

# Modelli predittivi di rischio

## Sensibilità, Specificità e Area ROC

Gabriele Cevenini

*Dipartimento di Biotecnologie Mediche - Università degli Studi di Siena*

I modelli predittivi per la valutazione del rischio di un evento sanitario sfavorevole producono in uscita una probabilità di rischio che può essere utilizzata per discriminare tra due singole condizioni: evento sfavorevole (p.es. insorgenza di una malattia) ed evento propizio (stato di buona salute o guarigione). Ciò semplicemente definendo una probabilità *cut-off* detta anche 'soglia di decisione', sopra la quale viene previsto l'evento sfavorevole. Per questo motivo la valutazione della capacità discriminante del modello può essere ricondotta allo studio delle prestazioni di un test di laboratorio per scopi diagnostici che, come noto, fornisce un risultato negativo (nel nostro caso associabile alla predizione dello stato di salute) o positivo (associabile alla predizione di malattia).

Un test diagnostico di laboratorio trova giustificazione nella sua rilevanza clinica: deve cioè essere in grado di fornire utili informazioni che non possono essere ottenute altrimenti o che sono ottenute a minor costo o a minor rischio per il paziente. Come esempio si può riportare il dosaggio degli enzimi per la diagnosi dell'infarto acuto del miocardio che è definito come una necrosi del tessuto miocardico; tuttavia, per la sua diagnosi non è pensabile un'osservazione diretta al microscopio del tessuto d'ogni paziente che si presenta con dolore toracico. Quindi, si ricorre ad altre procedure diagnostiche come il dosaggio degli enzimi cardiaci, il cui innalzamento, in spazi temporali ben definiti, si correla all'evento infartuale.

Per essere clinicamente rilevante, l'anormalità di un parametro di laboratorio deve avere un'elevata associazione con la malattia che dovrebbe confermare e contemporaneamente non dovrebbe presentarsi in altre malattie e in particolare in quelle che sono caratterizzate da una sintomatologia analoga, e possono essere argomento di diagnosi differenziale. Per definire la performance di un test di laboratorio se ne misura la sensibilità e la specificità diagnostica, i valori predittivi in caso di positività e di negatività e infine l'accuratezza o efficienza.

**Sensibilità diagnostica.** Con sensibilità (SE) di un test s'intende la sua capacità di rivelare una certa malattia nei soggetti che hanno effettivamente la malattia (malati). In termini statistici essa corrisponde alla probabilità che il test risulti alterato nel paziente con la malattia in esame, e si esprime con il rapporto tra i malati che risultano positivi al test (veri positivi, VP) ed il totale dei malati considerati nello studio. Una sensibilità del 95% significa che il 5% dei pazienti con la malattia (falsi negativi, FN) avranno un test falsamente negativo. Più alta è SE, maggiore è la probabilità che la malattia venga identificata e minore il numero dei FN.  $SE = VP / (VP + FN)$

**Specificità diagnostica.** Con specificità (SP) di un test s'intende la capacità di risultare negativo (entro l'intervallo di riferimento) negli individui che non hanno la malattia (sani). In termini statistici è la probabilità che il test risulti negativo in un soggetto che non ha la malattia in esame e si esprime come il rapporto tra il numero dei soggetti sani che risultano negativi al test (veri negativi, VN) e quello totale dei sani considerati nello studio. Se la specificità è del 98%, il 2% dei soggetti non malati ha il test falsamente positivo (falsi positivi, FP).  $SP = VN / (VN + FP)$

**Efficienza.** Con efficienza (EF), detta anche accuratezza, del test s'intende la percentuale dei risultati corretti, cioè i malati positivi al test più i sani negativi al test e si esprime con il rapporto fra il totale dei risultati corretti e il totale generale delle prove.  $EF = (VP + VN) / (VP + FN + VN + FP)$

Il termine efficienza è tuttavia fuorviante perché sembra indicare l'utilità diagnostica, il che non è necessariamente vero. Benché questo indice di accuratezza rappresenti una sorta di media pesata tra sensibilità e specificità, il suo utilizzo per giudicare la validità di un test diagnostico suscita diverse controversie. Vedremo più avanti come l'introduzione dell'area sottesa alla curva ROC permetta di definire meglio l'accuratezza globale del test.

