

## L'iniziativa STARD per una corretta valutazione dell'accuratezza dei test diagnostici

**Giuseppe Giocoli\***

\*Anche a nome degli altri componenti del Gruppo di Lavoro EBM AMCLI: Paola Pauri, Lidia Ricci, Pasquale Urbano

**Cambiare dispiace: spesso non vuol farlo neanche chi soffre a causa della propria condizione. Nessuno ha descritto questa situazione meglio di Tomasi di Lampedusa.**

(Da un editoriale del British Medical Journal)

Il 2002 è stato un anno speciale per la diagnostica basata sull'evidenza. Dal campo terapeutico, l'interesse degli studiosi si è spostata finalmente su questo importante settore (7, 10). Delle varie pubblicazioni sul tema ve n'è una che merita particolare attenzione: *The STARD Initiative*, a cura del gruppo *Standards for Reporting Diagnostic Accuracy*, comprendente studiosi di tre continenti; in essa si auspica la totale trasparenza degli studi sull'accuratezza dei test diagnostici (11; versione italiana a <http://www.gimbe.org/link/diagnosis.htm>). La proposta (si tratta di un documento provvisorio) è di migliorare la qualità strutturale ed espositiva del testo con l'ausilio di una *checklist* e di un diagramma di flusso. La descrizione completa e scrupolosa dell'architettura di un'indagine permette infatti di scoprirne i potenziali vizi strutturali (*bias*) e di valutare se i suoi risultati sono generalizzabili e applicabili nella pratica clinica.

Come l'iniziativa CONSORT in campo terapeutico (per i trials randomizzati e controllati), lo STARD nasce dall'esigenza di disporre in campo diagnostico di studi di validazione (osservazionali) eseguiti e descritti in modo coerente e uniforme per semplificare – tra l'altro - la preparazione delle revisioni sistematiche, le sintesi periodiche dei dati della ricerca in un determinato settore.

### LA RICERCA NEI SERVIZI DIAGNOSTICI

Nei servizi diagnostici gran parte della ricerca si svolge nell'alveo tradizionale delle rispettive scienze di base ed è orientata allo studio dei meccanismi fisio-patologici e delle caratteristiche degli agenti patogeni (BOX 1, studio n. 1). Gli strumenti d'indagine oggi disponibili finiscono dotati di grande obiettività e permettono una notevole approssimazione alla certezza dei risultati (5). Ne garantiscono l'affidabilità (riproducibilità e accuratezza analitica) l'aderenza ai protocolli metodologici e il controllo di qualità.

Ma le cose si complicano quando la ricerca è applicata alla pratica clinica. Quando l'oggetto di essa è l'**efficacia diagnostica** dei test, bisogna

considerare due componenti: l'accuratezza (la capacità di discriminare in una popolazione i sani dai malati) e l'utilità (la capacità di confermare o escludere una malattia nel singolo individuo).

I due requisiti sono valutati con studi osservazionali su gruppi eterogenei di persone (5); in conseguenza le conclusioni della ricerca applicata sono più esposte ai vizi d'interpretazione essendo basate su valutazioni probabilistiche e non su certezze<sup>1</sup> (BOX 1, studi n. 2 e 3).

È sull'accuratezza e l'utilità dei test che si concentra l'interesse della medicina dell'evidenza nel settore diagnostico e il campo d'azione appartiene alla ricerca applicata. Al di là dell'accuratezza analitica, della capacità di dosare sostanze, riconoscere agenti patogeni, interpretare immagini o microstrutture di organi, si tratta infatti di studiare la capacità di un test di confermare o escludere la malattia in situazioni assai diverse e con un margine di ragionevole sicurezza.

