manager&evidence health

Technology assessment *Tanti i vantaggi di un'accorta valutazione di nuovi test*

Diagnostica sotto la lente

Oggi la maturazione del concetto di "prova di efficacia" e la limitazione delle risorse disponibili rendono indispensabile un adeguato e continuo check di bontà e sicurezza delle tecnologie

di **Antonino Cartabellotta** * *e* **Gianni Rossi** **

Il criterio fondamentale per decidere quando utilizzare un test diagnostico e quale è rappresentato dalla sua utilità clinica, definita come «la capacità intrinseca di un test, qualunque sia il risultato, di modificare la decisione del medico in senso diagnostico, prognostico o terapeutico». Infatti, è bene tener presente che, nonostante l'aggettivo, le informazioni ottenute dai test "diagnostici" possono essere utilizzate per diversi obiettivi:
- screening;
- diagnosi;
- follow up di malattia;
- prognosi;
- monitoraggio del trattamento (efficacia e/o effetti collaterali).

Screening

I test di screening hanno l'obiettivo di identificare, in pazienti asintomatici, patologie la cui morbilità e mortalità possano essere ridotte da un trattamento precoce. Perché ciò sia possibile, è necessario che la malattia, oltre a essere abbastanza diffusa da giustificarne la diagnosi precoce, sia gravata da un'importante morbilità/mortalità, se non trattata. Infine, il trattamento dei soggetti asintomatici deve essere più efficace rispetto a quello di pazienti inizialmente sintomatici.
Lo screening deve essere in grado di modificare favorevolmente

Il punto

Sackett definisce la diagnosi come «... un processo critico che etichetta il paziente, classifica le sue malattie, accerta (e talora segna) la prognosi, e suggerisce - con una confidenza spesso non fondata - un trattamento specifico allo scopo di determinare più beneficio che danno».
Ciò che sia razionale per una rigorosa valutazione dei test diagnostici è stato spesso oggetto di discussioni, ma oggi la limitazione delle risorse economiche e la maturazione nei clinici e negli amministratori sanitari del concetto di "prova di efficacia", rende indispensabile un'adeguata e continua valutazione delle tecnologie diagnostiche relativamente alla loro efficacia, sicurezza e costo-efficacia.
Dopo aver definito il concetto, gli ambiti di applicazione e le attività internazionali di technology assessment (Ta) - si veda Il Sole-24 Ore Sanità e Management n. 6 del 2000 - ci occuperemo in questo numero delle metodologie per valutare l'efficacia delle tecnologie diagnostiche.

** Gimbe - Gruppo italiano per la medicina basata sulle evidenze*
*** Unità ospedaliera di anestesia e rianimazione, azienda Usl di Imola*

la storia naturale della malattia, non solo anticipando il momento dalla diagnosi (lead time bias), ma migliorando la sopravvivenza e/o la qualità di vita dei pazienti. Altrimenti, le uniche conseguenze della diagnosi precoce saranno gli eventuali danni fisici e psicologici dello screening e l'aumento dei costi per il sistema sanitario. Secondo tali presupposti, il test di screening, generalmente unico e mirato all'identificazione di una sola malattia, può essere applicato sia alla popolazione generale (screening neonatale per fenilchetonuria e ipotiroidismo nei neonati), sia in particolari sottogruppi a rischio (mammografia nelle donne di età superiore a 49 anni, HbsAg nelle donne gravide).

La "distorsione" del concetto di screening ha progressivamente configurato il "case finding". Con risultati deludenti

La "distorsione" del concetto di screening, attraverso l'utilizzo dei cosiddetti "esami di routine", ha progressivamente configurato il "case finding", ossia il tentativo di identificare una malattia asintomatica, o un dato patologico, grazie all'esecuzione di un certo numero di test. In tali condizioni, la speranza di identificare «troppe malattie in tutti i pazienti» ha dato risultati molto deludenti: infatti per numerosi "test di routine" (Ecg, radiografia del torace, emocromo, formula leucocitaria, profili biochimici, esame delle urine), esistono adeguate evidenze che la loro richiesta senza sospetto clinico è di scarsa utilità, perché raramente conduce a nuove diagnosi e quasi mai a nuovi trattamenti. Inoltre, i risultati falsamente positivi, la cui frequenza cresce esponenzialmente con il numero dei test richiesti, scatenano una cascata di ulteriori indagini e "peripezie diagnostiche", configurando la "sindrome di Ulisse", fonte di apprensione per il paziente e causa di incremento dei costi.