Una procedura diagnostica o un test di laboratorio ideale dovrebbe dare sempre risultati nell'intervallo di riferimento per i soggetti non ammalati e risultati anormali per quelli che hanno la malattia: sia SE, sia SP dovrebbero essere il 100%. In pratica, un tale test non esiste. Talvolta si ha l'opportunità di scegliere tra un test con alta SE e bassa SP e uno con più alta SP a spese di una ridotta SE. La scelta dovrebbe essere fatta a seconda che il test venga impiegato con finalità di riconoscimento o di conferma.

1. Un test di riconoscimento (*screening*) è utile quando si voglia studiare un gruppo di soggetti in cui la malattia ha una discreta prevalenza; in altre parole in soggetti con buona probabilità di avere la malattia. In questo caso il test dovrebbe avere la più alta SE possibile (pochi FN) in modo da consentire il riconoscimento del maggior numero dei casi con malattia. Alcuni FP sono attesi, come sempre quando si ha a che fare con un test ad alta SE, ma il vantaggio è che in presenza di un risultato negativo si potrà escludere tassativamente la malattia (alta predittività del test negativo). Un risultato normale richiederà quindi la conferma con un test dotato di una più alta SP. L'impiego di un test di questo tipo è illustrato dal seguente esempio: la ricerca di anticorpi antinucleo è un test sensibile per il lupus eritematoso sistemico, tuttavia può risultare positivo anche in pazienti con altre malattie del collagene e in taluni soggetti sani. La positività al test non è pertanto diagnostica per il sospetto di lupus, ma un test negativo escluderà la presenza della malattia. Per minimizzare il numero dei FP bisognerebbe includere nello studio solamente i soggetti che hanno maggiore probabilità di avere la malattia.
2. Un test di conferma serve appunto per confermare una diagnosi sospetta (alta predittività del test positivo). Un buon test di questo tipo deve avere pochi FP, deve cioè essere caratterizzato da un'alta SP; tuttavia, un risultato negativo non escluderà la malattia in questione. Anche in questo caso l'esempio seguente ne illustra l'impiego: l'infarto acuto del miocardio induce caratteristiche alterazioni elettrocardiografiche e la loro presenza conferma la diagnosi di infarto. Tuttavia, in molti casi di infarto, non si osservano queste modificazioni, così l'elettrocardiogramma da solo non può essere impiegato per escludere un infarto recente.

SE e SP sono caratteristiche di un test e descrivono quanto questo sia in grado di identificare soggetti affetti o no da una determinata malattia; entrambe vengono desunte da casistiche studiate a posteriori, dopo che la diagnosi è stata accertata con altre procedure. In altre parole rispondono al quesito: una volta che so che un paziente ha o non ha una malattia, qual è la probabilità che il test sia fuori, o dentro, l'intervallo di riferimento? Questo è un dato di grande interesse per conoscere le qualità del test, ma non segue la procedura diagnostica dato che nella pratica il medico è interessato al quesito inverso: una volta che so che un paziente ha il test positivo, qual è la probabilità che sia affetto dalla malattia? Come è noto, è dal sintomo che si giunge alla diagnosi e non accade che data una diagnosi si riesamini il sintomo. Lo strumento più potente, attualmente disponibile, per giudicare l'utilità di un test clinico, è il valore predittivo.

**Il valore predittivo di un test positivo (VPP)** è la probabilità che la malattia sia presente quando il risultato del test è positivo (fuori dall'intervallo di riferimento) e si esprime dal rapporto tra i VP e il totale dei positivi. E' tanto più alto quanto più il test è specifico.  $VPP = VP / (VP + FP)$

**Il valore predittivo di un test negativo (VPN)** è la probabilità che la malattia sia assente quando il test risulta negativo (entro l'intervallo di riferimento) e si esprime dal rapporto tra i VN e il totale dei negativi. E' tanto più alto quanto più il test è sensibile.  $VPN = VN / (VN + FN)$

Il VPN di un esame di laboratorio è spesso poco considerato, specie in confronto al VPP. Si ritiene comunemente che solo un esame con risultato positivo dia informazioni utili per fini diagnostici e terapeutici. Viceversa, molti esami, quando risultano negativi, hanno un ruolo essenziale nella strategia di esclusione di una malattia e presentano un forte impatto non solo nel porre una diagnosi corretta ma anche nell'evitare ricoveri e terapie inappropriate. Un esempio molto attuale è rappresentato dal test del D-dimero, che è assolutamente dirimente nel sospetto di patologie tromboemboliche di tipo venoso. Infatti, se la determinazione del D-dimero produce un risultato negativo, questo dato ha valore predittivo pressochè del 100% nell'escludere la malattia.