### BOX 1. STUDI SULL'*HELICOBACTER PYLORI*: IL PAZIENTE E LA MALATTIA\*

**STUDIO N. 1: Monstein H, et al. Div. of Clinical Microbiology, Molecular Biology Lab-LMO, Linköping, Sweden. Profiling of bacterial flora in gastric biopsies from patients with *H. pylori*-associated gastritis and histologically normal control individuals by temperature gradient gel electrophoresis and 16S rDNA sequence analysis. J Med Microbiol 2000; 49: 817-22.**

*Lo scopo dello studio è stato di stabilire i profili batterici in campioni da biopsia gastrica in pazienti con gastrite associata a *H. pylori* e in controlli normali. Con la gel elettroforesi a gradiente temporale di temperatura (TTGE) dei*

<sup>1</sup> Secondo la logica bayesiana (teorema di Bayes) ottenuto il risultato di un test, le probabilità di malattia dipendono dalla previsione di malattia prima del test e dal suo potere discriminante (sensibilità e specificità).

*frammenti 16S rDNA amplificati con PCR è stato dimostrato un aumento di DNA Helicobacter specifico nelle biopsie della maggior parte dei pazienti gastritici.*

**STUDIO N. 2: Queiroz DM et al. Lab. of Research in Bacteriology, Univ. Fed. de Minas Gerais, Brazil. Serological and direct diagnosis of *H. pylori* in gastric carcinoma: a case-control study. J Med Microbiol 1999; 48 (5): 501-6.**

*In questo studio si è valutata la sensibilità di alcuni test per la diagnosi d'infezione da *H. pylori* in più di 100 pazienti con Ca. gastrico e altrettanti controlli sani. Con la combinazione di diversi test (ureasi, microscopia e coltura) si è dimostrato che l'infezione da *H. pylori* è significativamente più frequente nei pazienti con Ca. gastrico che nei controlli; essa sembra pertanto associata ad un aumentato rischio di carcinoma.*

**STUDIO N. 3: Weijnen CF, et al. University Medical Center, Utrecht, Netherlands. Testing for *Helicobacter pylori* in dyspeptic patients suspected of peptic ulcer disease in primary care: a cross sectional study. BMJ 2001; 323: 71-5.**

*L'intento era di constatare se un test sierologico per l'*H. pylori* aggiunge informazioni utili all'anamnesi, al fine di una diagnosi non invasiva di ulcera peptica e dell'applicazione della strategia test&treat in medicina di base. Ciò non risulta, se non in un gruppo di pazienti dispeptici ad alto rischio di ulcera.*

\*La scelta degli articoli sull'*H. pylori* è stata guidata dall'intento d'illustrare la gerarchia delle evidenze e di commentarne la logica. Non riflette invece particolari opinioni del GdL EBM sull'argomento.

## LA VALUTAZIONE DELL'EFFICACIA

La **valutazione dell'efficacia** di un intervento medico secondo i criteri dell'*Evidence-based Medicine* (EBM) differisce dal modello tradizionale, i cui elementi chiave sono la comprensione dei meccanismi di malattia e l'esperienza, simboleggiata dall'autorità di un esperto. O meglio da un'autorità collettiva, come una società professionale che formula linee guida basate sull'opinione collegiale di un gruppo di esperti.

L'EBM ha assunto a riguardo un atteggiamento piuttosto critico. Bryan Haynes, della *McMaster University* di Toronto, ha scritto: "In campo assi-

stenziale vi possono essere più errori nelle raccomandazioni degli esperti che nelle prove ricavate da un'adeguata osservazione sistematica della letteratura. Specie quando la ricerca è applicata ad un ambiente clinico sempre più complesso e sempre più attento alle influenze esterne" (5). Era scontato che posizioni così decise suscitassero malumori e, a riguardo, l'epidemiologo canadese ha di recente commentato (6): "Il termine di *evidence-based medicine* è stato creato per indurre medici e pazienti al dovuto rispetto (parola da intendersi alla lettera) della migliore evidenza scientifica disponibile. Una definizione alternativa (per alcuni forse più accettabile) potrebbe essere *assistenza sanitaria potenziata dalla ricerca (research enhanced health care)*".

## L'OSSERVAZIONE SISTEMATICA DELLA RICERCA

Quale sia la definizione preferita, è comprensibile l'esigenza di validi criteri sia per cogliere i frutti della ricerca (da parte dei lettori di articoli scientifici), sia per offrirne (di non avvelenati) da parte di chi li scrive.

