Rassegne sistematiche degli esami di laboratorio: stato dell'arte

R. M. Dorizzi
Laboratorio Analisi Chimico Cliniche ed Ematologiche, Azienda Ospedaliera di Verona

Generalità

Le rassegne sistematiche (RS) dei risultati della ricerca rappresentano delle attività molto importanti. Il loro scopo è chiaro: integrare in modo efficiente le conoscenze e fornire agli operatori sanitari, ai ricercatori ed agli amministratori dati per rendere razionale il processo decisionale in sanità.
Ogni anno la letteratura biomedica si arricchisce di quasi 20.000 volumi e di circa 2 milioni di articoli su oltre 30.000 giornali. È stato stimato che questi articoli, messi uno sopra l'altro, formerebbero una montagna di 500 metri di altezza[1]. Questo fenomeno non accenna a rallentare ed anzi sta accelerando facendo sì che in molte aree è semplicemente impossibile per il singolo leggere, valutare criticamente e fare una sintesi delle conoscenze attuali. Le rassegne sono pertanto diventate uno strumento essenziale per chi vuole mantenersi aggiornato sui temi di interesse generale e particolare. Le RS chiariscono se i risultati ottenuti sono "consistenti e se possono essere generalizzati in popolazioni, ambienti e terapie diverse". Rendendo conto dello "stato dell'arte" sull'argomento rappresentano anche uno strumento che consente di individuare i campi in cui le evidenze non sono sufficienti e sono necessari ulteriori studi.
Secondo la Mulrow[2]

a) La RS, attraverso una esame critico, la valutazione e la sintesi separa "la sterpaglia" senza significato, non attendibile o inutile dagli studi importanti e critici meritevoli di riflessione.
b) È sempre maggiore il numero di "decisori", non solo clinici ma anche amministratori e politici sanitari che necessitano di valutazioni critiche di informazioni sanitarie di vario tipo. Il ricercatore impiega la RS per selezionare, sostenere e perfezionare ipotesi, riconoscere errori presenti nei lavori precedenti, valutare le dimensioni del campione necessarie ed anticipare effetti collaterali importanti meritevoli di ulteriori studi. Chi si occupa di politica sanitaria usa la RS per preparare linee guida e regolamenti relativi all'impiego di esami diagnostici e strategie terapeutiche
c) La RS rappresenta una tecnica scientifica efficiente e, nonostante sia talvolta impegnativa e richieda tempo, è meno costosa e più rapida rispetto ad uno studio sperimentale. In un'epoca in cui competitività ed efficienza sono sempre più importanti inoltre la RS indica quando una via è già stata battuta. Il continuo aggiornamento di rassegne, come nel Cochrane Database of Systematic Reviews, consente di accorciare l'intervallo tra la scoperta medica e l'implementazione della strategia diagnostica o terapeutica conseguente. Un esempio è quello costituito dei benefici della meta-analisi è illustrato nella Figura 1 che mostra gli Odds Ratio e gli intervalli di confidenza al 95% di 33 trial che confrontavano la streptokinasi per via endovenosa rispetto al placebo in pazienti ospedalizzati per infarto acuto del miocardio. La parte sinistra della figura indica come la terapia con streptokinasi era favorevole in 25 dei 33 trial ma solo in 6 in modo significativo. La stima complessiva, riportata alla base del grafico, mostra come il trattamento aveva un effetto significativo. La parte destra dell'immagine mostra come, se fosse stata eseguita una meta-analisi cumulativa dopo la pubblicazione di ognuno degli articoli, l'effetto avrebbe raggiunto una significatività < 0.05 nel 1971, < 0.01 nel 1973 e < 0.001 nel 1977.
d) La RS consente una generalizzabilità molto superiore al singolo articolo poiché i diversi articoli sono simili ma divergono per criteri di selezione, definizione della malattia, variazioni del trattamento e disegno dello studio.
e) Correlata alla generalizzabilità è la consistenza degli effetti in malattie diverse con analogie fisiopatologiche ed in termini di fattori di rischio.
f) La RS può spiegare inconsistenze e contrasti nei dati. Si può indagare perché una terapia è efficace in un contesto e non in un altro o perché uno studio produce dei risultati discordanti dagli altri.

Corrispondenza a: Azienda Ospedaliera di Verona, Piazzale A. Stefani, 37126 Verona - dorizzi@easyasp.it

g) La RS aggiunge potenza allo studio singolo. È stato scritto che la RS è simile ad una "torre di forza statistica che consente al ricercatore di salire sul corpo delle prove, esaminare il paesaggio e progettare nuove direzioni"[2]. Lo stesso logo della Cochrane Collaboration (Fig. 2) ci ricorda il caso classico di 7 trial che valutavano gli effetti di un breve ciclo di terapia cortisonica in donne a rischio di parto prematuro. Solo due trial davano dei risultati significativi, ma il pool di tutti gli studi aumentava le dimensioni e quindi la forza del campione. Questo vantaggio è particolarmente importante quando si indagano condizioni poco frequenti. L'aumento della numerosità consentiva inoltre di aumentare la precisione della stima, come è mostrato nella Figura 1, dalla riduzione dell'intervallo di confidenza.
h) La RS è meno influenzata dalle idiosincrasie, dalla predilezione o anche semplicemente dalle opinioni del ricercatore. Impiega infatti metodi espliciti e quindi replicabili ed interpretabili. Corregge infine il ritardo nella applicazione di quanto è dimostrato nella pratica clinica. La Figura 3 mostra come la lidocaina continuava a trovare proponenti nella profilassi per ridurre la mortalità dopo l'infarto anche dopo che 15 trial randomizzati avevano dimostrato prima del 1990 che non aveva effetto.

Le revisioni sistematiche degli esami diagnostici

A differenza di quanto accade per quanto riguarda i farmaci, l'introduzione di un esame diagnostico non deve possedere dei requisiti formali per essere introdotto in routine. Durante gli ultimi 25 anni questa situazione ha coinciso con l'accelerazione dello sviluppo tecnologico e con l'introduzione di un numero straordinario di analizzatori e di metodiche. Si verifica sempre più frequentemente che questi metodi, una volta commercializzati, si dimostrano deludenti. Casi classici sono quelli dell'antigene carcinoembrionario per la diagnosi di cancro del colon-retto, del test al desametasone per la diagnosi di depressione, della tecnica dell'immunofluorescenza per la diagnosi della malattia di Lyme e della scintigrafia con iodio 125 nelle diagnosi di trombosi venosa profonda[3].
Un ulteriore limite è costituito dal fatto che spesso la ricerca in questo ambito non è concentrata su un particolare ambito ristretto o su una particolare patologia o gruppi di patologie ma originano dalla necessità di chiarire un problema segnalato o di risolvere uno specifico problema clinico. Poiché il primo intervento cruciale nell'affrontare una qualunque patologia da parte di un clinico è quello di porre una ipotesi di diagnosi, vale a dire di interpretare i sintomi ed i segni, è vitale superare i limiti metodologici dell'impiego degli esami.
Per esame diagnostico si intende la raccolta di informazioni con lo scopo di chiarire il carattere e la prognosi della condizione del paziente; ne deriva che, oltre alle caratteristiche dell'esame, deve essere sempre considerato il quesito specifico a cui l'esame deve dare risposta. È fondamentale avere bene presente che un esame deve essere valutato secondo gli obiettivi per i quali è stato eseguito e richiesto:

Figura 1. Meta-analisi convenzionale e cumulativa di 33 trial dedicati all'impiego di streptokinasi nell'infarto acuto del miocardio. Odds ratio ed intervallo di confidenza al 95% espressi in logaritmo[2].

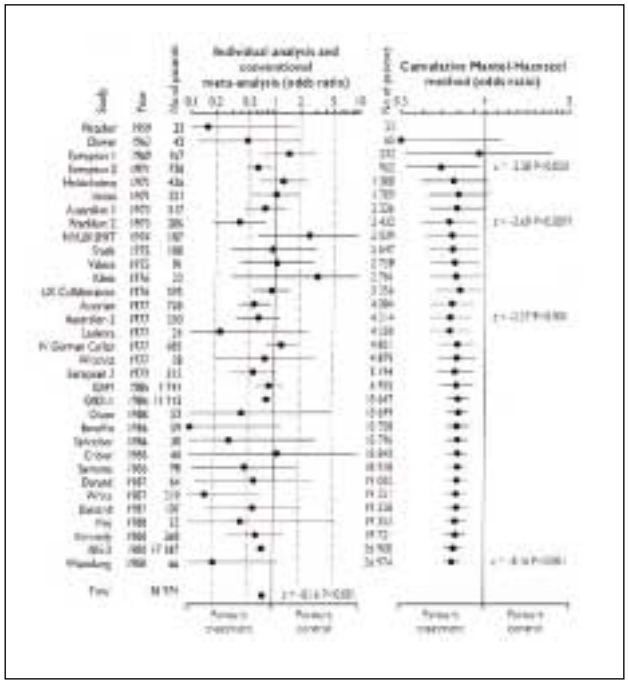


Figura 2. Logo della Cochrane Collaboration.

1) Aumentare la certezza della presenza o assenza di una malattia. Questo può essere ottenuto solo se la capacità discriminante della malattia è sufficiente. Recentemente il gruppo del Dipartimento di Medicina Generale di Maastricht [4,5] ha riassunto bene i parametri rilevanti in questo contesto mediante l'esempio dell'efficacia diagnostica dell'esame radiologico nella diagnosi di frattura (Tabella 1).