Ben superiore a esami invasivi e molto più costosi quali ecodoppler dei vasi degli arti inferiori, tomografia computerizzata a spirale e scintigrafia polmonare perfusoria. Al contrario, nel caso di positività del test il sospetto di tromboembolismo si rafforza, ma sono possibili numerosi FP dovuti a traumi, interventi chirurgici, neoplasie e stati infiammatori e, quindi, il sospetto diagnostico deve essere confermato con esami successivi. Altro esempio di esame ad elevato VPN è quello della mioglobina nella sindrome coronarica acuta. È ben noto come la mioglobina non sia miocardio-specifica e aumenti in tutte le patologie muscolari e renali. Tuttavia, è il marcatore di danno miocardico più precoce, almeno fra quelli per i quali esistono sistemi diagnostici nei laboratori clinici. Un aumento della mioglobina, quindi, non può essere ascrivibile esclusivamente a patologie del miocardio; in questo senso il VPP è modesto e certamente non giustificherebbe l'inclusione del marcatore nel profilo diagnostico del dolore toracico acuto. La determinazione della mioglobina presenta tuttavia un elevato VPN, vicino al 100%, ed è quindi utilissima nella strategia di esclusione. Se effettuato entro poche ore dall'insorgere del dolore toracico, un esito negativo di tale esame permette di escludere, con elevata probabilità, la presenza di un infarto acuto del miocardio. Inoltre, la caratteristica sensibilità e precocità dell'aumento della mioglobina, quando questa determinazione viene associata ad un esame caratterizzato da un'elevata SP (troponina cardiaca), assicurano un'efficacia diagnostica molto alta.

Per il corretto impiego dei valori predittivi di un test si dovrebbero conoscere non solo le sue SE e SP, ma anche la probabilità di identificare la malattia nel gruppo di soggetti studiato. Questa è la probabilità *a-priori* che nello studio di popolazioni coincide con la prevalenza della malattia. Quando ci si riferisce ad un singolo malato la probabilità *a-priori* di essere in presenza di una determinata malattia dipende da vari elementi: oltre alla popolazione cui il soggetto appartiene, anche le condizioni ambientali, la stagione, gli elementi anamnestici (razza, sesso, età, ambiente di vita e di lavoro, ceto sociale, precedenti morbosità, ecc.) e obiettivi e, infine, gli eventuali referti già acquisiti. Benché tutti questi elementi abbiano un peso rilevante è difficile nella pratica quantificarli e, pertanto, pur venendo utilizzati ai fini dell'orientamento, per quanto riguarda la probabilità *a-priori* si considera (quando è noto) solo il criterio della prevalenza. Quando si applica lo stesso test a due popolazioni con prevalenza differente della malattia (per esempio i pazienti sospetti di infarto in Pronto Soccorso e Unità Coronarica) si hanno forti differenze nei valori predittivi. I valori predittivi, infatti, variano a seconda del numero di persone che hanno la malattia nella popolazione cui appartiene il soggetto in studio. In altre parole, il numero dei FP e dei FN varierà nelle popolazioni con differente probabilità (*a-priori*) di malattia: lo stesso test avrà più alto VPP nella popolazione con più alta prevalenza. Questo spiega perché un buon test diagnostico fallisce come test di *screening* quando viene applicato ad una popolazione in cui la prevalenza della malattia è molto bassa.

Con riferimento alla Tabella 1, consideriamo un test per la diagnosi di infarto del miocardio caratterizzato dal 95% di SE e di SP; ciò significa che è positivo nel 95% dei pazienti con infarto e nel 95% dei sani risulta negativo; vogliamo conoscere i valori predittivi nel caso venga applicato sia sui soggetti che si presentano al Pronto Soccorso con una sintomatologia dolorosa al torace, sia sui pazienti che arrivano in Unità Coronarica. È nota la prevalenza degli infarti nei due reparti: 0.05 in Pronto Soccorso e 0.50 in Unità Coronarica.

