1 – specificity) = 66.67% / (1 – 91%) = 3.53

Likelihood ratio negative = (1 – sensitivity) / specificity = (1 – 66.67%) / 91% = 0.37

In base a questi numeri, se la probabilità di un paziente di avere un cancro all'intestino è di 1.48% prima del test, e la LR+ è 3.53 mediante il normogramma potete apprendere che la probabilità per quel paziente di avere proprio quel tipo di tumore sale con il test positivo a circa il 5%.



In altri termini, ciò suggerisce che l'esame cui il paziente è stato sottoposto non è un test molto efficace per diagnosticare tale malattia.

Alternativamente all'uso del normogramma, e note la probabilità pre-test e il rapporto di verosimiglianza, si può calcolare la probabilità post-test in tre passi successivi nel modo seguente:

1) Probabilità pre-test / (1 - Probabilità pre-test) = Pretest odds
2) Pre-test odds * rapporto di verosimiglianza = Post-test odds
3) Post-test odds / (Post-test odds + 1) = Probabilità Post-test

Dalla tabella 2x2 deduciamo 20 veri positivi, 10 falsi negativi, and 2030 total patients. Il positive pre-test probability viene calcolato nel seguente modo:

Pretest probability = (20 + 10) / 2030 = 0.0148
Pretest odds = 0.0148 / (1 - 0.0148) =0.015
Posttest odds = 0.015 * 3.53 = 0.053
Posttest probability = 0.053 / (0.053 + 1) =0.05 or 5%

Abbiamo un numero elevato di falsi positivi e un numero basso di falsi negativi, quindi un risultato positivo del test non è sufficiente per confermare la diagnosi (PPV = 10%). e occorre eseguire ulteriori indagini. Il test ha identificato correttamente il 66.7% dei casi positivi (sensibilità) e il 91% dei casi negativi (specificità) e ha un NPV (99.5%) molto elevato. Quindi, come test di screening, va detto che un risultato negativo è importante perchè ci rassicura che il paziente molto probabilmente non ha la malattia in questione.

Un altro esempio in cui andiamo a calcolare tutti gli indici.

Questo esempio è preso dai risultati di una rassegna sistematica sulla ferritina sierica come indice diagnostico per la anemia sideropenica

		TARGET DISORDER (Iron deficieny anaemia)		
		Present	Absent	TOTALS
DIAGNOSTIC TEST RESULT (serum ferritin)	Positive (<65 mmol/L)	731 (a)	270 (b)	1001 (a+b)
	Negative (≥65 mmol/L)	78 (c)	1500 (d)	1578 (c+d)
	TOTALS	809 (a+c)	1770 (b+d)	2579 (a+b+c+d)

Calcoli

Sensitivity = VP / VP + FN = 731/809 = 90%

Specificity = TN / TN + FP = 1500/1770 = 85%

Positive Predictive Value = VP / VP + FP = 731/1001 = 73%

Negative Predictive value = TN / TN + FN = 1500/1578 = 95%

Prevalence = VP + FN / VP + FN + TN + FP = 809/2579 = 32%

LR+ = sensibilità / (1-specificità) = 90/15 = 6

LR- = (1- sensibilità) / specificità = 10/85 = 0.12

Pre-test odds = prevalence / (1-prevalence) = 32/68 = 0.45

Pre-test odds = pre-test probability / (1-pre-test probability)

Post-test odds = pre-test odds * LR = 0.45*6 = 2.7

Post-test Probability = post-test odds / (post-test odds + 1) = 2.7/3.7 = 0.73 = 73%

L’analisi ROC (*Receiver Operating Characteristic* o *Relative Operating Characteristic*).

Ma torniamo ora al problema di come trovare un punto di equilibrio ottimale tra sensibilità e specificità in relazione al punto di cut-off da impiegare, ovvero la analisi ROC.

Si tratta di una curva che rappresenta per ogni punto di cut-off la combinazione di sensibilità, il tasso dei veri positivi, sull'asse verticale, con il valore di 1- la specificità, il tasso di falsi positivi, sull'asse orizzontale.