PREMESSE:
La SENSIBILITÀ è la probabilità di un risultato positivo in un paziente affetto dalla patologia D (nell'esempio della Tabella 190/200 = 0.95).
La SPECIFICITÀ è la probabilità di un risultato negativo in un soggetto senza la patologia D (720/800 = 0.90).

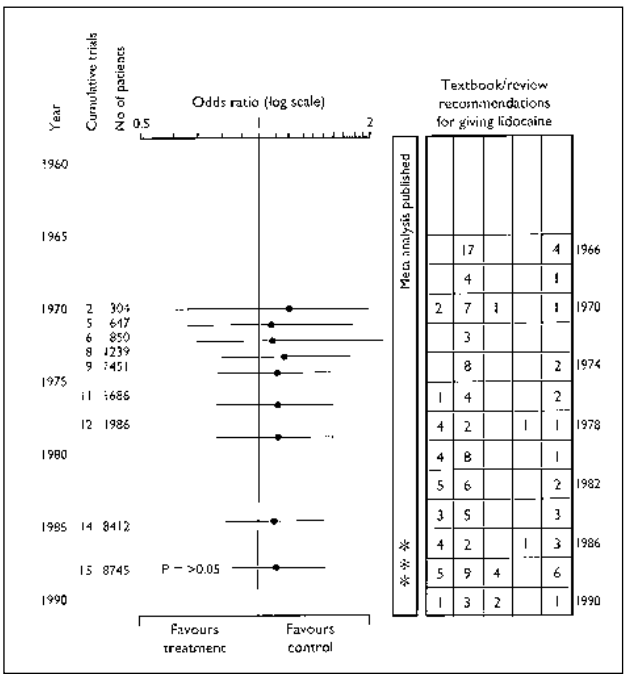
Tabella I. Parametri comuni usati per misurare la capacità discriminante dell'esame diagnostico T per la malattia D; l'esempio usato è quello della valutazione clinica di un trauma della caviglia usando la radiografia come riferimento.

	Risultato della radiografia (D)		
Risultato valutazione clinica (T)	Positivo (frattura)	Negativo (non frattura)	Totale
Positivo (frattura)	190	80	270
Negativo (non frattura)	10	720	730
Totale	200	800	1000

Il VALORE PREDITTIVO di un risultato POSITIVO (PPV) è la probabilità di malattia nei soggetti con un risultato dell'esame positivo (190/270 = 0.70)
Il VALORE PREDITTIVO di un risultato NEGATIVO è la probabilità di assenza di malattia nei soggetti con un risultato dell'esame negativo (720/730 = 0.99)

La capacità discriminante dell'esame è buona quando la differenza tra valore predittivo posteriore (o probabilità post-test) e la probabilità a priori o pretest (che si può indicare anche come prevalenza della malattia; 200/800 = 0.2 nell'esempio) è elevata.
Il QUOZIENTE DI PROBABILITÀ (LIKELIHOOD RATIO, LR) del risultato di un esame è il rapporto tra la probabilità di un risultato di un esame Tx in un paziente con la malattia D e la probabilità di un risultato in un paziente senza la malattia.

Figura III. Meta-analisi per anno di pubblicazione o trial randomizzati controllati dell'impiego di lidocaina nell'infarto acuto del miocardio[2].



LR POSITIVO=
= (SENSIBILITÀ/1-SPECIFICITÀ)=
= 190/200/1-(720/800) =9.5

=LR NEGATIVO=
=(1-SENSIBILITÀ/SPECIFICITÀ)=
=1-190/200/720/800) =0.06

Tabella II. Capacità discriminante di alcuni test diagnostici espressi in sensibilità, specificità, quoziente di probabilità positiva e negativa ed Odds Ratio[7]

				Quoziente di probabilità		
Esame	**Malattia**	**Sens (%)**	**Spec (%)**	**Positivo**	**Negativo**	**Odds Ratio**
ECG sforzo	Stenosi coronarica	65	89	5.9	0.39	15.0
Scintigrafia tallio	Stenosi coronarica	85	85	5.7	0.18	32.1
Ecografia	Cancro pancreas	70	85	4.7	0.35	13.2
TAC	Cancro pancreas	85	90	8.5	0.17	51.0
Angiografia	Cancro pancreas	75	80	3.8	0.31	12.0
VES > 28 mm/1 h	Cancro	78	94	13.0	0.23	56.0
VES > 28 mm/1 h	Infiammazione	46	95	9.2	0.57	16.2
Claudicatio intermittens	Malattia occlusiva arteriosa periferica	31	93	4.4	0.74	5.6
Polso arteria tibio-dorsale piede	Malattia occlusiva arteriosa periferica	73	92	9.1	0.29	30.4
Alterazioni alvo	Cancro colon-retto	88	72	3.1	0.17	18.4
Perdita di peso	Cancro colon-retto	44	85	2.9	0.66	4.6
VES > 30 mm/ 1 h	Cancro colon-retto	40	96	10	0.42	14
GB > 10^9/L	Cancro colon-retto	75	90	7.5	0.28	26.3
SOF positivo 1 su 3	Cancro colon-retto	50	82	2.7	0.61	4.6

L'LR è quindi una misura complessiva della capacità discriminante di un esame. L'esame è inutile se l'LR è 1 ed è tanto migliore quanto è più alto di 1 nel caso dell'LR+ e quanto è più basso di 1 nel caso dell'LR- [6].
L'Odds Ratio (OR) rappresenta la capacità discriminante complessiva ed è equivalente al rapporto tra LR+ e LR- (9.5/0.05555555555=171) L'OR può essere calcolato anche con una formula diversa:

SENSIBILITÀ/(1-SENSIBILITÀ)/(1-SPECIFICITÀ)/SPECIFICITÀ = (0.95/0.05)/(0.10/0.90) = 171.
L'esame è inutile se l'OR è 1 ed è tanto migliore quanto è più alto di 1.
La curva ROC (Receiver Operating Characteristic) rappresenta la relazione tra SENSIBILITÀ e SPECIFICITÀ per gli esami con punto di cut-off variabile su una scala ordinale (per esempio quando consideriamo 5 livelli di sospetto di frattura) o in una scala di intervallo (per esempio se il livello di sospetto di una patologia è espresso in percentuale). L'esame è inutile se l'area sotto la curva = 0.5, l'esame è perfetto se è 1.
La tabella 2 contiene degli esempi di come possono essere confrontati esami diversi e consente delle riflessioni molto interessanti. In alcuni casi l'esame meno invasivo, come l'ecografia, ha una efficienza uguale o addirittura superiore di un esame più invasivo e pericoloso come l'angiografia. I dati dell'anamnesi (come modificazioni dell'alvo intestinale) possono risultare almeno altrettanto utili degli esami di laboratorio. Quello che è importante non è tanto la capacità di discriminare in sé ma piuttosto di individuare quanta informazione un esame può aggiungere a quella che un esame meno costoso ed invasivo fornisce già al processo diagnostico. Knottnerus e van Weel[7] fanno l'esempio dello scarso contributo di informazione che gli esami di funzionalità epatica danno alla raccolta dell'anamnesi ed alla visita medica nei casi in cui i pazienti presentano sintomatologia sfumata.

2) Contribuire al processo decisionale relativamente alla prosecuzione della gestione diagnostica e terapeutica. Per esempio indicando la sede e la dimensione di una lesione e il migliore approccio terapeutico.

3) Valutare la prognosi. La natura e la gravità dei risultati consentono di pianificare il follow-up e per informare correttamente ed, eventualmente, rassicurare il paziente.

4) Monitorare il decorso clinico di una malattia o di una condizione fisiologica come la gravidanza.
Valutare un esame prima della sua introduzione favorisce la correttezza del processo anche dal punto di vista formale rispetto al farlo dopo che l'esame è entrato in routine. È comune del resto la falsa idea che i costi sanitari sono aumentati solo dalla introduzione di tecnologie costose e rivoluzionarie; in realtà la maggior parte dei costi diretti sono legati ad esami poco costosi ma richiesti in grande numero (i cosiddetti esami di routine). Paradossalmente, anche se questi esami sono quelli sottoposti a valutazione meno accurata, il loro risultato può avere un ruolo importante nel processo decisionale del medico e portare alla esecuzione di esami più costosi e più "rischiosi" per il paziente.