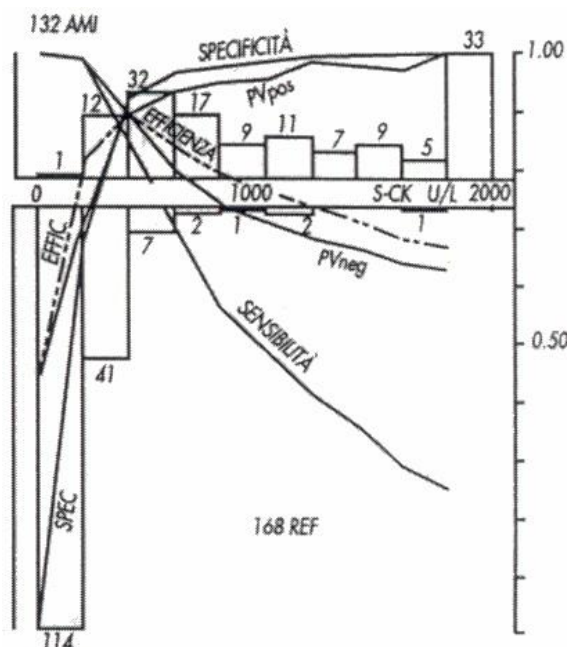
**Tabella 1.** Sensibilità (SE), specificità (SP) e valore predittivo del test positivo (VPP) in due popolazioni con differente prevalenza (PV); IMA = Infarto Acuto del Miocardio.

Bassa prevalenza				Alta prevalenza			
Test	+	-	Totale	Test	+	-	Totale
IMA	48	2	50	IMA	475	25	500
Non IMA	47	903	950	Non IMA	25	475	500
Totale	95	905	1000	Totale	500	500	1000
Pronto soccorso: PV di IMA = 50/1000 = 5%				Unità coronarica: PV di IMA = 500/1000 = 50%			
SE = 48/50 = 96%				SE = 475/500 = 95%			
SP = 903/950 = 95%				SP = 475/500 = 95%			
VPP = 48/95 = 50%				VPP = 475/500 = 95%			

Come mostrato dalla Tabella 2, quando il test viene applicato in Pronto Soccorso in cui la prevalenza è del 5%, esso ha un VPP del 50%, e cioè il 50% dei pazienti con valori fuori dell'intervallo di riferimento è affetto da infarto e il rimanente 50% quindi è rappresentato da FP. Quando questo stesso test viene applicato ai pazienti dell'Unità Coronarica (prevalenza del 50%) avrà un VPP del 95% e cioè il 95% dei pazienti con il test sopra i valori di riferimento avrà l'infarto con soltanto il 5% di FP (Tabella 2).

Come esempio si riporta il caso di un test per HIV che ha SE=100%, SP= 99.7% e viene eseguito su una popolazione con prevalenza della malattia di 0.003 (3/1000). Ciò vuol dire che su 1000 determinazioni si avranno 6 positivi (3 veri, e 3 falsi). VPP è quindi  $3/6 = 0.5$  che in percentuale è 50%. Questo valore può sembrare poco utile dal momento che una probabilità di prevedere un evento del 50% corrisponde a quella del lancio di una moneta. Tuttavia, essendo partiti da una probabilità *a-priori* di 3/1000 (0.003) e avendo ora una probabilità *a-posteriori* pari a 50/100 (0.5), il progresso è evidente sulla conoscenza dello stato clinico del paziente.

Un'altra variabile che influenza la performance clinica di un test diagnostico è il *cut-off* che può corrispondere al limite superiore della norma o ad un opportuno livello scelto arbitrariamente per ridurre FP o FN. Il modo come si modificano sensibilità, specificità e valori predittivi è riportato nel seguente esempio. La Figura 1 riporta la distribuzione dei risultati dell'enzima creatinichinasi nel siero (S-CK) ottenuti su una popolazione di infartuati (AMI), nella parte superiore, con altre procedure diagnostiche e su una di soggetti normali (REF), nella parte inferiore.