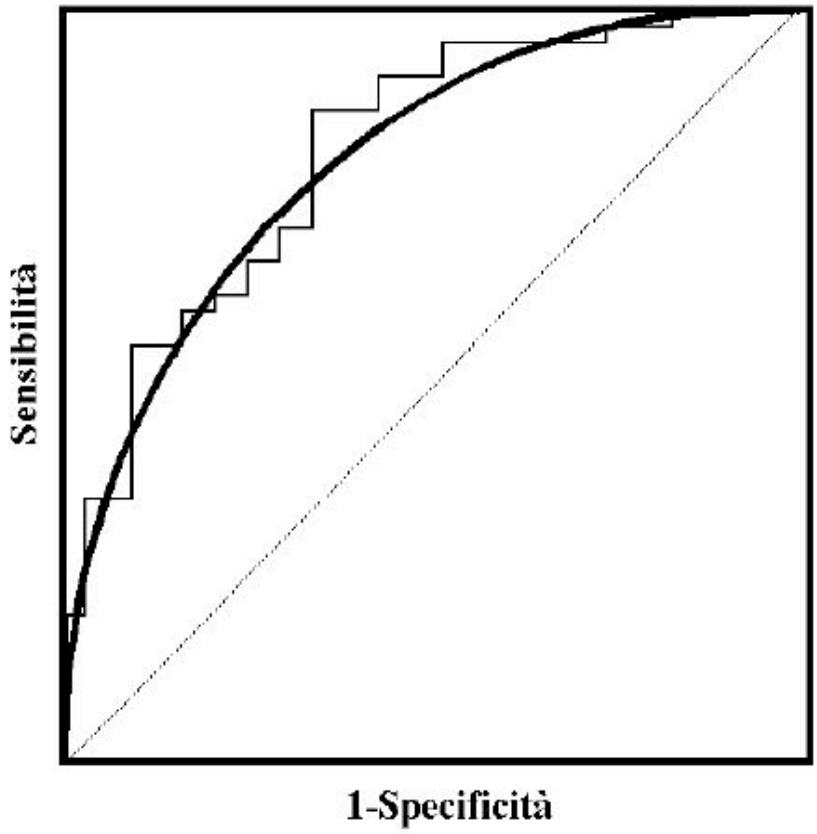
Il valore di cut-off ottimale è quel punto sulla curva che si trova più vicino all'angolo in alto a sinistra del sistema di assi. Questo è il punto che massimizza l'area sotto la curva. In pratica si calcola per ogni punto di cut-off il valore dell'area sotto la curva e si sceglie quello più elevato.

Le curve ROC: il principio di base

Si tratta di una metodologia che in tempi più recenti è divenuta relativamente comune per la valutazione non solo delle immagini, ma anche dei più svariati test nel settore medico (con particolare riguardo alla valutazione dei test clinici di laboratorio). L’analisi ROC viene effettuata attraverso lo studio della funzione che – in un test quantitativo – lega la probabilità di ottenere un risultato vero-positivo nella classe dei malati-veri (ossia la sensibilità) alla probabilità di ottenere un risultato falso-positivo nella classe dei non-malati (ossia 1-specificità).

La relazione tra i suddetti parametri può venire raffigurata attraverso una linea che si ottiene riportando, in un sistema di assi cartesiani e per ogni possibile valore di *cut off*, la proporzione di veri positivi in ordinata e la proporzione di falsi positivi in ascissa.

Si possono calcolare i valori di sensibilità e 1-specificità per ogni valore registrato (cut-off). L’unione dei punti ottenuti riportando nel piano cartesiano ciascuna coppia (Se) e (1-Sp) genera una curva spezzata da cui è possibile per interpolazione ottenere una curva (ROC *curve*).



La capacità discriminante di un test, ossia la sua attitudine a separare propriamente la popolazione in studio in “malati” e “sani” è proporzionale all’estensione dell’area sottesa alla curva ROC (*Area Under Curve*, AUC).

Nel caso di un test perfetto, ossia che non restituisce alcun falso positivo né falso negativo (capacità discriminante = 100%), la AUC passa attraverso le coordinate {0;1} ed il suo valore corrisponde all’area dell’intero quadrato delimitato dai punti di coordinate (0,0), (0,1), (1,0) (1,1), che assume valore 1 corrispondendo ad una probabilità del 100% di una corretta classificazione.