Aspetti metodologici

Relazioni con altri esami
Di solito gli esami diagnostici hanno più di una indicazione e sono eseguiti non da soli ma insieme ad altri esami.
Gold Standard, metodo di riferimento
La necessità di stimare il potere discriminante di un esame confrontandolo con un metodo di riferimento indipendente è ostacolata nella maggior parte delle circostanze dalla difficoltà di avere a disposizione un metodo di riferimento, un gold standard. Anche la Tomografia Assiale (TAC), la Risonanza Magnetica Nucleare (RMN) o i gli esami istologici sono suscettibili di risultati falsi positivi e falsi negativi anche senza considerare che la biopsia è un tecnica invasiva e che non può essere impiegata sistematicamente nella valutazione degli esami. Studiare il valore diagnostico dell'esame clinico di un paziente che presenta disturbi addominali non acuti con un approccio invasivo e sistematico, anche ammettendo che questo fosse possibile ed accettabile dal punto di vista etico, porterebbe ad un numero tale di reperti irrilevanti dal punto di vista clinico da non essere sicuramente raccomandabile. In un'epoca di rapida evoluzione, in alcuni casi anche rivoluzione, tecnologica va considerato anche l'effetto dello standard di riferimento prevalente. Per esempio, dato che l'angiografia classica è considerata il metodo di riferimento per validare una nuova tecnica di imaging vascolare, si può verificare il paradosso che la tecnica più avanzata risulterà sempre meno valida di quella meno avanzata che viene usata come riferimento. Solo quando la nuova tecnica viene riconosciuta come metodo di riferimento, le differenze saranno interpretate a suo favore e non a suo discapito. Questo esempio ci porta anche a considerare che quando si confronta una tecnica ecografia avanzata con l'angiografia nella diagnostica vascolare si deve tenere presente che le due tecnologie misurano cose diverse: una misura il flusso sanguigno che spiega i sintomi dal punto di vista clinico, la seconda riflette la situazione anatomica che è fondamentale per il chirurgo.
Bias di spettro e di selezione
Il *bias di spettro* si verifica quando la capacità discriminante di un esame è valutata in una popolazio-

ne con uno spettro clinico di malattia diverso (per esempio più grave) rispetto a quella in cui l'esame sarà eseguito nella pratica. Questo avviene quando esami che sono stati validati in ambito ospedaliero vengono poi eseguiti in ambito ambulatoriale, ovvero quando si stima la sensibilità di un esame in pazienti affetti da malattia grave e la specificità in soggetti chiaramente sani. In questo caso sensibilità e specificità sono grossolanamente sovrastimate mentre tutti gli esami sono tanto più necessari quanto più è difficile distinguere clinicamente sani e malati.
Il *bias di selezione* si può verificare quando il risultato dell'esame condiziona la probabilità di essere compreso nella popolazione nella quale l'esame è "calibrato". Per esempio, è più probabile che i soggetti con ECG da sforzo anomalo siano selezionati per eseguire l'angiografia e quindi questo esame, a causa di questa preselezione, presenta una sensibilità più alta ed una specificità più bassa.
Variabilità e bias legati all'osservatore
La variabilità tra osservatori va considerata tipicamente nell'interpretazione di esami come radiografie, TAC e preparati istologici ma anche in risultati ottenuti impiegando lo stesso strumento in tempi diversi o impiegando strumenti diversi. L'osservatore deve dare il proprio giudizio in cieco e non deve avere pregiudizi nei confronti del metodo (questo può costituire uno svantaggio quando sono considerate nuove tecnologie).
Impatto clinico
Si può verificare che l'informazione fornita dall'esame non sia sufficiente a modificare la gestione del paziente; questo è stato dimostrato avvenire anche nel caso in cui la diagnosi sia stata modificata dall'esito dell'esame. Deve quindi essere considerata non solo l'accuratezza dell'esame ma anche il suo valore clinico pratico. Un esame con bassa specificità è inappropriato nello screening (con probabilità di malattia bassa) a causa dell'alto rischio di risultati falsi positivi. Nella condizione opposta, quando la probabilità di malattia è molto alta un esame con bassa sensibilità ha il rischio di produrre risultati falsi negativi.
Effetto del rapido progresso tecnologico
La velocità delle innovazioni nella tecnologia diagnostica è tale che per una sua valutazione esaustiva può richiedere un tempo più lungo di quello richiesto dallo sviluppo di una tecnologia ancora più avanzata. Uno dei casi più evidenti è avvenuto nell'ambito della diagnostica per immagini in cui prima che l'efficacia della TAC in termini di costi fosse stata dimostrata in modo definitivo sono state introdotte RMN e Positron Emission Tomography (PET).

Gli esami diagnostici possono essere studiati con approcci metodologici diversi che vanno dalla ricerca clinica originale alla sintesi sistematica di quanto è stato segnalato dalla letteratura e di quanto deriva dalla pratica clinica. Questa sintesi può avere la struttura di una rassegna sistematica, di una metanalisi, di una analisi della decisione clinica, di studi di efficienza economica e metodi di consenso.
È esperienza quotidiana che non mancano tanto gli articoli di ricerca ma una buona sintesi di questi che fornisca una panoramica complessiva del valore di una procedura diagnostica.
L'interfaccia tra medicina clinica e metodi scientifici per ottenere esami con la massima validità ed utilità è sicuramente molto importante. Il minimo numero e la tipologia di domande a cui un articolo deve rispondere sono stati di recente sintetizzati da Sackett e Haynes[8,9] che hanno impiegato come esempio quello, di grande attualità, dell'uso del peptide natriuretico di tipo B (BNP) nella diagnosi di disfunzione ventricolare sinistra.
Fase I: i risultati ottenuti nei pazienti sono diversi da quelli ottenuti nei soggetti sani? In letteratura è stato riportato che la concentrazione di BNP in un gruppo di soggetti sani aveva una mediana di 129.4 ng/L (intervallo: 53.6-159.7 ng/L) mentre quella in un gruppo di pazienti con disfunzione ventricolare sinistra era di 439.5 ng/L (intervallo: 248.9-909.0 ng/L) senza alcuna sovrapposizione (*gli autori hanno concluso che il BNP è un utile aiuto nella diagnosi della disfunzione ventricolare sinistra*). Questa fase contribuisce alla conoscenza della terapia e della diagnosi della malattia ma i risultati che sono ottenuti non possono essere trasferiti direttamente alla diagnosi.
Fase II: i soggetti in cui sono stati ottenuti determinati risultati hanno più probabilità di essere ammalati rispetto a quelli in cui sono stati ottenuti risultati diversi? È stato pubblicato un articolo che confronta la concentrazione di BNP in un gruppo di soggetti sani e in tre gruppi di pazienti con gravità diversa di disfunzione del ventricolo sinistro. Sensibilità e specificità sono risultate rispettivamente 98% (intervallo di confidenza (IC) del 95%: 87-100%) e 96% (IC 95%: 77-98%); la predittività di un risultato positivo e di un risultato negativo sono risultate rispettivamente del 95% (IC 95%: 84-99%) e del 96% (IC del 95%: 81-100%) mentre l'LR+ e l'LR- sono risultate rispettivamente 13 (IC del 95%: 3.5-50) e 0.3 (IC del 95%: 0.0003-0.19) e gli autori commentano che *la concentrazione di BNP è un buon indicatore della gravità e della prognosi di insufficienza cardiaca congestizia* anche se in realtà è stata dimostrata solo l'utilità dell'esame in condizioni ottimali poiché confronta due estremi: soggetti sani e pazienti con una grave malattia.
Fase III: il risultato dell'esame distingue i pazienti con una determinata malattia se eseguito in un gruppo di pazienti in cui la malattia è sospettata? Recentemente è stato descritto l'efficacia diagnostica del BNP se misurato in cieco in pazienti in un terzo dei quali era stato dimostrata una disfunzione del ventricolo sinistro mediante ecocardiografia. In questi pazienti sensibilità e specificità sono risultate rispettivamente 88% (IC 95%: 74-94%) e 34% (IC 95%: 25-44%); la predittività di un risultato positivo

e di un risultato negativo sono risultate rispettivamente del 38% (IC 95%: 29-48%) e del 85% (IC del 95%: 70-94%) mentre l'LR+ e l'LR- sono risultate rispettivamente 1.3 (IC del 95%: 1.1-1.6) e 0.4 (IC del 95%: 0.2-0.9). Gli autori commentano che *è improbabile che l'introduzione della determinazione di routine del BNP migliori la diagnosi di disfunzione ventricolare sinistra.* La validità delle stime di accuratezza generate dagli studi di Fase III può essere compromessa se il metodo in esame non è effettivamente confrontato in cieco con il metodo di riferimento in tutti i casi e se il valore di riferimento o di cut-off viene fissato in modo da ottimizzare le prestazioni di sensibilità e specificità.

È un concetto ben noto che i valori predittivi si possono modificare quando noi ci spostiamo dall'ambito dello screening e della medicina di base (caratterizzati da bassa probabilità pre-test) alla medicina ospedaliera (caratterizzata da probabilità pre-test più alta).

La tabella 3 illustra il caso dei segni clinici impiegati nella diagnosi di appendicite confrontati ad una combinazione di referto istologico (nei casi in cui era stato eseguito l'intervento) ed andamento clinico positivo della malattia (nei casi in cui l'intervento non era eseguito). La percentuale di pazienti con appendicite è più alta in ospedale (63%) che in ambulatorio (14%) anche perché i pazienti con dolenzia del quadrante addominale inferiore destro erano "passati" allo step diagnostico successivo diversamente da quelli che non lo presentavano (il sintomo era infatti presente nel 21% di pazienti ambulatoriali e nell'82% dei pazienti ospedalizzati). È ben noto che la predittività positiva dell'esame è aumentata ma portare ad un livello diagnostico più approfondito pazienti con risultati falsi positivi diminuisce la specificità che nell'esempio in Tabella 3 passa dall'89% al 16%. Ne deriva che un segno clinico importante in ambulatorio (LR+: 8; LR-: 0.2) diventa inutile in ospedale (LR+ ed LR-: 1).

Non è tuttavia costante il fenomeno di una riduzione della specificità di un esame spostandoci da un ambito diagnostico di primo livello (ambulatoriale) ad uno di secondo livello (ospedaliero) come è illustrato dall'esempio della rigidità addominale illustrato nella Tabella 4. In questo caso un segno clinico inutile in ambulatorio (LR+ ed LR- intorno a 1) è molto utile in ospedale (LR+ : 5); la specificità in questo caso non è diminuita ma è passata dal 74% al 95%; la frequenza del segno è aumentata probabilmente per il fatto che, per evitare di non diagnosticare una appendicite, i medici di famiglia l'hanno "sovradiagnosticato".