Diagnosi

La strategia esaustiva concepisce il processo diagnostico in due fasi: nella prima, il medico costruisce, in maniera acritica, il data base del paziente; quindi, sulla base di una quantità rilevante di dati raccolti, inizia il ragionamento diagnostico. Tale strategia, tipica del medico principiante e invalidata da numerosi studi, dovrebbe essere definitivamente abbandonata.

Il metodo che meglio si accorda con il concetto di efficienza diagnostica, è quello ipotetico-deduttivo: i primi dati del paziente (sintomi/segni principali, età, sesso, fattori di rischio) suggeriscono precocemente una o più ipotesi, che indirizzano l'ulteriore ricerca dei dati, sia clinici, sia di laboratorio e strumentali. La loro progressiva acquisizione rimodella l'ipotesi iniziale e indirizza la ricerca di nuovi elementi sino alla formulazione della diagnosi corretta. L'ipotesi iniziale è comunque molto flessibile, soggetta a continui aggiustamenti ed eventualmente a essere sostituita da altre ipotesi se emergono nuovi dati (secondo Kassirer la generazione delle ipotesi diagnostiche è più simile alle sequenze di un film che a un singolo fotogramma). Secondo la strategia ipotetico-deduttiva, i test sono utili quando il grado di certezza diagnostica che deriva dalla storia e dall'esame obiettivo è insufficiente per prendere decisioni cliniche. Tuttavia, la capacità intrinseca di un test di ridurre i margini di incertezza è ampiamente variabile, per cui la sua richiesta a scopi diagnostici deve risultare dalla valutazione di numerosi fattori:

- probabilità pre test di malattia. Se la probabilità di malattia prima di eseguire un test diagnostico è troppo bassa o troppo alta, il suo risultato difficilmente modificherà la probabilità post test e, quindi, la decisione clinica;
- sensibilità e specificità del test. Sono caratteristiche proprie di ciascun test, che riflettono le informazioni attese in pazienti con o senza la malattia presa in esame (si veda infra). Un risultato positivo non fornisce la certezza assoluta di malattia, considerato che può esserlo sia in pazienti malati, proporzionalmente alla sua sensibilità, sia in soggetti sani, in relazione inversa alla sua specificità; analogamente, la negatività di un test, proporzionale alla sua specificità in un soggetto sano e in relazione inversa alla sensibilità nei soggetti malati, non è sempre associata all'assenza di malattia. Pertanto, i test molto sensibili sono utili per escludere, se negativi, una malattia, mentre quelli a elevata specificità, quando positivi, servono per confermare l'ipotesi diagnostica.

La capacità intrinseca di una "prova" di ridurre i margini di incertezza è molto variabile per cui esso va attentamente meditato

Considerato che per numerosi test di laboratorio i risultati si dispongono lungo una scala di valori numerici (a esempio, glicemia, creatininemia), è opportuno distinguerli in tre gruppi: quelli che rientrano ampiamente in un ran-

ge di normalità, che indicano verosimilmente l'assenza di malattia; quelli nettamente patologici e una fascia intermedia, da interpretare adeguatamente in relazione al contesto clinico del paziente. Infatti, questo gruppo comprende numerosi risultati falsamente positivi che, per un fenomeno puramente statistico (regressione verso la media), risultano normali a una successiva determinazione;

• accettabilità del test in relazione al vantaggio diagnostico. Un'attenta valutazione del rischio, dell'invasività e dei costi del test diagnostico può suggerire l'ipotesi di prendere una decisione terapeutica senza eseguirlo: a esempio, alla luce delle nuove evidenze sull'efficacia delle eparine a basso peso molecolare nei pazienti con sospetta embolia polmonare (*Low molecular weight heparin in pulmonary embolism*, in Clinical Evidence, Bmj Publishing Group, 2000), è opportuno iniziare il trattamento anche a un livello intermedio di probabilità diagnostica;

