

Cannoni d'agosto

Blog di mezza estate

G. Giocoli



"*Likelihood ratio*: non usateli per calcolare numeri senza significato, ma come misura della validità dei test. Considerate i loro limiti di confidenza come misura dell'accuratezza di questa stima ...".

Leggo sul *British Medical Journal* del 7 agosto questo incisivo commento di un oncologo canadese (1) ad un articolo sul corretto impiego dei *Likelihood ratio* (LR) (2) e l'associo alle voci di iniziative - attuate o in via di attuazione in alcuni laboratori italiani - per la refertazione di questi indici assieme ai risultati dei test.

A mio parere, i rilievi del ricercatore hanno colpito nel segno e, per quanto io possa aver capito dalla quasi quotidiana consultazione della letteratura sull'argomento, l'idea di segnalare i *likelihood ratio* nei referti di laboratorio mi sembra prematura, inopportuna e rischiosa.

Se - nell'attuale realtà italiana - questi indici (e relative istruzioni per l'uso) saranno a disposizione di tutti, quali saranno le conseguenze per i pazienti? Quale sarà il futuro dei *likelihood ratio* nella diagnostica? Li attende il riconoscimento di indicatori di qualità o una squallida fine nel dimenticatoio delle cose inutili? Oppure una sinistra fama di apportatori di danno iatrogeno?

Likelihood ratio

I *likelihood ratio* sono indici che esprimono la potenza (accuratezza) di un test diagnostico, ossia la capacità di distinguere chi ha una malattia da chi non ce l'ha. Qualsiasi test dicotomico (con risultato positivo/negativo, presente/assente, o simili) può dare risultati veri o falsi e, se ammettiamo che il numero UNO esprima l'equivalenza di queste possibilità, i test clinicamente più utili sono quelli caratterizzati da LR che si allontanano dall'unità, in direzione dell'infinito per i risultati positivi, in direzione dello zero per i risultati negativi. Il perchè è facile da comprendere, esprimendo entrambe le tendenze lo sbilanciamento in favore dei risultati veri nei confronti di quelli falsi.

Per le loro qualità, i *likelihood ratio* rendono realizzabile l'enunciato di Bayes "Le probabilità che il risultato di un test corrisponda o meno alla presenza della malattia dipendono dalle probabilità pre-test e dalla potenza del test".

Bayes: quando e come

Sackett ha auspicato che il medico, alle prese con il dilemma diagnostico, adoperi i *likelihood ratio* di un test per modificare le sue ipotesi sulla malattia. Egli può utilizzare un calcolo aritmetico o il nomogramma di Fagan oppure, in casi particolari, gli *SPin* o gli *SNout*.

SPin e *SNout* sono proprietà conferite ad alcuni test da una specificità o una sensibilità talmente prossime al 100% (o coincidenti con esso) da renderne assiomatici i risultati positivi o negativi, per cui una certa diagnosi sarebbe da essi confermabile o escludibile senza necessità di ulteriori indagini. Tuttavia, la validità di molti *SPin* e *SNout* riportati nel libro di Sackett e nel sito web del Centro EBM di Oxford risulta chiaramente contestabile, come provato dall'epidemiologo svizzero Daniel Pewsner (2).

Ma non basta: Andrew Robinson, il citato oncologo di Vancouver, ha esteso le sue critiche all'uso attuale e generalizzato dei *likelihood ratio* – grandi o piccoli che siano – per calcolare il valore predittivo dei test al letto del malato: i loro limiti di confidenza sono così ampi e le incertezze sulle probabilità pre-test così sensibili da annullare il significato pratico di tale procedura (1).

Statu quo

La logica bayesiana e i relativi strumenti di attuazione – LR compresi – sono oggetto di intense ricerche, soprattutto nei Paesi dell'area anglosassone – Australia, Canada, Regno Unito, USA – e dell'Europa settentrionale (Paesi Bassi). Non so quanto i *likelihood ratio* siano realmente entrati nella locale pratica clinica, ma è probabile che minoranze elette ne facciano uso.