I clinici possono applicare l'approccio bayesiano all'interpretazione degli esami migliorando la stima della probabilità pre-test sulla base dell'esperienza personale, di statistiche di prevalenza nella popolazione, di banche dati dedicate alla pratica medica, di articoli che hanno valutato l'esame in oggetto e di studi primari sulla probabilità pre-test in contesti diversi. È da rilevare comunque che è stato anche recentemente dimostrato quanto poco siano applicate le tecniche diagnostiche bayesiane nella pratica clinica sia da parte dello specialista che del medico di medicina generale [10].

Fase IV: l'outcome dei pazienti in cui è eseguito questo esame è migliore rispetto all'outcome di quelli in cui non è eseguito? Costituisce l'aspetto più cruciale della valutazione di un esame anche se solo di rado può essere evidente ed immediatamente verificabile; più spesso può essere dimostrato solo mediante il follow-up.

Negli ultimi anni sono apparse rassegne sistematiche e metanalisi dedicate a studi che valutavano l'accuratezza degli esami diagnostici. Per esempio 19 delle 26 rassegne pubblicate tra il 1996 ed il 1997 erano rassegne sistematiche o meta-analisi [11] e 23 delle 45 rassegne dedicate alla chimica clinica ed alla ematologia pubblicate tra il 1985 ed il 1998 erano rassegne sistematiche[12].

Tabella III. Accuratezza del sintomo clinico di dolenzia al quadrante inferiore destro nella diagnosi di appendicite.

	Appendicite (ambulatorio)		Appendicite (ospedale)	
Dolenzia quadrante inferiore dx	Si (%)	No (%)	Si (%)	No (%)
Presente	84	11	81	84
Assente	16	89	19	16
Totale	100	100	100	100
Frequenza di appendicite	14%		63%	
Frequenza di un segno positivo	21%		82%	
Sensibilità	84%		81%	
Specificità	89%		16%	
LR+	7.6		1	
LR-	0.2		1	

Tabella IV. Accuratezza del sintomo clinico di rigidità addominale nella diagnosi di appendicite.

	Appendicite (ambulatorio)		Appendicite (ospedale)	
Addome rigido	Si (%)	No (%)	Si (%)	No (%)
Presente	40	26	23	86
Assente	60	74	77	94
Totale	100	100	100	100
Frequenza di appendicite	14%		47%	
Frequenza di un segno positivo	28%		14%	
Sensibilità	40%		81%	
Specificità	74%		16%	
LR+	1.5		5	
LR-	0.8		0.8	

Figura 4. Strategia di ricerca di articoli dedicati alla accuratezza diagnostica in PubMed (MEDLINE)

((((((((((((sensitivity and specificity [All fields] OR sensitivity and specificity/standards [All fields] OR specificity [All fields]) OR screening [All fields]) OR false positive [All fields]) OR false negative [All fields]) OR accuracy [All fields]) OR ((((predictive value [All fields] OR predictive value of tests [All fields]) OR predictive value of tests/standards [All fields]) OR predictive values [All fields] OR predictive values of tests [All fields])) OR ((reference value [All fields] OR reference values [All fields]) OR reference values/standards [All fields])) OR ((((((((((roc [All fields] OR roc analyses [All fields]) OR roc analysis [All fields]) OR roc and [All fields]) OR roc area [All fields]) OR roc auc [All fields]) OR roc characteristics [All fields]) OR roc curve [All fields]) OR roc curve method [All fields] OR roc curves [All fields]) OR roc estimated [All fields]) OR roc evaluation [All fields])) OR likelihood ratio [All fields] AND notpubref [sb]) AND human [MesH Terms])

Tabella V. Ricerche su Medline di studi diagnostici sulle strisce reattive per gli anni 1990-1995. (Da Ref.14 modificato).

Termini MeSH	
#1 Reagent strips/all subheadings	
#2 Esterases/urine	
#3 Nitrites/urine	
#4 Urinary tract infections/all s.	
#5 Cystitis/all s.	
#6 Pyelocystitis/all s.	
#7 Schistosomiasis/all s.	
#8 (#1 OR #2 OR #3) AND (#4 OR #5 OR #6)	
#9 #8 NOT #7	**57**
#10 dipstick* in (ti or ab)	
#11 esterase* in (ti or ab)	
#12 nitrite* in (ti or ab)	
#13 (#10 OR #11 OR #12) AND (#4 OR #5 OR #6)	
#14 #13 NOT #7	
#15 #9 OR #14	**88**

Tabella VI. Ricerche su Medline di studi diagnostici sulle strisce reattive senza limiti temporali (Meline-Web).

#15 Search #14 OR #9	**104**
#14 Search #13 NOT #7 Field: Title/Abstract	**86**
#13 Search (#10 OR #11 OR #12) AND (#4 OR #5 OR #6) Field: Title/Abstract	**94**
#12 Search nitrite$ Field: Title/Abstract 9348	
#11 Search esterase$ Field: Title/Abstract 11543	
#10 Search dipstick$ Field: Title/Abstract 765	
#9 Search #8 NOT #7	**57**
#8 Search (#1 OR #2 OR #3) AND (#4 OR #5 OR #6)	**87**
#7 Search *schistosomiasis/all s. 5492	
#6 Search *pyelocystitis/all s. 1284	
#5 Search *cystitis/all s. 1045	
#4 Search *urinary tract infections/all s.	**4983**
#3 Search *nitrites/urine 397	
#2 Search *esterases/urine	**999**
#1 Search *reagent strips/all s. 499	

Le linee guida

Come fare ricerche della letteratura dedicata all'accuratezza degli esami diagnostici

Una rassegna sistematica deve comprendere tutte le evidenze disponibili e richiede quindi una ricerca esaustiva e sistematica della letteratura. L'autore della rassegna deve quindi indicare una strategia di ricerca basata sulla descrizione chiara ed esplicita dei soggetti in cui viene eseguito l'esame di interesse, dell'esame stesso e della stima della sua accuratezza, della malattia in cui viene applicato l'esame e dell'impostazione dello studio. Questi elementi devono essere specificati nei criteri di inclusione degli studi primari nella rassegna.

La ricerca della letteratura per identificare gli studi primari può comprendere le fasi seguenti:

1) Una ricerca su una banca dati come MEDLINE (website PubMed http://www.ncbi.nlm.nih.gov/PUBMED) o EMBASE. La strategia di ricerca comincia creando una lista di parole chiave che descrivono l'esame e le malattie di interesse; in questo tipo di ricerca è necessario trovare un compromesso tra il numero limitato degli studi specifici dedicati alla accuratezza degli esami ed il numero talvolta elevatissimo degli studi generali. Recentemente è stato proposto l'impiego di una combinazione di una ricerca generica relativa agli articoli dedicati alla diagnostica associata ad una specifica al soggetto o di due strategie di ricerca generica (Figura 4)[13]. Il gruppo del Dipartimento di Medicina Generale dell'Università di Maastricht ha esaminato la sensibilità ed il valore predittivo positivo di una ricerca per studi diagnostici relativi alla velocità di eritrosedimentazione e l'esame urine impiegando una striscia reattiva. L'interessante articolo dimostra che combinando il vocabolario Medical Subject Headings della National Library of Medicine, i cosiddetti termini MeSH, con una ricerca a testo libero si ottengono risultati migliori rispetto all'impiego dei soli termini MeSH. In Tabella 5 è mostrata la ricerca su Medline relativa agli anni 1990-1995 riportata dal lavoro[14] mentre in Tabella 6 è mostrata una ricerca analoga eseguita non su Medline OVID-CD ma eseguita dall'autore usando Medline Web.
2) La bibliografia degli articoli primari individuati con la ricerca elettronica, delle rassegne narrative e delle rassegne sistematiche deve essere controllata per ricercare ulteriori studi primari che potrebbero non essere stati individuati dalla ricerca elettronica. Sarà tra breve disponibile la banca dati MEDION che può essere richiesta all'indirizzo berna.schouten@hag.unimaas.nl[14] .

3) Ulteriori studi primari anche non pubblicati possono essere richiesti ad esperti della materia; questo può risultare rilevante nell'ambito degli studi diagnostici in quanto gli studi di accuratezza sono spesso basati su studi eseguiti routinariamente nell'ambito della consueta pratica professionale del laboratorio.

Il primo passo in una ricerca della letteratura disponibile, identificare gli articoli rilevanti, può risultare impegnativa soprattutto perché gli articoli di tipo diagnostico, quelli non recenti in particolare, non sono bene indexati. Una volta che la ricerca è completata, due revisori indipendenti devono esaminare i titoli e gli abstract degli articoli identificati applicando i criteri di inclusione precedentemente specificati. Se i due revisori non trovano l'accordo su un articolo o se le informazioni non sono sufficienti viene interpellato un terzo revisore o viene esaminato l'intero articolo.

Criteri di inclusione

- *Esame di riferimento.* La descrizione dell'esame di riferimento è un prerequisito essenziale per la valutazione di un esame diagnostico. L'esame di riferimento può consistere in un singolo esame, di una combinazione di esami diversi o del follow-up del paziente.
- *Popolazione.* I partecipanti devono essere definiti in modo esplicito per quanto riguarda età, sesso, sintomi e segni clinici e loro durata. Deve essere indicata anche una stima dell'accuratezza diagnostica e della precisione di questi parametri.
- *Outcome.* Deve essere presentata una Tabella diagnostica 2 X 2 con 4 celle: veri positivi, falsi negativi, falsi positivi e veri negativi o devono essere fornite informazioni per calcolare questi parametri.
- *Lingua.* Deve essere dichiarato se la rassegna è limitata a pubblicazioni in una o più lingue.