Ai primi (soprattutto ai clinici) si sono rivolti con diverse linee guida gli esperti del gruppo EBM anglocanadese che (per non contraddire se stessi) hanno lavorato in base all'evidenza di estese ricerche cliniche ed epidemiologiche. Si tratta di raccomandazioni per valutare in modo critico la letteratura primaria – i lavori originali – e ve ne sono anche per il campo diagnostico (8).

Bisognerebbe aggiornarsi continuamente, scegliendo i test che la ricerca dimostra efficaci – ovvero sia accurati, utili e non dannosi per il paziente. Si è però presto capito che la valutazione critica è un compito sovrumano per il medico pratico (4), per motivi intuibili (tempo, istruzione, dimestichezza con il linguaggio scientifico), ai quali si aggiunge per tanti l'insufficiente conoscenza dell'inglese.

È nata pertanto l'**editoria secondaria**: gruppi di addetti preparano periodicamente rassegne sintetiche della letteratura su determinati temi, utilizzando criteri codificati e standardizzati, e le pubblicano sotto forma di Revisioni Sistematiche (RS). Ne è un esempio il periodico *Clinical Evidence*, di recente tradotto in italiano (2) e dedicato finora alla terapia, ma che presto lo sarà anche ad argomenti diagnostici. La sintesi critica è un compito assai difficile perché, rastrellati in letteratura gli articoli sul tema prescelto, bisogna valutarne l'efficacia con criteri che – almeno per certi aspetti della diagnostica (impatto clinico) non sono ancora ben definiti (7). In linea generale essi ricalcano quelli esistenti per la valutazione dell'editoria primaria (8), ma in realtà sono molto

più complicati, in quanto si tratta qui di scovare prove di efficacia che nel singolo lavoro possono essere parziali o poco manifeste, ma apprezzabili se confrontate e combinate con quelle provenienti da altri articoli (7a).

### LA TRASPARENZA NELLA RICERCA

Per coloro che sono sull'altro versante (chi descrive i risultati delle indagini) si è mossa ora l'iniziativa STARD. È forse necessario ribadire che il documento riguarda gli articoli scientifici che descrivono la valutazione dell'accuratezza e della predittività di test diagnostici di qualsiasi natura (biochimici, microbiologici, *imaging*, anatomia patologica, ecc...).

Lo STARD non prende in considerazione le caratteristiche strutturali del test in sé (principi, ingredienti, metodologia analitica). Pone invece l'accento sulla correttezza del confronto "test in esame vs gold standard", sull'ambiente di reclutamento dei soggetti in esame e sullo spettro clinico che li caratterizza.

Le osservazioni più valide provengono infatti da una serie "consecutiva" di pazienti. Il termine si riferisce al reclutamento quando la malattia è sospetta, non accertata, il che riproduce le circostanze di applicazione del test nella pratica clinica, su casi lievi o gravi ed anche su persone con disturbi simili, ma di diversa natura (possibili falsi positivi). Tale tipo di studio (osservazionale-trasversale consecutivo) fornisce valutazioni del potere discriminante di un test e del suo potere predittivo assai più attendibili di quelle basate sul confronto, spesso retrospettivo, tra un gruppo di pazienti con malattia conclamata e un gruppo di volontari "normali". Architettura, quest'ultima, assai diffusa in campo diagnostico (studi caso-controllo) e affetta di frequente dallo *spectrum bias* (8 a), con una sopravvalutazione dell'accuratezza (BOX 2).

### BOX 2. L'EVIDENZA CHE CONTA AL LETTO DEL MALATO: UNA SCELTA INEVITABILE, NON UNA DISCRIMINAZIONE

I risultati di un'adeguata osservazione sistematica della letteratura devono essere utili a chi s'interroga sul modo migliore per restituire salute e benessere a una persona che soffre.

Per dare risposte concrete ai suoi quesiti sono indispensabili drastiche scelte, sia da parte dei singoli clinici che degli studiosi addetti alle revisioni sistematiche (9):

- privilegiare la produzione scientifica orientata al paziente in modo diretto, sui problemi

che s'incontrano nella pratica clinica, e che offre risultati in grado d'influire sull'assistenza (studio n.3, BOX 1);

- trascurare articoli di minor rilevanza, perché i risultati sono meno concreti o di tipo preliminare (studi caso-controllo, es. studio n.2, BOX 1), o perché l'evidenza è aneddotica (*case-report*) o perché l'indagine riguarda problemi di fondo (studio n.1, BOX 1).