• effetti del test sulla decisione clinica. Se si prevede che questa non venga modificata dal risultato del test è inutile eseguirlo, specie se costoso e/o invasivo. A esempio, in pazienti con metastasi da neoplasia di origine sconosciuta, le evidenze indicano discrete probabilità di identificare la sede primitiva del tumore, ma prospettive terapeutiche molto scarse. Infatti, in una valutazione cumulativa su 2.283 pazienti (Shapira D.V., et al., *Arch Intern Med*, 1995;155:2050-4), la sede primitiva del tumore è stata identificata nel 21% dei casi, ma solo 25 pazienti (1,1%) erano affetti da neoplasie potenzialmente curabili (linfomi, tumori del testicolo).

Solo un'accurata conoscenza della storia naturale della malattia permette un adeguato utilizzo dei test

Follow up di malattia

L'identificazione di nuove alterazioni non rilevabili soggettivamente (dal paziente) né obiettivamente (dal medico), può richiedere l'utilizzo di test diagnostici. Tuttavia, solo un'accurata conoscenza della storia naturale della malattia può consentire un loro adeguato utilizzo, sia nella selezione delle indagini, sia nella frequenza del follow up.

Prognosi

L'utilizzo dei test diagnostici fornisce anche informazioni prognostiche, anzi talvolta essi vengono utilizzati con questo obiettivo primario. Infatti se Archie Cochrane insegna che bisogna «diagnosticare solo ciò che è possibile trattare», talvolta la necessità di fare una prognosi adeguata, specie se è il paziente a richiederla, impone l'esecuzione di test invasivi e/o costosi (a esempio, per ragioni assicurative e/o pensionistiche).

Monitorare i risultati di un trattamento ha il doppio obiettivo di verificare la risposta terapeutica e di rilevare effetti collaterali

Monitoraggio del trattamento

Monitorare gli effetti di un trattamento ha un duplice obiettivo: verificare se la risposta terapeutica è adeguata e rilevare precocemente l'eventuale comparsa di effetti collaterali. Spesso gli indici di monitoraggio della risposta terapeutica sono clinici: a esempio, in pazienti con polmonite pneumococcica in trattamento antibiotico è inutile ripetere ogni 2-3 giorni una radiografia del torace, visto che la risposta clinica è più precoce del miglioramento radiologico.

Un singolo test può essere impiegato sia per il monitoraggio della terapia che degli effetti collaterali (a esempio, Inr nei pazienti che assumono anticoagulanti orali), ma più frequentemente sono necessari più test, contemporaneamente o in momenti differenti.

La valutazione delle tecnologie diagnostiche

Un'adeguata valutazione delle evidenze relative ai test diagnostici richiede la conoscenza di alcuni standard metodologici che gli studi devono possedere già al momento della loro pianificazione.

Criteri di validità degli studi che misurano l'accuratezza dei test diagnostici

Nonostante le difficoltà a pianificare studi per misurare l'accuratezza dei test diagnostici, esistono ormai alcuni criteri condivisi per giudicarne la qualità metodologica (Jaeschke R., et al. Jama 1994; 271:389-91 e 703-7):

• confronto cieco e indipendente del test diagnostico con un standard (gold) di riferimento. Un nuovo test deve essere confrontato con una metodica diagnostica standard (con sensibilità e specificità ~ 100%). A esempio, l'accuratezza diagnostica di un Ecg da sforzo può essere determinata confrontandone i risultati con quelli della coronarografia (gold standard). Spesso, negli studi clinici, tale confronto viene eseguito

solo nei pazienti con risultato positivo (o negativo), a seconda delle circostanze, con conseguente distorsione (work up bias) degli indici di accuratezza: infatti, se a eseguire la coronarografia sono solo i pazienti con il test da sforzo positivo, la sensibilità del test risulterebbe falsata, poiché un certo numero di pazienti con test da sforzo negativo (falsi negativi), potrebbe avere lesioni coronariche significative. Quando il gold standard è costituito da un'indagine che presenta rischi elevati, può essere sostituito da un adeguato monitoraggio clinico che non sempre costituisce un indice affidabile, considerato che durante il periodo di follow up gli eventi clinici attesi possono non manifestarsi (specie per le malattie croniche). Infine, per evitare che venga influenzata l'interpretazione del test in studio, la sua valutazione deve essere effettuata da personale medico che non è a conoscenza ("cieco") dello standard di riferimento;