Indicare l'ambito specifico in cui è applicata la rassegna consente di condurre una analisi in sottogruppi nel caso si rilevi una eterogeneità. Mentre tutte le evidenze disponibili devono essere considerate indipendentemente dalla lingua, gli articoli in lingua diversa dall'inglese sono spesso indexati nelle banche dati informatiche in modo non completo causando un bias. Questo avviene soprattutto, ma non solo, negli studi basati su piccole popolazioni che raramente sono costituiti da pazienti consecutivi o selezionati in modo casuale.

Qualità metodologica

Esempi di criteri di validità degli esami diagnostici sono stati pubblicati dal Cochrane Methods Group on Screening and Diagnostic Tests (http://www.som.fmc.flinders.edu.au/fusa/cochrane) e da altri autori[15] (Tabella 7). I criteri di validità interna ed esterna devono essere codificati e descritti in modo esplicito nella rassegna. I criteri interni si riferiscono alle caratteristiche che salvaguardano da errore o bias sistematico mentre quelli esterni si riferiscono alla possibilità di generalizzare lo studio ed alla esecuzione della valutazione dello studio secondo standard accettati. La valutazione metodologica degli studi primari è ostacolata frequentemente da mancanza di informazione. In questi casi i revisori possono decidere o di prendere contatto con gli autori o di sostituire l'informazione mancante con commenti del tipo "non so" o "non chiaro".

L'esame urine con striscia reattiva è eseguito di solito prima dell'esame colturale. L'esame urine standard può essere valutato senza conoscere i risultati dell'esame colturale ma se questo non è specificatamente dichiarato chi esegue la rassegna può decidere di valutare questo punto con "non noto" o "non in cieco". Reid et al[16] hanno riportato come solo una minoranza dei 112 studi pubblicati su 4 giornali autorevoli di medicina interna ha soddisfatto gli standard metodologici appropriati. Lijmer et al[11] hanno dimostrato che l'accuratezza diagnostica di un esame è stato sovrastimato in studi caso-controllo, che usavano esami diversi di riferimento per i risultati positivi e negativi dell'esame in valutazione, che accettavano risultati non ottenuti in cieco, che non descrivevano i criteri diagnostici per l'esame considerato ed in cui i partecipanti non erano descritti in modo adeguato.

Gli studi di screening sono frequentemente interessati dal bias di verifica in quanto solo i risultati positivi sono verificati con l'esame di conferma cosicché può essere calcolato solo il valore predittivo positivo a meno che non siano a disposizione dei registri accurati di malattia (come avviene talvolta nel caso dei Registri Tumori).

Estrazione dei dati

Due revisori devono estrarre indipendentemente le informazioni richieste dagli studi primari. Devono essere estratte informazioni dettagliate sui partecipanti inclusi nello studio e circa le procedure di esame. Devono essere sempre indicati i valori di cut-off usati negli esami dicotomici, le ragioni delle esclusioni ed il numero dei partecipanti esclusi. L'eterogeneità riscontrata in una meta-analisi dell'esame standard delle urine mediante strisce reattive può essere attribuita a: procedure di raccolta del materiale esaminato (modalità di raccolta delle urine, tempo intercorso tra raccolta delle urine ed esame colturale), esecutore dell'esame e modalità di esecuzione (manuale o automatica), produttore diverso dei reagenti.

L'accuratezza può essere presentata in diversi modi. Per la meta-analisi di esami dicotomici è necessario costruire una tabella 2 X 2 indicando numeri assoluti nelle quattro celle. Sono necessari il totale dei partecipanti "malati" e "non-malati" per calcolare la probabilità a priori (probabilità pre-test) e per costruire la Tabella 2 X 2 dai dati di sensibilità, specificità, quoziente di probabilità, valori predittivi e curve ROC.

Tabella VII. Esempio di criteri di validità per articoli che riportano dati sulla accuratezza delle strisce reattive nella diagnosi di infezione delle vie urinarie (UTI) o batteriuria.

CRITERI DI VALIDITÀ INTERNA	Score positivo
1 Standard di riferimento valido	Coltura (semi) quantitativa (2 punti) o qualitativa (1 punto)
2 Definizione del punto di cut-off per lo standard di riferimento	Definizione di infezione del tratto urinario/ batteriuria mediante unità formanti colonie per ml (1 punto)
3 Misura in cieco dell'esame in valutazione e dell'esame di riferimento	Nelle due direzioni (2 punti) o solo esame in valutazione o di riferimento
4 Mancanza del bias di verifica	Valutazione mediante standard di riferimento indipendente dai risultati dell'esame in valutazione (1 punto)
5 Esame in valutazione interpretato indipendentemente da tutte le informazioni cliniche	Menzionato in modo esplicito nell'articolo o in campioni di urine provenienti da popolazione mista esaminata in un laboratorio generale (1 punto)
6 Progetto	Raccolta dei dati prospettica (serie consecutiva) (1 punto) o retrospettiva (0 punti)
CRITERI DI VALIDITÀ ESTERNA	
1 Spettro di malattia	Criteri di inclusione ed esclusione menzionati
2 Ambito	Informazione sufficiente per identificare l'ambito (medicina del territorio vs medicina ospedaliera) (1 punto)
3 Esami precedenti/filtro di riferimento	Dettagli circa le informazioni cliniche e altre informazioni diagnostiche come il tipo di pazienti in cui viene eseguito l'esame (pazienti sintomatici o asintomatici) (1 punto)
4 Durata della malattia prima della diagnosi	Durata menzionata (1 punto)
5 Condizioni comorbose	Dettagli forniti (tipo della popolazione) (1 punto)
6 Informazioni demografiche	Dati relativi ad età (1 punto) e/o sesso (1 punto) forniti
7 Modalità di esecuzione dell'esame	Informazioni sulla procedura in esame direttamente o indirettamente disponibili, procedura di raccolta delle urine, urine della prima minzione, distribuzione dei microrganismi, procedure per i campioni di urine contaminati, tempo di trasporto dei campioni di urine, modalità di lettura dell'esame e persone addette (1 punto)
8 Spiegazione del punto di cut-off dell'esame	Traccia, 2 o più 1 (1 punto se applicabile)
9 Percentuale di dati mancanti	Se appropriato: menzione dei dati mancanti (1 punto)
10 Riproducibilità dell'esame	Riproducibilità studiata o riferimento citato (1punto)

Tabella VIII. Tabella di presentazione dei dati degli studi relativi alle strisce reattive per urine

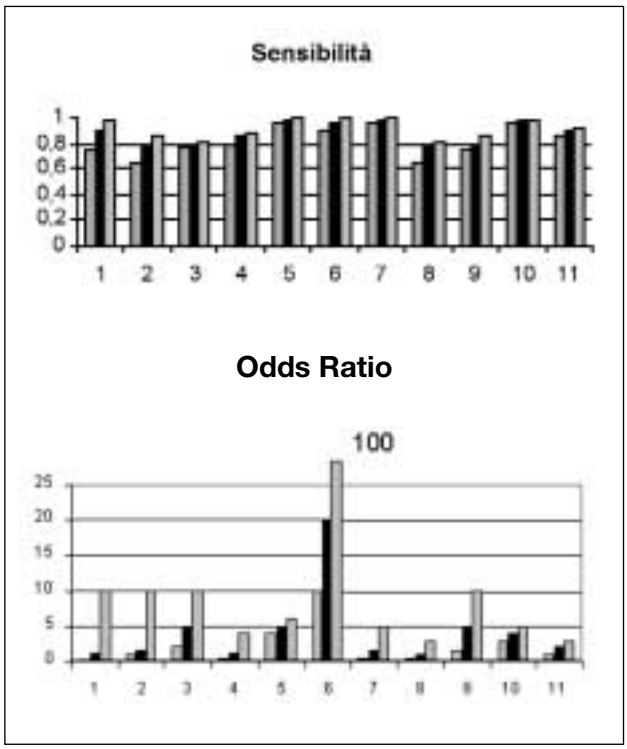
Fattore	DOR (95% IC)	Sensibilità (95% IC)	Specificità (95% IC)	Probabilità Pre-test	Predittività Valore Positivo	Predittività Valore Negativo
Popolazione mista	11 (6-21)	0.50 (0.44-0.58)	0.82 (0.71-0.95)	0.32	0.57	0.78
Chirurgia	34 (25-47)	0.54 (0.39-0.74)	0.96 (0.93-0.99)	0.2	0.76	0.89

Analisi dei dati

Poiché gli studi di accuratezza sono spesso eterogenei e contengono informazioni limitate risulta di solito difficile completare una meta-analisi e la rassegna può essere limitata alla analisi descrittiva qualitativa della ricerca diagnostica disponibile (sintesi della migliore evidenza). Le fasi raccomandate dall'analisi sono le seguenti:

1. *Presentazione dei risultati dei singoli studi*
2. *Ricerca della presenza di eterogeneità*
3. *Esame per la presenza di un effetto cut-off implicito*
4. *Esame del problema dell'eterogeneità*
5. *Decisione circa il modello di cumulo (pooling) appropriato*
6. *Cumulo (pooling) statistico*

Figura 5. Stima di sensibilità e Odds Ratio di 11 studi sulla validità di un esame diagnostco (Da Ref. 17 modificato)



1) *Presentazione dei risultati dei singoli studi.* Gli studi sono presentati con alcune informazioni (anno di pubblicazione, regione geografica, numero di pazienti malati e non, scelta dei pazienti, caratteristiche metodologiche) e riassunto dei risultati. Poiché la gran parte degli esami diagnostici sono asimmetrici (alcuni sono efficaci nell'escludere la malattia, altri a confermarla) è importante riportare coppie di misure complementari (sensibilità e specificità, valore predittivo positivo e negativo, quoziente di probabilità positivo e negativo). Può essere aggiunto anche l'Odds Ratio Diagnostico (DOR) che si può calcolare, come si è visto in precedenza in due modi:

   DOR = SENSIBILITÀ/ (1-SENSIBILITÀ)/ (1-SPECIFICITÀ)/ SPECIFICITÀ

   DOR = LR+/LR- = SENSIBILITÀ/ (1-SPECIFICITÀ)/(1-SENSIBILITÀ)/ SPECIFICITÀ

2) *Ricerca della presenza di eterogeneità.* Anche se i revisori cercano di definire un insieme più o meno omogeneo di studi, spesso si registra una ampia eterogeneità. L'omogeneità di sensibilità e specificità può essere valutata con un test del chi quadro o del test esatto di Fischer. Un metodo semplice ma molto informativo per valutare l'eterogeneità è quello di produrre un grafico in cui gli outcome dei singoli studi sono raccolti insieme all'intervallo di confidenza al 95% in un grafico (Figura 5).