Il problema è che la stima dell'accuratezza di un test (validazione) deve poggiare su solide fondamenta scientifiche edificate con studi di disegno appropriato, concernenti sia la validazione interna del test (corretto confronto con un idoneo gold standard, popolazione appropriata, numerosità adeguata), sia la sua validazione esterna (generalizzabilità e applicabilità dei risultati). Sono attualmente ben pochi i test che soddisfano a pieno tali requisiti.

I recenti protocolli STARD e QUADAS forniscono indicazioni sulle corrette procedure da seguire ed è probabile, dicevo, che in alcuni centri si ottengano stime affidabili dell'accuratezza (con relativi limiti di confidenza) dei test effettuati nei locali laboratori. Se a ciò si uniscono la disponibilità di rilievi epidemiologici e di regole predittive per definire il secondo termine dell'enunciato bayesiano (le probabilità pre-test), non può stupire la notizia che in ospedali australiani o canadesi i clinici usano il palmare o il diagramma di Fagan per "aggiornare" le loro opinioni sulle condizioni del malato all'arrivo (in tempo reale) del risultato di un test.

Ed è certo che in diversi centri (specie universitari) del Nordamerica sono disponibili mini-protocolli (CATs) per la diagnosi rapida di malattie, anche da infezione (es. le meningiti). Nei CATs (*Critically Appraised Topics*) sono riportate in sintesi le prestazioni di alcuni test (sotto forma di LR) per aiutare i medici, specie nei reparti di emergenza, a ottenere risposte critiche mediante un semplice calcolo probabilistico.

Tali procedure restano tuttavia fuori portata dei più, sconosciute o addirittura avversate. Non tanto (in quest'ultima evenienza) per i motivi esposti, che presuppongono una non superficiale conoscenza dell'argomento, quanto per la riluttanza dei medici ad abbandonare tradizionali percorsi diagnostici ispirati alle opinioni personali piuttosto che all'evidenza obiettiva.

How good is a test?

Le argomentazioni di Pewsner e di Robinson non appartengono ad una delle tante campagne di demolizione dell'EBM, ma invocano un uso più adeguato alle conoscenze attuali di uno dei suoi strumenti più preziosi.

"... piuttosto serviamoci dei *likelihood ratio* per valutare la bontà dei test" (*as a barometer of how good a test is*)," esorta Robinson.

I migliori frutti della ricerca per una corretta definizione della potenza di un test - sono le Revisioni sistematiche (RS) e delle metanalisi in campo diagnostico. La revisione sistematica è una "summa" dei dati scientifici su un dato argomento, ottenuta dalla selezione della letteratura secondo criteri standardizzati e riproducibili. In molte RS sono ormai riportati sia la sensibilità e la specificità dei test che i relativi LR positivi e negativi, con i relativi limiti di confidenza.

Cito esempi di patologie per le quali sono state compiute profonde analisi critiche dei test pertinenti mediante le RS, spesso rivedute e chiosate da eminenti istituti quali il CRD dell'università di York: le infezioni genitali, ematiche, urinarie, da HIV1, il cancro della prostata e del colon, il diabete, l'insufficienza cardiaca, le fratture del malleolo e del piede ("regole di Ottawa"). Dopo anni dedicati alla revisione dei dati in campo terapeutico, anche la *Cochrane Collaboration* ha deciso di aprire una sezione per le RS in diagnostica.

Nelle revisioni sistematiche le stime dei *likelihood ratio* e dei relativi limiti di confidenza forniscono informazioni rilevanti per comprendere potenzialità e limiti dei test, che si traducono in elementi di giudizio e di confronto per istituzioni e singoli professionisti intenzionati a migliorare il proprio armamentario per la diagnosi e lo screening delle malattie.