**Dev'essere chiaro che la distinzione è pragmatica.** Non è messa in discussione l'utilità della ricerca di base nelle fasi di progettazione e costruzione dei test diagnostici, inoltre nelle prove preliminari di validazione. È importante però comprenderne il significato e i limiti. Esiste un'estesa e spesso dolorosa evidenza in campo terapeutico: quella delle vittime di terapie prematuramente giudicate efficaci sulla base di risultati lacunosi o preliminari.

### *Adapt or die*

Il detto richiama - con più efficacia dell'antico *publish or perish* - i rischi ai quali il processo evolutivo in atto nel *Diagnostic decision making* dei clinici espone chi lavora nei servizi diagnostici. Oggi si tende a considerare lo schema decisionale come "la revisione delle opinioni mediante informazioni imperfette", mentre "la normativa di questo processo s'ispira alla logica bayesiana" (7b). La realtà ci dice che i test di qualsiasi natura sono informazioni imperfette e chi, nei laboratori che si fregiano dell'attributo "clinico", li studia, li propone o li esegue, deve ben rendersene conto (3). Le revisioni sistematiche sono gli strumenti più idonei per realizzare il messaggio di Cochrane (al paziente conviene che la pratica medica sia influenzata dalla ricerca) (v. BOX 3), ma l'inclusione in esse di un articolo come contributo alla ricerca della verità dipendono sia dal contenuto che dal modo con il quale viene esposto. In questo senso l'iniziativa STARD appare di certo un avvenimento da segnare *cum albo lapillo*.

Se è permesso rivolgere un invito ai colleghi italiani, è quello di scrivere anche nella nostra lingua, secondo gli standard auspicati dalla comunità scientifica. *Pubmed*, il più importante motore di ricerca nel mondo, inserisce nei propri archivi anche gli *abstract* di lavori originali in lingue diverse dall'inglese e v'è chi precisa - a garanzia dell'assenza di un *Babel bias* in una rassegna sistematica: "*we had no language restrictions*"<sup>2</sup>. Parlare nel proprio idioma, oltre che più semplice e conveniente, è anche raccomandabile perché apre la strada ai meno esperti e stimola il con-

<sup>2</sup> H. Honest et al. *BMJ* 2002; 325: 301-11.

fronto e il dibattito, la vera linfa vitale della ricerca. E se i bilanci dell'editoria nelle società scientifiche italiane sono in rosso, è pur vero che "l'editoria è parte integrante di esse e il suo mantenimento deve essere garantito" (1).

Direi il potenziamento, anche a costo di sacrifici su altri aspetti di facciata. Ai nostri periodici dovrebbero essere conferiti i requisiti che ne rendano il contenuto inseribile nelle banche dati internazionali. Molto di quanto viene ora a mancare sarebbe recuperato, in termini di prestigio, e non solo. I nostri ricercatori sono in grado di portare le testate scientifiche italiane nell'arena internazionale: il loro patrimonio d'ingegno lo permette, in aggiunta alle pur meritevoli imprese da cavalieri solitari sotto insegna straniera.

### BOX 3. COME L'EVIDENZA PUÒ INTEGRARE L'ESPERIENZA (in attesa del nuovo)

**Roberts AP, et al. Centre for Health Studies, University of Durham, UK. Tests for *Helicobacter pylori* infection: a critical appraisal from primary care. Fam Pract 2000 Aug;17 Suppl 2:S12-20.**

I test per *H. pylori* sono assai richiesti in medicina di base. Questa **revisione sistematica** è una valutazione critica dell'accuratezza di tali test e delle iniziative idonee al trattamento dei pazienti con sospetta infezione.

Le conclusioni sono che i test per *H. pylori* sono utili in medicina di base quando **le probabilità pre-test d'infezione** non sono né troppo basse né troppo alte. Ciò significa che i test possono non essere utili per uno screening, ma solo per una diagnosi differenziale.