3) *Esame per la presenza di un effetto cut-off implicito.* La stima dell'accuratezza diagnostica cambia se il cut-off per un risultato positivo usato negli studi cambia. L'interpretazione del risultato di un esame dipende dal processo analitico e studi diversi possono usare cut-off diversi; la variazione dell'accuratezza può dipendere dalla variazione del cut-off. Una possibilità per valutare questo aspetto è quella di calcolare il coefficiente di correlazione di Spearman tra la sensibilità e la specificità di tutti gli studi considerati. Se la coppia di parametri presenta una forte correlazione negativa ($p < -0.4$) questi rappresentano la stessa DOR ed il logaritmo della DOR (lnDOR) risulterà omogeneo e può essere ottenuta una curva ROC riassuntiva.

4) *Esame del problema dell'eterogeneità.* Valutare il problema della eterogeneità è spesso uno degli aspetti più interessanti di una metanalisi. Il semplice esame del grafico dei parametri di prestazione con l'intervallo di confidenza del 95% può dimostrare la presenza di outlier che deve essere investigata. In una rassegna dedicata al valore diagnostico dell'ematuria macroscopica per la diagnosi del cancro in ambito di medicina primaria è stato dimostrato che il valore predittivo positivo complessivo di 5 studi è risultato di 0.19 (IC del 95%: 0.17-0.23) e quello di un singolo studio di 0.4. Gli autori hanno ipotizzato che i medici di medicina generale dell'area in cui è stato eseguito lo studio con il PPV circa doppio degli altri erano particolarmente competenti e quindi valutavano personalmente i casi ed inviavano allo specialista solo i casi più gravi portando ad una popolazione molto selezionata e con una PPV molto alta[17]. In alcuni casi gli outlier possono essere esclusi e l'analisi viene continuata considerando solo i restanti studi ma il processo della esclusione va compiuta solo con grande cautela. È possibile condurre una analisi per sottogruppi in modo da individuare sottogruppi omogenei. Gli studi considerati in una rassegna sull'impiego di strisce reattive per l'analisi delle urine hanno presentato sensibilità e specificità debolmente associate ($p = -0.227$) e molto eterogenee[13]. L'analisi dei sottogruppi ha mostrato significative differenze del lnDOR tra sei diverse popolazioni di partecipanti. In tre popolazioni si registrava una forte associazione negativa tra sensibilità e specificità ($p= -0.539$; $p= -0.559$; $p= -1.000$) producendo lnDOR omogenei nei tre sottogruppi.

5) *Decisione circa il modello di cumulo (pooling) appropriato.* Il cumulo dei risultati di studi diversi può essere fatto con due modelli. Il *modello di effetto fisso* assume che tutti gli studi costituiscano un campione casuale di un unico grande studio e che le differenze di risultati siano casuali. In questo caso il cumulo è semplice: viene calcolata una media pesata per l'inverso della varianza dell'accuratezza dell'esame o per il numero dei partecipanti dei risultati dei diversi studi. Il *mo-*

dello dell'effetto casuale prevede invece che, oltre alla presenza dell'errore casuale, le differenze tra gli studi possa risultare anche da differenze tra le popolazioni e le procedure. In questo caso il calcolo è più complesso e comprende anche la variazione intra- ed inter-studio. Se i parametri sono omogenei e non presentano effetto (implicito) di cut-off, i loro risultati possono essere cumulati e si può usare un modello di effetto fisso mentre se vi sono evidenze di effetto di cut-off possono essere costruite curve SROC o si possono cumulare le curve ROC. Nel caso invece si registri eterogeneità, il revisore o non cumula i risultati degli studi e si limita ad una osservazione qualitativa o, se possibile, esegue una analisi in sottogruppi ed esegue il cumulo all'interno dei sottogruppi o tenendo conto dell'effetto casuale. Considerata la scarsa qualità della maggior parte degli studi diagnostici, frequentemente si consiglia di cumulare tutti gli studi usando modelli di effetto casuale anche se non vi sono segni di eterogeneità.

6) *Cumulo (pooling) statistico*

Cumulo delle proporzioni

Sensibilità e specificità omogenee. Se si può usare il cumulo ad effetto fisso, le proporzioni cumulative sono la media di tutti i risultati degli studi individuali, pesate per le dimensioni del campione. Questo si ottiene facilmente sommando tutti i numeratori e dividendo il totale per la somma di tutti i denominatori (Figura 6).

Effetto cut-off; curva SROC
Irwig et al ricordano che sensibilità e specificità media di tre studi di dimensioni analoghe e con sensibilità e specificità rispettivamente di 100% e 0%, 99% e 99% e 0% e 99% risultano del 66%[18]. Tuttavia se la percentuale dei risultati veri positivi (sensibilità) è graficata contro la percentuale dei falsi positivi (1-specificità), vale a dire gli assi che sono usati anche nella curva ROC, l'accuratezza diagnostica risulta elevata con una AUC vicina ad 1.

In generale, stimando sensibilità e specificità media separatamente, si riduce l'accuratezza dell'esame. Uno scattergramma dei risultati veri positivi versus i risultati falsi positivi consente un primo approccio ragionevole nel valutare una meta-analisi. Questo metodo viene denominato Summary Receiving Operating Characteristic Curve (SROC).
La curva SROC presenta 1-specificità sull'asse x e la sensibilità sull'asse y, nel caso in cui ogni studio fornisce un valore per la sensibilità ed uno per la specificità. Se si può ottenere il fittaggio di una curva SROC si usa un modello di regressione con il logaritmo naturale del DOR (lnDOR) degli studi come variabile dipendente e due parametri come variabili indipendenti, uno per l'intercetta (lnDOR medio) ed uno per la slope della curva (variazione degli lnDOR tra gli studi a causa delle differenze della soglia).

Figura 6. Formule statistiche impiegate per la valutazione dell'eterogeneità degli studi

Sensibilità e/o specificità omogenee

$\text{Sensibilità}_{\text{cumulata}} = \Sigma^{k}_{i=1}\ a_i / \Sigma\ (a_i + c_i)$

a = veri positivi, c = falsi negativi, i = numero dello studio; k = numero totale degli studi

$SE = \sqrt{\frac{p\ (1-p)}{n}}$

dove

$p = \text{sensibilità}_{\text{cumulata}}$ e

$n = \Sigma^{k}_{i=1}\ (a_i + c_i)$

Effetto del punto di cut-off : curva SROC

$\ln(\text{DOR})\text{cumulata} = \alpha + \beta S$

α = intercetta, β = coefficiente di regressione
S = Stima del punto di cut-off = ln (sensibilità/(1-sensibilità) +ln (specificità/(1- specificità)
Con errore standard:

$SE_{\ln(\text{DOR})} = \sqrt{1/a + 1/b + 1/c + 1/d}$

Cumulo dei quozienti di probabilità

$\log(LR)_{\text{cumulati}} = \log$ (numero dei non malati/ numero dei malati) $+ \alpha + \beta x$

dove LR = quoziente di probabilità
log (numero dei non malati/ numero dei malati) = fattore di conversione
α = intercetta nella regressione logistica; β coefficiente di regressione; x = misura dell'esame

Cumulo delle curve ROC

I risultati degli studi con un risultato dicotomico rispetto ad un gold standard ed un risultato che è refertato su una scala continua sono di solito rappresentati come curva ROC accompagnata o meno dall'area sotto la curva (AUC) correlata e l'intervallo di confidenza al 95%. Il revisore può cumulare sensibilità e specificità per i valori di cut-off rilevanti, cumulare le AUC e cumulare le curve ROC stesse.
Una curva ROC cumulativa e il suo intervallo di confidenza possono essere costruite sulla base della sensibilità e specificità cumulate per un determinato punto di cut-off (per fare questo occorre un numero sufficiente di dati che è di rado possibile).
Le curva AUC come tutte le misure unidimensionali non danno informazioni circa la natura asimmetrica di un esame diagnostico. Non può, per esempio, distinguere un esame con alta sensibilità e bassa specificità ed uno con bassa sensibilità ed elevata specificità.
Le curve ROC sono robuste per quanto riguarda le

variazioni tra i diversi studi circa il valore ed il significato dei punti di cut-off e forniscono anche informazioni circa la natura asimmetrica dell'informazione dell'esame. Il cumulo delle curve ROC può essere eseguito solo con le curve pubblicate ed il numero dei partecipanti positivi e negativi rispetto al metodo di riferimento. Il cumulo può essere convertito in una serie di dati di sensibilità e specificità con software dedicati[13].