**Figura 1.** Curve di sensibilità, specificità, valori predittivi (PV) ed efficienza al variare del *cut-off* della creatinichinasi nel siero (S-CK).

**Tabella 2.** Valore predittivo positivo percentuale (VPP%) di un test con sensibilità e specificità del 95%, in funzione della prevalenza percentuale (PV%) della malattia.

PV%	VPP%
1	16.1
2	27.9
5	50.0
10	67.9
15	77.0
20	82.6
25	86.4
50	95.0

**Tabella 3.** Valori numerici derivati dalla Figura 1.

Cut-off U/l	SE%	SP%	VPP%	VPN%	EF%
0	100	0	44	100	44
200	99	68	71	99	82
400	86	92	90	90	90
600	69	96	94	80	84
800	56	98	95	74	79
1000	49	98	96	71	77
1200	41	99	98	68	74
1400	36	99	98	66	71
1600	29	99	97	64	68
1800	25	100	100	63	67

Come si vede in Fig. 1, è impossibile trovare un *cut-off* che consenta una completa separazione tra i due gruppi dal momento che le due popolazioni hanno risultati che si sovrappongono. Per questo possiamo scegliere il *cut-off* in qualunque punto della zona di sovrapposizione. In Tabella 3

sono riportati i valori estratti dalla Figura 1 delle variazioni di SE, SP, VPP, VPN e EF al variare del *cut-off*: SE e SP variano in maniera opposta così come VPP e VPN. Aumentando il *cut-off*, SP e VPN calano mentre SE e VPP aumentano. La posizione del *cut-off* andrebbe scelta com'è già stato detto in base allo scopo clinico: infatti, se si usa il test come *screening* sarà opportuno posizionare il *cut-off* ad un valore basso per avere la massima SE e il più alto VPN anche a costo di una più bassa SP e quindi ad un alto numero di FP. Viceversa per finalità di conferma.

Da quanto riportato risulta chiaro che per giudicare il valore diagnostico di un test bisogna innanzitutto conoscere la prevalenza della malattia studiata e non sono sufficienti dati sperimentali ottenuti su casistiche con rapporto malati/sani lontano da quelli della realtà in cui si opera. Un errore frequente è quello di riferirsi a studi pilota con una prevalenza relativamente alta (per esempio 0.5) e quindi trasferire i risultati direttamente a situazioni cliniche con prevalenza molto più bassa, come nel caso di indagini di *screening* che di solito non hanno una prevalenza che supera lo 0.05. E' tuttavia possibile, applicando il teorema di Bayes, ricalcolare i valori predittivi del test a qualunque altra prevalenza: tenendo costante il numero dei malati si moltiplica il numero dei sani per un appropriato fattore in modo da diluire il numero dei malati riportandolo al rapporto con i sani della nuova realtà. Il fattore F in questione si calcola come segue:

$$F = [(1 - \text{nuova prevalenza}) - 1] \times (\text{malati/sani})$$

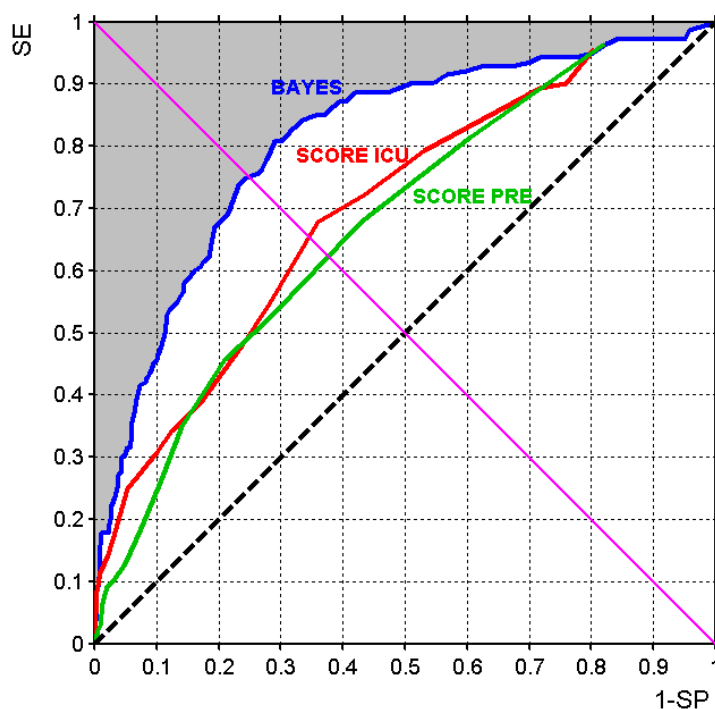
Così, nel caso si vogliano trasformare i valori predittivi dell'esempio precedente (Fig. 1), Unità Coronarica, prevalenza 0.44, ad una nuova prevalenza di 0.1 (Pronto Soccorso), il fattore sarà uguale a:

$$F = [(1 - 0.1) - 1] \times (132/168) = 7.07$$

Avremo così un totale di  $168 \times 7.07 = 1188$  sani e dovremo conseguentemente moltiplicare per 7.07 tutte le classi della popolazione sana e ricalcolare i valori predittivi con i nuovi dati.

**La curva ROC** (Receiver Operating Characteristics, caratteristiche operative del ricevente) è un diagramma in cui vengono riportati, per diversi *cut-off* in ordinata SE (frazione di VP rispetto al totale dei malati) ed in ascissa il complemento ad 1 di SP, cioè  $1 - SP$  (frazione di FP rispetto al totale dei sani). La costruzione della curva ROC mostra probabilmente meglio di ogni altro procedimento il compromesso tra SE e SP. Va osservato che, qualora le distribuzioni dei sani e dei malati non siano affatto sovrapposte, il *cut-off* ideale è un qualunque valore compreso fra le due distribuzioni e che in questo caso fortunato, ma purtroppo assai raro, nessun compromesso tra SE e SP è necessario.

In Figura 2 sono riportate le curve ROC relative a tre diversi modelli predittivi del rischio di insorgenza di complicanza post-operatoria in pazienti critici di terapia intensiva (Intensive Care Unit, ICU) dopo chirurgia cardio-toracica. In verde e rosso sono tracciate le curve ROC di due modelli a punteggio



**Figura 2.** Curve ROC di modelli predittivi di complicanza post-cardiochirurgica

intero (*score*) con variabili rispettivamente preoperatorie (PRE) e postoperatorie (ICU). In blu è mostrata la curva ROC di un modello predittivo bayesiano. Ad ogni soglia di decisione corrisponde una coppia di valori di SE e SP. Nel modello bayesiano sarà una probabilità compresa tra 0 e 1. La curva ROC parte e finisce sempre dagli angoli basso-sinistro e alto-destro. Il primo angolo coincide con SE nulla e SP massima, che corrisponde a considerare tutti i soggetti sani (ovviamente nessun FP). Il secondo angolo è relativo a SP massima e SP nulla, corrispondente a tutti i soggetti valutati come positivi (nessun FN).

Le informazioni che si possono trarre da una curva ROC, per quanto concerne la Medicina di Laboratorio o un modello predittivo (come nel caso di Fig. 2), sono le seguenti:

**Livello di validità di un esame o di bontà di adattamento di un modello predittivo.** Quanto più una curva si avvicina ai lati sinistro e superiore del diagramma, tanto maggiore sarà la validità del metodo diagnostico: l'angolo in alto a sinistra del diagramma rappresenta l'ottimo assoluto, indicando la combinazione di massima SE=1 e massima SP=1; viceversa, una curva ROC che si avvicini alla diagonale nera tratteggiata del diagramma di Fig. 2 indicherà un esame la cui validità è nulla; ciò si realizza quando le distribuzioni dei malati e dei sani sono completamente sovrapposte. Sotto tale diagonale la curva ROC non può esistere. Pertanto più la curva ROC è arcuata verso il punto di ottimo, migliore è il test decisionale. La quantificazione della validità di un esame può essere dedotta dall'**area sottesa alla curva ROC** stessa (Area Under ROC Curve, **AUC**): tale area sarà dunque compresa nell'intervallo tra 0.5 (diagonale nera tratteggiata), test non valido, e 1.0, test ideale; ai valori di area si può associare un'interpretazione qualitativa del livello di validità dell'esame mediante l'applicazione della scala seguente, empirica ma efficace: 0.9–1.0 eccellente, 0.8–0.9 buono, 0.7–0.8 discreto, 0.6–0.7 scarso, 0.5–0.6 nullo. Con riferimento alla Fig. 2 e al modello probabilistico di Bayes (linea blu), AUC è l'area bianca sottesa alla curva stessa e vale circa 0.81, cioè il modello diagnostico può essere giudicato buono. L'area grigia, essendo il complemento a 1 di AUC, varrà 0.19 e rappresenta l'errore diagnostico globale. I modelli a *score*, fornendo entrambi un'area AUC pari a circa 0.66 hanno scarso potere discriminante.

