Confronto di metodi: accordo e associazione tra i valori ottenuti con due metodi differenti

Carlo Franzini

Università degli Studi di Milano. Dipartimento di Scienze Cliniche L.Sacco. Milano

L'ultimo fascicolo di BC dell'anno 2000 riportava una breve nota simolante di Marco Rocchi dal titolo: "*Nota in margine al congresso SIBioC 2000: sull'uso distorto del coefficiente di correlazione nel confronto tra metodiche*" (Biochim Clin 2000;24:509-510). L'articolo, che in realtà esaminava un problema da anni discusso nella letteratura biochimico-clinica, e che ci si sarebbe sentiti autorizzati a considerare definitivamente risolto nell'anno duemila, partiva dalla constatazione che in occasione del 32° Congresso Nazionale della nostra Associazione (Rimini, settembre 2000) in ben 29 poster su i 102 esposti in un sessione veniva ripetuto l'errore metodologico di giudicare l'attendibilità analitica relativa di due metodi confrontati basandosi esclusivamente sul calcolo del coefficiente di correlazione "r".

L'anno scorso (settembre 2002) siamo ritornati a Rimini per il nostro incontro nazionale: non so se Rocchi c'era, e sa ha voluto ripetere l'esperienza. Non ho esaminato con tanta attenzione i poster esposti, ma ho consultato i riassunti delle comunicazioni libere (presentazioni poster), pubblicati sul fascicolo di BC dedicato al Congresso [Biochim Clin 2002;26(3):185-327]. Mi sono detto in partenza: mi sembra interessante documentare il sicuro effetto che, insieme ai lavori in merito che ancora compaiono in letteratura, avrà avuto il rihiamo all'ordine di Rocchi, anche in relazione alla circostanza della ripetizione del congresso nella medesima sede.

Il risultato della "scansione" del citato fascicolo è stato: su 282 riassunti pubblicati, dei quali un certo numero specificamente relativi a confronti intermetodi, 15 presentavano la medesima scorrettezza metodologica (dal punto di vista statistico) di valutare la concordanza tra metodi con il solo calcolo di "r". Va da sè che la frequenza della scorrettezza da me rilevata, spero con accettabile esattezza, non è confrontabile con quella di Rocchi. Non è uniforme il campionamento, ed è anche possibile che mentre nel riassunto fosse riportato solo il valore di "r" nel poster per esteso fossero poi indicati anche altri parametri statistici. In ogni caso, e anche se sul piano della frequenza la situazione apparirebbe migliorata, mi sembra che il problema debba essere ancora discusso.

Mi sembra che gli spetti critici della questione siano molto ben chiariti in una Lettera recente di D. G. Altman e J. M. Bland (*Commentary on quantifying agreement between two methods of measurement.* Clin Chem 2002;48:801-802), della quale qui riporto un brano, che ho cercato di tradurrre nel modo più letterale possibile. "... *la massima parte degli autori usavano il coefficiente di correlazione di Pearson. Ci sembra ovvio che questo metodo non valuta l'**accordo (agreement)** ma la **associazione (association)**, e che una elevata correlazione non era garanzia di buon accordo*".

In sostanza, Altman e Bland ribadiscono che il coefficiente di correlazione è un parametro di associazione tra due serie di misure, ma non fornisce nessuna informazione sull'accordo: Nel contesto specifico del confronto tra due metodi, ci informa su quanto le due serie di valori sono associate, ma non su come i valori di un metodo sono in accordo con quelli dell'altro: questa peraltro è la informazione che più ci interessa. Per illustrare questi concetti, ho predisposto alcuni semplici esempi, basati su reali valori sperimentali, originali e in parte successivamente modificati. Come misure di associazione sono stati

considerati il coefficiente di correlazione (di Pearson, "r") e la deviazione standard residua (o "deviazione standard dei residui", Syx), questa ultima considerata una migliore misura di associazione; come misure di accordo i parametri della equazione di regressione lineare calcolata con il metodo dei minimi quadrati (regressione ordinaria).