Presentazione dei dati

L'interpretazione di un DOR, combinazione di sensibilità e specificità, è difficile. Tuttavia è utile presentare sensibilità e specificità cumulate insieme agli Odds Ratio rilevanti per le diverse caratteristiche e sottogruppi degli studi.
L'informazione diventa più accessibile ai clinici se viene fornito anche il valore predittivo usando la probabilità a priori (pre-test) di ogni sottogruppo ovvero i quozienti di probabilità in modo che sia possibile calcolare la probabilità post-test sulla base della probabilità pre-test. I DOR diagnostici cumulati e gli intervalli di confidenza dei diversi sottogruppi possono essere presentati graficamente su una scala logaritmica in modo da fornire degli intervalli di confidenza simmetrici di ampiezza ridotta (Tabella 8).

Esempi di rassegne sistematiche degli esami di laboratorio
La revisione sistematica degli studi dedicati agli esami diagnostici presenta alcuni aspetti complessi: è difficile rintracciare tutti gli articoli pubblicati in quanto indexati nelle banche dati elettroniche in modo inadeguato, gli studi sono spesso di cattiva qualità (la definizione dei malati non è definita chiaramente, la descrizione dei partecipanti agli studi non è adeguata, il cieco non è rispettato, i dati non sono interpretati in modo indipendente), la stima dell'accuratezza è spesso molto eterogea, i risultati devono essere tradotti in informazioni clinicamente rilevanti considerando la realtà clinica in diversi gradi di malattia (prevalenza della malattia, spettro della malattia, disponibilità di informazioni cliniche e di altro tipo).
Il lavoro in questo ambito può essere fatto risalire all'articolo di Irwig at al[18] che hanno presentato le linee guida per le meta-analisi dedicate agli esami diagnostici. Questo autorevole gruppo sintetizzava i seguenti step per condurre una meta-analisi di un esame diagnostico:

1. Stabilire l'obiettivo e lo scopo di una meta-analisi
   1.1. Esiste una descrizione chiara su:
      1.1.1. Esame di interesse
      1.1.2. Malattia di interesse ed esame di riferimento
      1.1.3. Problema clinico e contesto
2. L'obiettivo è quello di valutare un solo esame o di paragonare esami diversi?
3. Individuare la letteratura rilevante
   3.1. È descritta la letteratura rilevante con i termini e le modalità di ricerca?
   3.2. Sono indicati i criteri di inclusione ed esclusione?
4. Estrarre e presentare i dati
   4.1. Gli studi sono valutati da due o più revisori?
   4.2. Gli autori spiegano come sono risolti i disaccordi tra i revisori?
   4.3. Vi è un elenco completo dell'accuratezza diagnostica e delle caratteristiche dello studio di ogni studio primario?
5. Stimare l'accuratezza diagnostica
   5.1. Il metodo per cumulare sensibilità e specificità tiene conto della loro interdipendenza?
   5.2. Quando sono disponibili categorie multiple di un esame sono usate nel riassunto?
6. Valutare l'effetto della variabilità nella validità degli studi sulla stima dell'accuratezza diagnostica
   6.1. È esaminata la relazione tra stima dell'accuratezza diagnostica e validità degli studi primari per ognuna delle seguenti caratteristiche:
      6.1.1. Esame di riferimento appropriato
      6.1.2. Valutazione dell'esame o dell'esame di riferimento indipendente
      6.1.3. Bias di verifica evitato
   6.2. Negli studi comparativi, gli esami sono stati eseguiti in tutti i pazienti o sono stati selezionati i pazienti in modo casuale?
   6.3. Sono usati dei metodi analitici che stimano se i difetti nello studio influenzano l'accuratezza diagnostica piuttosto che semplicemente la soglia dell'esame?
7. Stimare l'effetto della variabilità delle caratteristiche del paziente e dell'esame nella stima dell'accuratezza diagnostica
   7.1. È esaminata la relazione tra stima dell'accuratezza diagnostica e caratteristiche dei pazienti e dell'esame?
   7.2. I metodi analitici individuano se le caratteristiche influenzano accuratezza e soglia dell'esame?

Il lavoro del gruppo australiano-statunitense è servito da base per quello di un gruppo più ampio di autori condotto sotto l'egida della Cochrane Collaboration[19]. Secondo Cochrane Collaboration un esame (test) è "ogni determinazione che ha lo scopo di individuare soggetti che hanno la potenzialità di trarre giovamento da un intervento" e può quindi essere un sintomo, un segno, un esame di laboratorio o un valore di rischio per la presentazione di una malattia nel futuro. Coerentemente a questa definizione deve essere data la priorità a rassegne sistematiche di esami che hanno l'impatto maggiore nella diagnosi di malattia. Il riferimento per il gruppo rimane il lavoro di Irwig et al[18] ma vengono fatte delle precisazioni:

1. Il primo step nell'analisi è la semplice presentazione descrittiva comprendente

1.1. Tabelle che riassumono la qualità e l'applicabilità degli studi primari
1.2. Tabelle che riassumono la sensibilità, la specificità e gli Odds Ratio nelle categorie di qualità e l'applicabilità degli studi primari
1.3. Presentazione grafica di sensibilità e specificità degli studi primari, preferibilmente con l'indicazione delle dimensione e degli intervalli di confidenza

2. Anche se l'accuratezza dell'esame deve essere stimata lungo l'intervallo dei risultati piuttosto che in modo dicotomico la maggior parte degli studi primari presentano i dati in quest'ultimo modo ed i metodi statistici per i dati continui sono meno sviluppati.
3 Il modo migliore per valutare gli studi dicotomici è la SROC.
4. Lo scopo di una rassegna sistematica dell'accuratezza di un esame può essere definito restringendola agli studi di elevata qualità applicabili al problema di interesse immediato ovvero valutare l'effetto della variabilità della qualità dello studio di altre caratteristiche (ambito, tipo della popolazione, spettro di malattia) sulla stima della accuratezza.
5. Sono indicati i criteri necessari alla validità dello studio:
   5.1. L'esame è stato confrontato ad uno standard di riferimento valido?
   5.2. Esame ed esame di riferimento sono stati misurati in cieco?
   5.3. La scelta dei pazienti esaminati dall'esame di riferimento era indipendente dai risultati dell'esame (BIAS DI VERIFICA)?
   5.4. L'esame è stato misurato indipendentemente da tutte le altre informazioni cliniche?
   5.5. L'esame di riferimento è stato eseguito prima che fosse iniziato qualunque intervento dopo che il risultati dell'esame erano conosciuti (PARADOSSO DEL TRATTAMENTO)?

Sono indicati i criteri necessari alla applicabilità dei risultati:

Clinica
5.a.spettro della malattia
5.b.spettro della non-malattia
5.c.ambito in cui è condotto lo studio
5.d.durata della malattia prima dell'esecuzione dell'esame
5.e. filtro degli esami precedenti
5.f. condizioni co-morbose
5.g. informazioni demografiche

Esame
5.h.natura (biochimico,..)
5.i.indicare la soglia
5.j.indicare la percentuale di risultati esclusi per l'impossibilità di eseguire l'esame o di risultati indeterminati
5.k.riproducibilità.

Sono indicate misure indirette di qualità ed applicabilità:
5.A. Anno di pubblicazione dello studio
5.B. Prevalenza della malattia
5.C. Dimensioni del campione
5.D. Studio prospettico o retrospettivo
5.E. Articolo o abstract

Oosterheuis et al[12] hanno avuto il merito di esemplificare quanto era stato sostenuto negli articoli precedenti. La Tabella 9 è utile per esemplificare alcune applicazioni che sono state fatte fino ad oggi di rassegne sistematiche degli esami diagnostici. Gli autori affermano nella discussione che avevano applicato dei criteri di esclusione poco rigidi perché, in qualche caso, anche se le rassegne avevano l'obiettivo di essere sistematiche, avevano chiaramente fallito l'obiettivo probabilmente non tanto perchè le linee guida non erano state rispettate ma perché gli autori non avevano riferito in modo corretto i risultati. Come si vede nella Tabella 10 nessuno delle rassegne ha soddisfatto tutti i sei punti delle linee guida; sette ne rispettano quattro o cinque e dodici meno di tre. Questo riporta quindi al problema della scarsa qualità di molti studi primari in ambito diagnostico anche se il lavoro di molti gruppi ridurrà questo fenomeno. Lijmer e al[11] hanno recentemente dimostrato e quantificato in modo molto convincente che esami diagnostici con dei limiti metodologici, ed in particolare quelli che comprendono pazienti non rappresentativi e quelli che applicano metodi di riferimento diversi, possono sovrastimare l'accuratezza degli esami[10]. Anche Deeks sottolinea questo aspetto e sostiene che comunque la qualità degli studi compresi nella rassegna deve essere valutata e registrata, in modo da poterne tener conto nelle conclusioni[20]. Anche se le SROC risultano poco utili nella pratica clinica, possono infatti identificare se un esame ha una potenziale utilità clinica ma non possono essere usate per calcolare le probabilità di malattia in caso di un particolare risultato di un esame. D'altra parte il vantaggio dell'impiego delle SROC rispetto a cumulare sensibilità, specificità e quoziente di probabilità si registra solo in presenza dell'effetto soglia. Il tentativo di individuare dei metodi più semplici per cumulare questi parametri è assolutamente auspicabile.