- rappresentatività del campione di pazienti su cui viene sperimentato il test. Il test deve essere valutato in un campione di pazienti che includa un appropriato spettro di malattia: forme lievi e forme gravi, malattia trattata e non trattata, soggetti affetti da condizioni comunemente "confuse" con la malattia d'interesse. Infatti, gli indici di accuratezza risentono sensibilmente della variabilità della popolazione su cui vengono sperimentati e, successivamente, applicati. Pertanto, se il campione di soggetti studiati non è sufficientemente rappresentativo (spectrum bias), l'utilità del test può essere limitata. A esempio, i primi studi che valutavano l'antigene carcinoembrionario (Cea) come marker precoce di carcinoma del colon riportavano entusiasticamente un'elevata sensibilità del test, proposto "candidato ideale" per lo screening di malattia. In realtà questi studi includevano solo pazienti in fase avanzata di malattia e, quando venne sperimentato su un campione rappresentativo, la sensibilità del Cea nelle fasi iniziali di malattia (quando avrebbe un'utilità rilevante per la diagnosi precoce), era inferiore al 3%!;
- descrizione della riproducibilità dei risultati del test e della sua interpretazione. Considerato che spesso il risultato di un test dipende dalla variabilità dell'osservatore (imaging, istologia) o dalla standardizzazione delle procedure di laboratorio, la sua riproducibilità (si veda infra) è un elemento indispensabile per valutarne gli indici di accuratezza diagnostica;
- definizione accurata del concetto di "risultato normale" del test;
- precisione dei risultati per l'accuratezza del test. Considerata la variabilità degli indici di accuratezza in relazione alla numerosità del campione di pazienti studiato e alla prevalenza di malattia, i limiti di confidenza dovrebbero sempre essere riportati;
- determinazione dell'utilità del

Valutazione di una tecnologia diagnostica/ 1

	MALATTIA PRESENTE	MALATTIA ASSENTE
Test positivo	**Vero positivo (A)**	**Falso positivo (B)**
Test negativo	**Falso negativo (C)**	**Vero negativo (D)**

Valutazione di una tecnologia diagnostica/ 2

CARATTERISTICA	SIGNIFICATO	FORMULA
Sensibilità	**In che misura il test riesce a identificare i soggetti malati?**	**a/(a+c)**
Specificità	**In che misura il test riesce a escludere i soggetti sani?**	**d/(b+d)**
Valore predittivo +	**Se un soggetto ha il test positivo, qual è la probabilità che sia realmente malato?**	**a/(a+b)**
Valore predittivo -	**Se un soggetto ha il test negativo, qual è la probabilità che sia realmente sano?**	**d/(c+d)**
Accuratezza	**Quale percentuale della totalità dei test eseguiti ha fornito risposte corrette (veri positivi + veri negativi)**	**(a+d)/ (a+b+c+d)**
Likelihood ratio +	**In che misura è più probabile trovare un test positivo in un soggetto malato, rispetto a un soggetto sano?**	**Sensibilità/ (1-specificità)**
Likelihood ratio -	**In che misura è più probabile trovare un test negativo in un soggetto sano, rispetto a un soggetto malato?**	**(l-sensibilità)/ specificità**
+ del test positivo - del test negativo		

Valutazione dei giudizi di 100 accusati di omicidio/ 1

VERDETTO	VERITÀ	
	Colpevoli	Innocenti
Condannati	20	15
Assolti	10	55

Valutazione dei giudizi di 100 accusati di omicidio/ 2

INDICI DI ACCURATEZZA	FORMULA	VALORE
Sensibilità	a/(a+c)	67%
Specificità	d/(b+d)	79%
Valore predittivo del test positivo	a/(a+b)	57%
Valore predittivo del test negativo	d/(c+d)	85%
LR +	Sensibilità/ (1-specificità)	3,11
LR -	(l-sensibilità)/ specificità	0,42
Accuratezza	(a+d)/(a+b+c+d)	75%
Probabilità pre-test (o prevalenza)	(a+c)/(a+b+c+d)	30%