I dati originali si riferiscono ad un confronto di due metodi (metodo-x e metodo-y) per la misura della albumina del sangue, applicati ad un gruppo di 63 sieri (umani, freschi) la cui concentrazione di albumina, misurata con il metodo-x, variava da 2,41 a 4,11 g/dL. I risultati delle misure sono presentati graficamente, come grafico di dispersione, nella figura 1; i relativi parametri di associazione e di accordo nella tabella 1, prima riga. L'accordo ci dice che per un valore di 4,50 g/dL misurato con il metodo-x il valore più probabile misurato con il metodo-y è di 4,49 g/dL, mentre l'associazione ci dice che il 95% dei valori misurati con il metodo-y (sempre per y = 4,50) cadrà nell'intervallo 4,33-4,65 g/dL. Se ciò sia soddisfacente ai fini medici è un altro discorso. In genere un accordo di questo tipo, confermato da un valore di r pari a 0,988, è considerato ottimo. Il buon grado di accordo e di associazione sono confermati dall'esame visivo del grafico di dispersione: la variabilità dei singoli punti attorno alla retta di equivalenza (y = x) ed alla retta di regressione appare assi contenuta.

Se si modificano i medesimi dati, simulando due livelli di crescente imprecisione nelle misure con il metodo-y (errore casuale 1 e 2), già la osservazione del grafico di dispersione evidenzia la aumentata dispersione dei singoli punti alla retta di regressione, che peraltro non si scosta molto dalla retta di equivalenza (figure 2 e 3). I parametri di associazione (tabella 1, seconda e terza riga) denunciano numericamente il medesimo

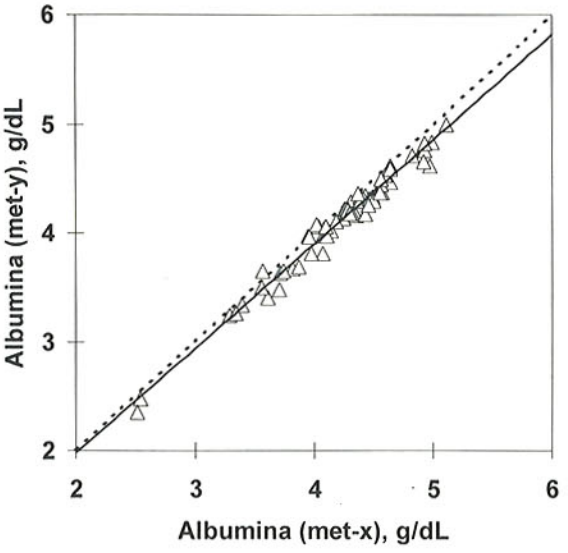

Figura 1
Grafico di dispersione dei dati orginali (n = 63). Le linee continua e tratteggiata rappresentano, rispettivamente, la retta di regressione dei dati sperimentali (regressione "ordinaria") e la retta di equivalenza (y = x

Tabella 1
Analisi statistica dei dati di confronto intermetodi, applicata ad un gruppo di risultati analitici originali (n = 63), ed ai medesimi dopo simulazione di errore casuale (due simulazioni) e di errore sistematico (2 simulazioni)

Dati	Figura	Parametri di accordo		Parametri di associazione	
		Pendenza # g/dL	Intercetta#	r	Syx g/dL
Originali	1	0,98 (ns)	0,07 (ns)	0,988	0,08
Errore Casuale 1	2	1,09 (ns)	-0,32 (ns)	0,866	0,33
Errore Casuale 2	3	0,94 (ns)	0,36 (ns)	0,557	0,72
Errore sistematico costante	4	0,96 (ns)	-0,69 (*)	0,988	0,08
Errore sistematico proporzionale	4	1,34 (*)	0,10 (ns)	0,986	0,11

(#) regressione lineare "ordinaria" (minimi quadrati)
(*) differenza da 0 (intercetta) o da 1 (pendenza) statisticamente significativa
(ns) differenza da 0 (intercetta) o da 1 (pendenza) statisticamente non significativa