Come regola generale, si può affermare che il punto sulla curva ROC più vicino all’angolo superiore sinistro rappresenta il miglior compromesso fra sensibilità e specificità.

Valutazione della *performance* di un singolo test mediante una curva ROC

L’area sottesa ad una curva ROC rappresenta un parametro fondamentale per la valutazione della *performance* di un test, in quanto costituisce una misura di accuratezza non dipendente dalla prevalenza (“*pure accuracy*”).

Poiché AUC rappresenta una stima, risulta quasi sempre necessario testare la significatività della capacità discriminante del test, ovvero se l'area sotto la curva eccede significativamente il suo valore atteso di 0.5. Tale procedura corrisponde a verificare se la proporzione dei veri positivi è superiore a quella dei falsi positivi.

Stima dell'area sottesa ad una curva ROC

Il calcolo dell'AUC per una curva empirica (cioè ottenuta da un campione finito) può venire effettuato semplicemente connettendo i diversi punti del ROC *plot* all'asse delle ascisse con segmenti verticali e sommando le aree dei risultanti poligoni generati nella zona sottostante.

Per quanto riguarda l'interpretazione del valore di AUC, essa è basata su criteri largamente soggettivi secondo lo schema seguente.

• AUC=0.5 test non informativo

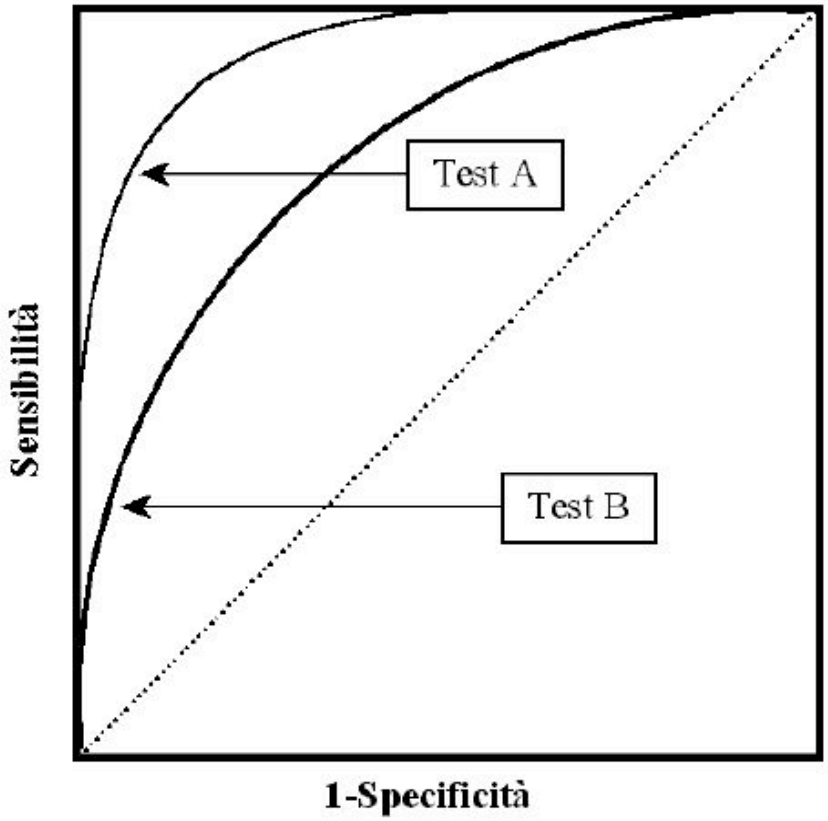
• $0.5<AUC\leq0.7$ test poco accurato

• $0.7<AUC\leq0.9$ test moderatamente accurato

• $0.9<AUC<1.0$ test altamente accurato

• AUC=1.0 test perfetto

Per finire, possiamo aggiungere che due test possono essere quindi confrontati tra di loro comparando le *accuracy* stimate mediante l'area sottesa alle corrispondenti curve ROC.



Risulta evidente la superiorità del test A la cui curva ROC teorica si trova interamente al di sopra di quella corrispondente al test B.