**Scelta del valore soglia ottimale per un esame.** La soglia ottimale, almeno teoricamente, per un esame è quella in corrispondenza alla quale la curva ROC è più vicina all'angolo in alto a sinistra; tale soglia ottimale corrisponde quasi sempre al punto della curva ROC con SE=SP identificato dall'intersezione con l'altra diagonale, color magenta in Fig. 2. Tutto ciò tuttavia va interpretato alla luce di considerazioni globali rispetto alla patologia indagata; per esempio, nel caso di patologie invalidanti ma curabili se diagnosticate tempestivamente (come l'ipotiroidismo congenito) la scelta del *cut-off* cadrà non sul valore teoricamente ottimale, ma si preferirà tenere il più alto possibile il livello di SE dell'esame perché nessun malato sfugga alla diagnosi (nessun FN) accettando di pagare lo scotto di una ridotta SP (alcuni FP).

**Confronti mediante curve ROC.** Una comparazione mediante l'area sottesa alla curva ROC può applicarsi ad esami e modelli diversi impiegati per la diagnosi di una patologia, alla validità dello stesso metodo diagnostico in situazioni patologiche differenti o in diversi stadi della stessa patologia, a sottogruppi di pazienti aventi la stessa patologia, ma che differiscono per altre caratteristiche quali l'età, il sesso, ecc.

Le curve ROC vennero impiegate per la prima volta durante la Seconda Guerra Mondiale per valutare la capacità dei tecnici addetti ai radar di distinguere sul monitor i segnali corrispondenti a un aereo nemico dai normali rumori di fondo. Un tecnico troppo "ansioso" faceva scattare l'allarme ad ogni segnale che compariva sul monitor; con questo comportamento nessun aereo nemico sfuggiva ma moltissimi allarmi erano a vuoto; il risultato può essere paragonato a quello di un test molto sensibile e poco specifico. Diversamente è il comportamento del tecnico "imperturbabile" che non si lascia impressionare dai piccoli segnali ed è molto più cauto nel fare scattare l'allarme che, quindi, difficilmente scatterà a vuoto (alta SP) ma qualche aereo nemico potrà sfuggire (bassa SE).

**Considerazioni finali.** Si può riassumere che i criteri di valutazione di ogni metodo diagnostico (esame di laboratorio o modello predittivo clinico) possono essere solitamente raffigurati come una piramide alla base della quale vi è la determinazione delle specifiche analitiche (in particolare riproducibilità e accuratezza). Successivamente vengono valutate le caratteristiche diagnostiche (sensibilità e specificità clinica), l'impatto clinico e organizzativo e infine il rapporto costo-efficacia. È sempre più evidente la necessità di seguire questa piramide gerarchica per rispondere non solo a chi ritiene vengano eseguiti troppi test, ma soprattutto per legare il metodo diagnostico stesso al ragionamento clinico e, ancor più, alla decisione clinica. È da sottolineare come attualmente tale valutazione venga di fatto eseguita per tutte le procedure diagnostiche innovative che vengono proposte e che, prima di essere introdotte nella pratica clinica, devono dimostrare non solo di essere valide dal punto di vista tecnico, ma soprattutto di possedere caratteristiche decisionali tali da giustificare, in termini di rapporto costo-beneficio, gli investimenti correlati. Tale valutazione andrebbe fatta anche per gli esami già consolidati per assicurare che, nel tempo, non siano divenuti obsoleti, superflui o non necessari.