SENSIBILITÀ E SPECIFICITÀ: OVVERO LA VERIFICA STATISTICA DI UN NUOVO METODO DIAGNOSTICO

O. Petrini, C. Limoni

Spesso il clinico, per diagnosticare la presenza di una grave malattia (ad esempio HIV, cancro, sclerosi multipla), è obbligato ad applicare test che sono sì affidabili e veloci, però spesso invasivi e talvolta pericolosi per il paziente. Un test di screening semplice e non invasivo, ma altrettanto affidabile, in questi casi, sarebbe altamente auspicabile. I medici sono sempre alla ricerca di nuove tecniche diagnostiche, e le ditte impegnate nella ricerca di tali metodi mettono regolarmente sul mercato kit di cui vantano la semplicità, affidabilità e sicurezza. Si tratta, in questi casi, di verificare, prima di passare ad un uso quotidiano del nuovo test, la sua affidabilità. Questo è possibile paragonando la sua efficienza (espressa normalmente attraverso la sua sensibilità e specificità) con quella del test corrente, chiamato spesso "Gold standard".

Il test diagnostico

Un test diagnostico deve rispondere ad alcuni requisiti fondamentali. Deve essere accurato, quindi fornire un risultato che rifletta lo stato effettivo del paziente. Inoltre i risultati devono essere affidabili, indipendenti quindi dalle condizioni di misurazione o dall'esperienza dello sperimentatore.

- Ci aspettiamo dunque che sia sensibile – che sia in grado cioè di identificare tutti i casi positivi, o, nella prassi quotidiana, la maggior parte di essi.
- Deve poi esser specifico – idealmente dovrebbe evidenziare solo i casi positivi; in pratica, ovviamente, la maggior parte dei casi che risultano positivi al test dovrebbero essere veramente malati.
- Al nuovo test, inoltre, si richiede di essere sicuro e se possibile meno invasivo di quelli già sul mercato.
- Infine, idealmente, la sua applicazione dovrebbe essere semplice e i suoi costi più ridotti o simili a quelli dei prodotti esistenti.

Un fattore addizionale secondario, ma estremamente importante per il successo del test nella pratica quotidiana, è poi la sua accettabilità socio-culturale: ad esempio,uno screening per il cancro del collo dell'utero, pur essendo di fondamentale importanza, può essere di difficile applicazione in alcune società, specialmente se il personale sanitario è prevalentemente maschile[1].

Alcune definizioni

Come possiamo definire l'efficienza di un test diagnostico? Di solito paragonando il suo rendimento (in inglese chiamato "performance" o "efficiency") con quello del "gold standard", il test cioè con cui 100% dei casi sono identificati correttamente come positivi o negativi. In pratica, tale test esiste solo molto raramente, e in generale si definisce come "gold standard" il metodo diagnostico comunemente accettato dalla comunità medica e scientifica. I risultati del nuovo metodo sono paragonati a quelli del gold standard per mezzo di quattro valori: la sensibilità ("sensitivity"), la specificità ("specificity"), il valore predittivo positivo ("positive predictive value", PPV) e il valore predittivo negativo ("negative predictive value", NPV). Vediamo come sono calcolati questi valori.

Supponiamo che in uno studio il nuovo metodo sia paragonato al gold standard in un numero N di pazienti. I risultati dello studio possono essere rappresentati nel modo seguente:

		gold standard: positivo	gold standard: negativo	totale
nuovo metodo	positivo	**a** (veri positivi)	**b** (falsi positivi)	**a+b**
	negativo	**c** (falsi negativi)	**d** (veri negativi)	**c+d**
	totale	**a+c**	**b+d**	**N**

Corrispondentemente possiamo definire i parametri seguenti:

Sensibilità (%): $\frac{a}{a+c}$ x 100

il numero di veri positivi diviso per il totale di pazienti con la malattia, espresso come percentuale.

Specificità (%): $\frac{d}{b+d}$ x 100

il numero di veri negativi diviso per il totale di pazienti sani, espresso come percentuale.

PPV (%): $\frac{a}{a+b}$ x 100

questo valore rappresenta la percentuale di tutti i soggetti veramente malati tra tutti quelli che sono risultati positivi al nuovo test.

NPV (%): $\frac{d}{c+d}$ x 100

questo valore rappresenta la percentuale di tutti i soggetti veramente sani tra tutti quelli che sono risultati negativi al nuovo test.