Ma si è appena all'inizio di un duro cammino. Le "spietate" procedure di revisione hanno messo in luce enormi carenze nel campo diagnostico, da tempo bollate al vetriolo dalla rivista britannica *Bandolier*. L'aspetto positivo della vicenda è un vivace impulso alla standardizzazione delle procedure di validazione dei test, ad esempio i citati protocolli STARD e QUADAS, o altre iniziative dell'FDA statunitense. L'intento è di elevare il livello degli studi di validazione sì da rendere più agevole ed efficace il lavoro dei revisori.

Chi legge il barometro?

A chi è dunque rivolto l'invito di Robinson? A chi spetta la lettura del "barometro dei test"? Il ricercatore non lo dice, ma suppongo che la risposta sia "Ai clinici". Credo sia giusto. Ma, a mio parere, ad essi vanno affiancati coloro che i test li propongono e li eseguono: gli specialisti dei servizi diagnostici.

Proprio nello scorso luglio mi son permesso di proporre al *British Medical Journal* un *Future Theme Issue* sull'applicazione dell'EBM nei laboratori ospedalieri (*The EBM in the basement*). Appare opportuno che, nei nostri, non ci si affidi soltanto alla luce di lontane stelle polari. Le revisioni sistematiche sono ancora poche e solo alcune includono risultati ottenuti nel nostro Paese.

Ogni centro di riferimento dovrebbe avviare una supervisione critica dei test, ad iniziare dai più importanti, per stabilire la validità locale di dati ottenuti altrove, spesso in altri continenti. A questo scopo esprimere la potenza dei test sotto forma di *likelihood ratio* può essere assai utile, ma lasciandone l'interpretazione agli addetti ai lavori. Questi, a loro volta, dovrebbero avvalersi di tecniche per il trasferimento della conoscenza alla compagine sanitaria e ai pazienti (*knowledge translation, knowledge brokering, shared-decision making*).

"I fear me both are false" (Shakespeare, Richard III, I-ii)

Il problema, a mio modesto parere, è che non è possibile eseguire siffatti studi senza un organico e comune impegno di laboratoristi, clinici, epidemiologi e altri specialisti: basti solo pensare all'uso di gold standard clinici (o misti) per alcune malattie da infezione per le quali la coltura era considerata fino a ieri il non plus ultra dei riferimenti.

Certo, ormai le regole sono note ma (secondo me) gli ostacoli per la loro applicazione qui da noi risiedono non tanto nella complessità degli adempimenti quanto nella nostra atavica carenza di requisiti per assolverli. Perché sono ineludibili (cito a caso) un'oculata gestione delle risorse umane e del loro tempo, l'ordinato sviluppo di coerenti e stabili reti operative, un'affidabile organizzazione (anche spicciola). Pratiche non certo comuni dietro la facciata dei nostri palazzi della salute. Ma anche un serio impegno, coordinato, tenace e spesso oscuro e gregario, attributi agli antipodi dell'improvvisazione, del protagonismo sensazionalistico e del campanilismo che stimolano e modellano molte iniziative nel Bel Paese.

E' questo il motivo della mia forte contrarietà ad una prematura immissione dei *likelihood ratio* nella pratica clinica da parte dei laboratori italiani. Un conto è - in parallelo con il lavoro di revisione critica - ragionare insieme ai medici (e ai pazienti) su rischi e vantaggi di una procedura diagnostica, un conto è ritenersi "all'avanguardia" e ostentare un "prodotto" con la "griffe" di Fagan. E chi controllerebbe le inevitabili emulazioni, nel quadro della legittima concorrenza?

A chi mi conosce come fautore dei *likelihood ratio* potranno sembrare strane e incoerenti queste critiche. È che vorrei per questi numeretti un adeguato posto nello strumentario diagnostico, non il futuro ostracismo per malasania.

La Spezia, ferragosto dell'anno 2004

1. Andrew G Robinson. Out with Snout. Rapid response to 'Daniel Pewsner, et al. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ 2004; 329: 209-213 (7 August 2004)
2. Daniel Pewsner, et al. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. BMJ 2004; 329: 209-213