COMMENTI

SENSIBILITÀ = 67%
- La Corte condanna giustamente un omicida nel 67% dei casi
- Il 33% dei colpevoli viene erroneamente assolto

SPECIFICITÀ = 43%
- La Corte assolve giustamente il 79% degli innocenti
- Il 21% degli innocenti viene erroneamente condannato

VALORE PREDITTIVO DEL TEST POSITIVO = 57%
- Se la Corte ha ritenuto colpevole un imputato, la probabilità che questo lo sia realmente è del 57%
- Il 43% dei condannati è innocente

VALORE PREDITTIVO DEL TEST NEGATIVO = 85%
- Se la Corte ha ritenuto innocente un imputato, la probabilità che questo lo sia realmente è dell'85%
- Il 15% degli assolti è colpevole

ACCURATEZZA = 75%
- La Corte emana un giudizio corretto nel 75% dei casi

test:
- ruolo del test in un percorso diagnostico: a esempio, se sostituisce o integra altre indagini;
- report di tutti gli esiti rilevanti: ritardo nell'instaurazione di un trattamento, complicanze, impatto psicologico del test;
- descrizione accurata delle tecniche per eseguire il test;
- presentazione dei risultati indeterminati: nella pratica clinica quando il risultato di un test è equivoco o indeterminato e pertanto "non diagnostico", sono necessarie ulteriori indagini; pertanto un'elevata frequenza di tali risultati rende un test di limitata efficacia, perché gli indici di accuratezza sono spesso calcolati sui risultati nettamente positivi o negativi.

Le misure di accuratezza dei test diagnostici

Tutti i dati clinici, di laboratorio e strumentali hanno un grado molto variabile di accuratezza, cioè sono più o meno attendibili rispetto alla situazione clinica che sono chiamati a rappresentare.

In particolare, qualunque test diagnostico possiede un set di caratteristiche che riflettono i risultati attesi in soggetti affetti o meno dalla malattia.

Poiché la maggior parte dei test non sono perfetti, c'è un variabile grado di sovrapposizione tra pazienti con e senza malattia, che determina un'erronea classificazione di soggetti sani (etichettati come malati = falsamente positivi) e di soggetti malati (etichettati come sani = falsamente negativi).

L'accuratezza dei test diagnostici risulta da tre componenti:
- riproducibilità (del test e della sua interpretazione);
- sensibilità: frequenza con cui il test è positivo nei pazienti con la malattia ipotizzata;
- specificità: frequenza con cui il il test è negativo al di fuori di quella malattia.

• Riproducibilità. Tutti i dati clinici (della storia, dell'esame fisico, dei test di laboratorio e strumentali) hanno una quota di non-concordanza: spesso, tale dato non è trascurabile bensì particolarmente importante per l'interpretazione dei test di imaging. La riproducibilità di un test è «la percentuale dei giudizi concordanti sul totale dei giudizi relativi a un dato diagnostico»; in pratica risulta più utile il complemento della riproducibilità (100% - riproducibilità), ossia la percentuale di giudizi non concordanti (inter observer variation) che è inversamente proporzionale:
- all'esperienza (del clinico che visita il paziente, del radiologo che