Un esempio

Supponiamo che si voglia introdurre una nuova metodica non invasiva, veloce e a buon mercato, per dia-

gnosticare la presenza di un cancro per la cui diagnosi si usa correntemente una tecnica invasiva e molto costosa. Lo sperimentatore esegue uno studio comparativo su 500 pazienti, applicando i due metodi a tutti i pazienti, con i risultati seguenti:

		tecnica corrente		
		positivo	negativo	totale
nuovo metodo	positivo	**45** (veri positivi)	**30** (falsi positivi)	**75**
	negativo	**5** (falsi negativi)	**420** (veri negativi)	**425**
	totale	**50**	**450**	**500**

Applicando le formule specifiche, otteniamo i risultati seguenti per l'efficienza del nuovo test:
sensibilità: [(45/50)×100]%= 90%;
specificità: [(420/450)×100]%= 90%;
PPV: [(45/75)×100]%= 60%
NPV: [(420/425)×100]%= 99%
Da questi valori possiamo dedurre che il nuovo test ha una buona efficienza e potrebbe rimpiazzare la tecnica tradizionale. Diciamo "potrebbe", in quanto diventa importante decidere se una sensibilità di 90% sia sufficiente per questo tipo di diagnosi, se sia cioè importante o no, da un punto di vista medico, rilevare il maggior numero possibile di casi e quindi sia necessaria una sensibilità maggiore di quella rilevata nello studio. Inoltre, visto che il metodo tradizionale potrebbe non essere così sensibile e specifico come si desidera, è consigliata una continuazione dello studio applicando altri metodi, per poter così verificare in modo più dettagliato i valori ottenuti con il nuovo test. Il valore PPV, inoltre, sembrerebbe indicare che il nuovo metodo tende a fornire un numero relativamente alto di falsi positivi. Anche questo deve essere valutato attentamente da un punto di vista medico. In altre parole, è meglio o accettabile diagnosticare una persona sana come malata e quindi iniziare un trattamento inutile piuttosto che il contrario? A questa domanda è il medico e non lo statistico che dovrà rispondere.

Sensibilità, specificità, PPV e NPV: tutti ugualmente importanti?

Come già accennato nella discussione dell'esempio precedente, la statistica può solamente fornire le basi per la decisione riguardo l'adozione o meno di un nuovo test diagnostico. L'esperienza del medico, legata alla convenienza o meno di accettare preferibilmente dei falsi positivi a dei falsi negativi, e quindi, rispettivamente, dei valori PPV o NPV bassi, rimane fondamentale per la decisione. Inoltre, non bisogna dimenticare che sensibilità, specificità, PPV e NPV sono delle misure operazionali che rispecchiano da una parte la validità del test (sensibilità e specificità), dall'altra la prevalenza effettiva della malattia nella popolazione. Un test altamente sensibile e specifico darà dei valori PPV deludenti nel caso di una bassa prevalenza della malattia. Ad esempio, nel caso del carcinoma della prostata in pazienti sessantenni, un test con sensibilità e specificità del 99% darà dei valori PPV bassi, intorno al 50%. In questo caso, però, il medico potrebbe preferire un alto numero di falsi positivi alla mancanza di trattamento di un paziente veramente malato. In sostanza, bisogna sempre trovare un equilibrio tra sensibilità e specificità di un test, secondo le conseguenze mediche derivanti dal mancare la diagnosi in alcuni casi o dal trattare dei pazienti sani. La soluzione ottimale potrebbe quindi essere l'adozione di un metodo in concomitanza con l'esecuzione di ulteriori analisi diagnostiche[2].

"Gold standard" o semplicemente test diagnostico accettato universalmente?

Vorremmo concludere con un breve cenno sull'adeguatezza dell'uso di un gold standard. Un documento pubblicato dall'agenzia americana FDA[3] rende attenti sul fatto che i concetti di specificità, sensibilità, PPV e NPV sono applicabili solo nel caso esista un vero gold standard, o almeno se ne possa "costruire" uno. Ad esempio, le emocolture sono da molti considerate lo standard per la diagnostica di batteriemie, ma è chiaro che alcune emocolture danno dei risultati negativi anche quando una batteriemia è presente. In casi come questo, il gold standard deve essere ridefinito, se possibile, usando test aggiuntivi e paragonando quindi i risultati ottenuti con il nuovo metodo con quelli derivanti da una combinazione di altre tecniche che permettano di garantire una diagnosi accurata ed affidabile. Nel caso ciò non fosse possibile, raccomandiamo di evitare l'uso dei parametri descritti sopra e di semplicemente indicare la concordanza percentuale tra il nuovo metodo e la tecnica diagnostica usata tradizionalmente. Questo evita di sopravvalutare il metodo esistente ed al medesimo tempo eventualmente penalizzare un nuovo metodo che, alla luce di nuove ricerche e conoscenze, potrebbe essere migliore di quello tradizionale ed eventualmente diventare il nuovo gold standard.

Orlando Petrini,
Istituto cantonale di microbiologia, Bellinzona
Costanzo Limoni,
Alpha 5-Biometrics, Riva San Vitale

Bibliografia

1 Lirri E. Uganda: Women shy away from cervical cancer screening. The Monitor, March 3, 2010. http://allafrica.com/stories/201003030625.html

2 Webb P, Bain C, Pirozzo S. Essential Epidemiology. An introduction for students and health professionals. Cambridge University Press, 2005: pp. 290-315.

3 FDA. Statistical guidance on reporting results from studies evaluating diagnostic tests. March 13, 2007. http://www.fda.gov/cdrh/osb/guidance/1620.pdf.