Se gli interventi terapeutici sono raccomandati nell'impiego clinico solo se è dimostrato che producono più vantaggi che danni, così gli esami di laboratorio dovranno essere introdotti solo se sono utili e producono più vantaggi che danno.

Tabella IX. Riassunto delle rassegne valutate da Oosterhuis et al[12].

Autore	Problema clinico	Conclusione
Aziz et al 1993	Valore diagnostico del PSA nella diagnosi di cancro della prostata	Un cut-off di 3 _g/L individua il 74% dei casi
Becker at al 1996	Uso del D-dimero nella diagnosi di tromboembolismo venoso acuto	Mancanza di standardizzazione, di standard di riferimento e di valutazione prospettica impediscono il suo uso clinico
Campens et al 1997	Confrontare i metodi per misurare la filtrazione glomerulare	La formula di Cockroft-Gault consente una stima ragionevole della funzione renale soprattutto in ambito ambulatoriale
Chien et al 1997	Accuratezza della fibronectina cervico-vaginale fetale nel predire il parto pre-termine	Accuratezza limitata
Craig 1998	La Lp(a) è un fattore di rischio per la malattia cardiaca ischemica?	Fattore indipendente di malattia cardiaca ischemica
Da Silva 1995	Proteina C reattiva e leucociti sono indici accurati di setticemia nei neonati in ambito di terapia intensiva?	La PCR è probabilmente il migliore esame anche se i risultati dipendono molto dal metodo usato e dalla popolazione studiata
Feldt-Rasmussen 1994	La presenza degli anticorpi contro il recettore per il TSH consente di anticipare la recidiva dei pazienti trattati per malattia di Graves? L'assenza di anticorpi alla fine del trattamento protegge dalle recidive?	L'assenza di anticorpi protegge contro la recidiva della malattia di Graves dopo la terapia anche se il 25% dei pazienti è "misclassificato" e il metodo non è quindi utile nel singolo paziente
Gerdes 1998	Accuratezza diagnostica della PCR nel liquido cerebro-spinale e nel siero per fare diagnosi di meningite batterica	Solo un esame negativo è molto informativo
Hallan et al 1997	Accuratezza diagnostica della PCR e del conteggio dei leucociti nella diagnosi dell'appendicite acuta	L'accuratezza della PCR è "media" anche se tende ad essere inferiore al conteggio dei leucociti
Hoeksem et al 1993	Valore degli esami di laboratorio nello screening e nel riconoscere l'abuso alcolico in ambito ambulatoriale	ALT/AST, GGT e MCV possono far sospettare ma non confermare l'abuso alcolico
Vd Hoogen et al 1995	Valutazione dell'accuratezza della anamnesi, esame fisico e VES nella diagnosi di radicolopatie, cancro vertebrale e spondilite anchilosante come causa di lombalgia	L'accuratezza è modesta
Hurlbut et al 1991	Accuratezza diagnostica nel test dei nitriti e dell'esterasi leucocitaria nel predire una infezione batterica del tratto urinario	La combinazione migliore è data dalla combinazione positiva di nitriti ed esterasi leucocitaria. Un risultato negativo non può escludere l'infezione quando la probabilità a priori è elevata
Jensen et al 1996	Utilità della striscia reattiva MICRAL per individuare la microalbuminuria	Il fattore decisivo nella utilità del MICRAL è stato dalla prevalenza della microalbuminuria
Najmey et al 1997	Gli anticorpi contro la beta-2 glicoproteina I è associata con la sindrome da antifosfolipidi ?	L'associazione è confermata e la determinazione degli anticorpi può aiutare la diagnosi
Offringa et al 1992	Eritrociti dismorfici o cellule con basso MCV indicano sanguinamento glomerulare?	Valore diagnostico limitato
Oosterhuis et al 2000	Valore diagnostico di un MCV elevato nella diagnosi di carenza di vitamina B12	La sensibilità è bassa lasciando il 20% dei casi non riconosciuti
Peters et al 1996	L'emoglobina glicata può essere usata al posto del carico orale di glucosio per diagnosticare il diabete mellito	La glicemia a digiuno seguita dalla determinazione selettiva dell'emoglobina glicata può esser il metodo migliore per la diagnosi di un diabete mellito che richiede trattamento.
Rao et al 1995	Utilità dei C-ANCA nella diagnosi della granulomatosi di Wegener	Vi sono ancora seri problemi di accuratezza
Van Schaik et al 1995	Valore diagnostico degli anticorpi contro i gangliosidi GM1 nei disordini del motoneurone e delle neuropatie	Gli ELISA per gli anticorpi GM1 sono utili se il metodo usato è buono
Spencer Green et al 1997	Sensibilità e specifictà degli anticorpi anti-centromero e Scl-70 nella sclerosi sistemica	Entrambi sono molto specifici. La determinazione è secondaria al quadro clinico.
Tenner et al 1994	Utilità dei dati di laboratorio nel distinguere pancreatite da colelitiasi	Una AST elevata e/o una ALT elevata in presenza di pancreatite sono molto utili nella diagnosi
Watine et al 1998	L'emocromo ha un valore prognostico indipendente nel cancro primario del polmone	Attualmente non si ritiene abbia valore con l'eccezione del conteggio dei linfociti e dei neutrofili nel cancro del polmone non a cellule piccole
Wu et al 1995	Utilità delle troponina I nella diagnosi e nella prognosi della malattia cardiaca ischemica	La troponina T è sensibile come il CK-MB nella diagnosi retrospettiva di infarto del miocardio anche se meno specifica. Troponina elevata è associata ad una prognosi peggiore.

Tabella X. Linee guida rispettate (+) e non rispettate (-) dalle rassegne sistematiche studiate da Oosterhuis[12]

Autore rassegna	Linee guida						Totalmente soddisfatte
	1	2	3	4	5	6	
Aziz et al 1993	-	+	-	+	-	-	2
Becker at al 1996	+	-	+	+	-	-	3
Campens et al 1997	+	-	-	+	-	-	2
Chien et al 1997	+	+	+	-	+	+	5
Craig 1998	-	+	-	-	-	+	2
Da Silva 1995	+	+	+	-	+	+	5
Feldt Rasm-ussen 1994	-	-	-	-	+	+	2
Gerdes 1998	-	+	+	+	+	+	5
Hallan et al 1997	+	+	+	+	-	+	5
Hoeksem et al 1993	+	+	-	-	-	-	2
Vd Hoogen et al 1995	+	+	+	+	-	-	4
Hurlbut et al 1991	+	+	-	+	+	-	4
Jensen et al 1996	-	+	-	-	-	-	1
Najmey et al 1997	-	+	-	-	-	-	1
Offringa et al 1992	+	-	-	-	-	-	1
Oosterhuis et al 2000	+	+	+	-	+	+	5
Peters et al 1996	+	-	-	-	-	-	2
Rao et al 1995	+	+	-	-	+	-	3
Van Schaik et al 1995	+	+	-	-	-	+	3
Spencer Gre-en et al 1997	-	+	-	-	+	+	3
Tenner et al 1994	-	-	-	+	-	-	1
Watine et al 1998	-	+	-	-	-	-	1
Wu et al 1995	-	+	-	+	-	-	2

Bibliografia

1. Sandberg S, Oosterhuis W, Freedman D, Kawai T. Systematic reviewing in laboratory medicine. JIFCC 1997; 9: 154-5.
2. Mulrow CD. Rationale for systematic reviews. In: Chalmers I, Altman DG (eds) Systematic reviews. London: BMJ Publishing Group, 1995: 1-8.
3. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995; 274: 645-51.
4. Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. BMJ 2002; 324: 477-80
5. Knottnerus JA (ed) The evidence base of clinical diagnosis. London: BMJ Books, 2002.
6. Dorizzi RM, Giavarina D, Venturini M. Presentazione del repertorio sull'efficienza diagnostica degli esami di laboratorio. Riv Med Lab-JLM 2001; 2 (S1): 106-12.
7. Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA (ed) The evidence base of clinical diagnosis. London: BMJ Books, 2002; pp. 1-17.
8. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002; 324: 539-41.
9. Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA (ed) The evidence base of clinical diagnosis. London: BMJ Books, 2002 ; pp. 19-38.
10. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgements: practising physicians' use of quantitative measures of test accuracy. Am J Med 1998; 104: 374-8.
11. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design related bias in studies of diagnostic tests. JAMA 1999; 282: 1061- 6.
12. Oosterhuis WP, Niessen RWLM, Bossuyt PMM. The science of systematic reviewing studies of diagnostic tests. Clin Chem Lab Med 2000; 38: 577-88.
13. Devillé WL, Buntinx F. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In: Knottnerus JA (ed) The evidence base of clinical diagnosis. London: BMJ Books, 2002 ; pp. 145-65.
14. Van der Weijden T, Ijzermans CJ, Dinant G-J, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in Medline. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. Family Practice 1997; 14: 204-8.
15. Greenhalgh T. How to read a paper: papers that report diagnostic o screening tests. BMJ 1997; 315: 540-3.
16. Reid CM, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995; 274: 645-51.
17. Buntinx F, Wauters H. The diagnostic value of macroscopic haematuria in diagnosing urological cancers: a meta-analysis. Family Practice 1997; 14: 63-8.
18. Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120:667-76.
19. The Cochrane methods group on systematic review of screening and diagnostic tests: Screening and diagnsotic tests: recommended methods. http://www.cochrane.org/cochrane/sadtdoc1.htm (ultimo accesso 21 aprile 2002)
20. Deeks J. Systematic reviews of evaluations of disanostic and screening tests. BMJ 2001; 323: 157- 62.