Valutazione delle tecniche di imaging

EFFICACIA TECNICA
- Grado di risoluzione di coppie di linee **- Gradazione della scala dei grigi** **- Quantità delle chiazze** **- Nitidezza** **- Variabilità analitica**
EFFICACIA DI ACCURATEZZA DIAGNOSTICA
- Percentuale di diagnosi corrette in una serie di casi **- Sensibilità e specificità del test** **- Rilevazioni della sommità della curva Roc o dell'area sottostante la curva**
EFFICACIA NELLA FORMULAZIONE DELLA DIAGNOSI
- Percentuale in una serie di casi in cui il test ha contribuito alla formulazione della diagnosi **- Variabilità della probabilità di distribuzione delle diagnosi differenziali** **- Differenza tra le stime soggettive dei clinici riguardo le informazioni sulle probabilità di diagnosi pre e post test** **- Stima empirica, in una serie di casi, del rapporto di verosimiglianza sia per il test negativo sia per quello positivo**
EFFICACIA TERAPEUTICA
- Percentuale di volte, in una serie di casi, in cui il test è stato utile per la gestione del paziente: **• il risultato del test ha evitato procedure mediche o ha modificato la terapia pianificata prima del test** **• i medici hanno dichiarato l'intenzione di modificare le scelte terapeutiche in seguito al risultato del test**
EFFICACIA SUGLI ESITI ASSISTENZIALI
- Percentuale dei pazienti migliorati a seguito dell'esecuzione del test, rispetto ai pazienti che non hanno eseguito il test **- Morbilità (o procedure) evitate in seguito all'esecuzione del test** **- Variazioni negli anni di vita "aggiustati" per la qualità (Qaly)** **- Costi risparmiati dal test per Qaly**
EFFICACIA SOCIALE
- Vantaggio nell'utilizzo del test dimostrato da analisi costo-beneficio, costo-efficacia, costo-utilità

legge la Tac); in fase di adozione di tecnologie diagnostiche innovative questo fattore è di grande rilevanza, poiché la scarsa esperienza degli operatori può risultare in una bassa riproducibilità del test;
- alla semplicità del quesito: a esempio, è molto più facile - e quindi riproducibile - il giudizio di "normalità" o "anormalità" di un esame radiologico del torace; lo è assai meno la definizione di natura dell'eventuale lesione (neo-plastica, tubercolare, micotica o altro);
- alla qualità dell'esame.

• Sensibilità e specificità. Nella sua forma più semplice, la valutazione di una tecnologia diagnostica considera due dicotomie: la presenza/assenza di malattia e la positività/negatività del test. Per la comprensione di tali relazioni, risulta di grande utilità la tabella a pagina 39, da cui si ricavano tutti gli indici di performance di un test.
La presenza o l'assenza di malattia viene determinata in relazione ai risultati del gold standard diagnostico, la cui applicazione rappresenta uno dei criteri di qualità degli studi sui test diagnostici.

Abbiamo voluto esemplificare tali concetti riportando la valutazione dei giudizi di una Corte nei confronti di 100 soggetti accusati di omicidio (tabella 1). Ovviamente, nella tabella 2 i colpevoli corrispondono ai malati, gli innocenti ai sani, i condannati al test positivo e gli assolti al test negativo.

La gerarchia dell'efficacia diagnostica

La stima dell'accuratezza diagnostica di un test rappresenta solo uno degli step nella valutazione delle tecnologie diagnostiche. Infatti, se il loro fine ultimo non è migliorare il grado di accuratezza diagnostica, ma gli esiti assistenziali, è necessario prendere in considerazione altri studi. Infatti, anche test diagnostici dall'accuratezza elevata potrebbero offrire un contributo minimo, o nullo, nel migliorare gli esiti assistenziali. Di contro, l'utilità clinica di test diagnostici molto meno accurati può essere più elevata. Pertanto, secondo il punto di vista della ricerca sugli esiti (outcomes research), è necessario prendere in considerazione il beneficio globale sulla salute che deriva dall'uso delle tecnologie diagnostiche. A esempio, nonostante l'ecografia addominale sia un test molto accurato (71-95%) per la diagnosi di appendicite acuta, il suo utilizzo non riduce le complicanze, né la durata della degenza (Douglas Cd, et al. Bmj 2000;321:1-7).
A questo scopo, Fryback e Thornbury hanno elaborato un modello gerarchico per valutare l'efficacia delle tecniche di imaging, i cui concetti sono estrapolabili a qualunque tecnologia diagnostica (tabella in pagina). Questa valutazione aderisce strettamente ai criteri dell'evidence based medicine, perché la tecnologia diagnostica "guadagna" il livello gerarchico superiore solo se studi adeguati forniscono le prove necessarie. ●