Al di fuori della fascia di probabilità pre-test 20-80% le probabilità di ottenere risultati falsamente positivi o negativi è molto alta. Sono inaffidabili i risultati negativi quando le probabilità pre-test sono >80% e i positivi, quando le probabilità pre-test sono <20%. Perciò, se le probabilità d'infezione sono alte a sufficienza (>80%), i pazienti dovrebbero ricevere il trattamento di eradicazione senza effettuare test. La preselezione clinica dovrebbe limitare l'uso degli esami ai pazienti con probabilità pre-test entro il citato intervallo 20-80%.

Tutte le esistenti raccomandazioni sull'uso di questi esami, specie di quelli del tipo *Near-patient test*, dovrebbero essere riconsiderate con attenzione.

Questa valutazione critica conferma però le linee guida dell'*European Society for Primary Care Gastroenterology* (basate sul **consensus**)

sull'uso dei test per *H. pylori* in medicina di base.

NOTA (data dell'invio del lavoro in redazione) "Successivamente alla presentazione dell'articolo a questa rivista (agosto 2002), numerose giornali scientifici hanno segnalato in contemporanea l'iniziativa STARD, mettendo in rilievo le sue potenziali conseguenze favorevoli sulla qualità della ricerca scientifica in campo diagnostico e sulle revisioni sistematiche ad essa dedicate: *Amer J Clin Pathology, Ann Int Medicine, British Med J, Clin Chemistry, J Clin Pathology, Radiology, The Lancet* ed altri, nel loro primo numero del 2003".

### BIBLIOGRAFIA

1. Clerici P. Relazione del tesoriere. *Notizie AMCLI* 2002; 5 (7-8): 3.
2. *Clinical Evidence*, una sintesi delle migliori prove di efficacia. Ed. Italiana, 2001 n.1 (a cura del Min. della Salute) 2 a 3.
3. Garber AM, Mulrow C, Sox HC. Articles on diagnostic tests: a new series and an invitation to authors (Editorial). *Ann Int Medicine* 2002; 137: 288-9.
4. Guyatt GH, Meade MO, Jaeschke RZ, Cook DJ, Haynes RB. Practitioners of evidence based care (Editorial). *BMJ* 2000; 320: 954-5.
5. Haynes RB. What kind of evidence is it that Evidence-Based Medicine advocates want health care providers and consumers to pay attention to? *BMC Health Services Research* 2002; 2: 3 [www.biomedcentral.com/1472-6963/2/3](http://www.biomedcentral.com/1472-6963/2/3)
6. Haynes RB, Devereaux PJ, Guyatt GH. Physicians' and patients' choices in evidence based practice (Editorial). *BMJ* 2002; 324: 1350.
7. Knotterus JA. The evidence base of clinical diagnosis. 2002 *BMJ Bookshop* [www.bmjbooks.com](http://www.bmjbooks.com) In questo volume: (a) Devillé WL, Buntinx FB. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests (a pag. 145). Il capitolo è disponibile sotto forma di articolo a: [www.biomedcentral.com/14712288/2/9](http://www.biomedcentral.com/14712288/2/9). (b) Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: a selective review of the cognitive research literature (a pag. 179).
8. Jaescke RZ, Guyatt GH, Sackett D, for the Evidence-Based Medicine Working Group. Users' Guides to the medical literature. III. How to use an article about a diagnostic test. (a) Are the results of the study valid? *JAMA* 1994; 271: 389-91. (b) What Are the results and will they help me in caring my patients? *JAMA* 1994; 271: 703-6.
9. Shaughnessy AF, Slawson DC. POEMs: patient-oriented evidence that matters. *Ann Intern Med.* 1996;125: 605-13. Si può anche consultare il *tutorial "Infomastery"* (Univ. del Michigan, USA) a: <http://www.poems.msu.edu/InfoMastery/default.ht>.
10. Smith R. Diagnosis, diagnosis, diagnosis (Editorial). *BMJ* 2002; 324.
11. The STARD Group. Towards complete and accurate reporting of studies on diagnostic accuracy: the STARD initiative. <http://www.consort-statement.org/stardstatement.htm>.