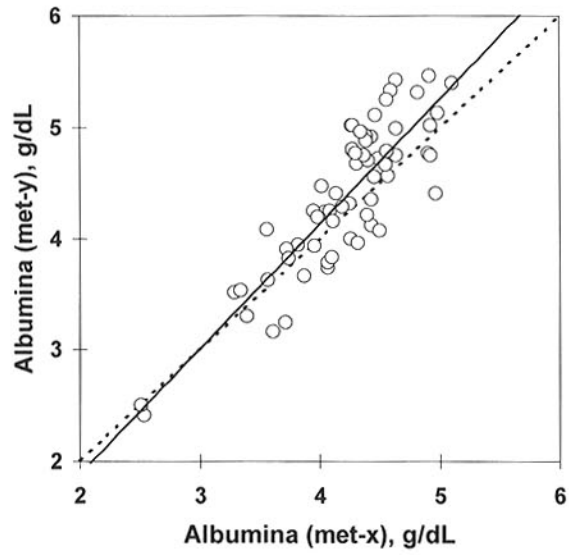

Figura 2
Come figura 1, dopo simulazione di un moderato aumento di imprecisione nel metodo-y

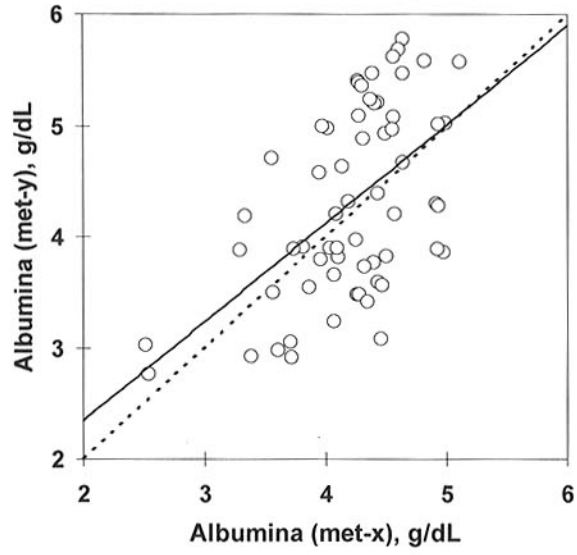

Figura 3
Come figura 1, dopo simulazione di un più accentuato aumento di imprecisione nel metodo-y

fatto: r cala visibilmente (da 0,988 a 0,866, a 0,557), mentre Syx aumenta (da 0,08 g/dL a 0,33 g/dL a 0,72 g/dL). Peraltro, i singoli punti si distribuiscono quasi simmetricamente sopra e sotto la retta di equivalenza: i parametri di accordo si discostano alquanto da quelli ottenuti per i dati originali, senza tuttavia raggiungere la significatività statistica. Si osservi che pendenza maggiore di uno e intercetta negativa si compensano a vicenda nella zona centrale dei valori, come pure si compensano pendenza inferiore a 1 e intercetta postiva. Per un valore di 4,50 g/dL misurato con il metodo-x, in presenza dell'errore casuale 1 e dell'errore casuale 2 il metodo-y misurerà, rispettivamente, 4,58 g/dL e 4,39 g/dL. Si può quindi concludere che in questi due casi l'associazione è debole ma l'accordo (medio) ancora non così compromesso.

I dati originali sono stati successivamente modificati in due maniere differenti, simulando nelle misure con il metodo-y la presenza di un errore sistematico costante (negativo) e, rispettivamente, di un errore sistematico proporzionale (positivo). I valori sono riportati graficamente nelle figura 4, in confronto con i dati originali. Si osserva il tipico comportamento delle rette di regressione dei dati modificati ad includere un errore sistematico, rispettivamente costante e proporzionale, in confronto alla retta di equivalenza. Tuttavia, in questi due casi i singoli dati si dispongono con piccolissima dispersione attorno alla relativa retta di regressione: i parametri di associazione (r e Syx, quarta e quinta riga) non differiscono sostanzialmente da quelli calcolati per i dati originali. Al contrario, i parametri di accordo raggiungono valori differenti da quelli dei dati originali,

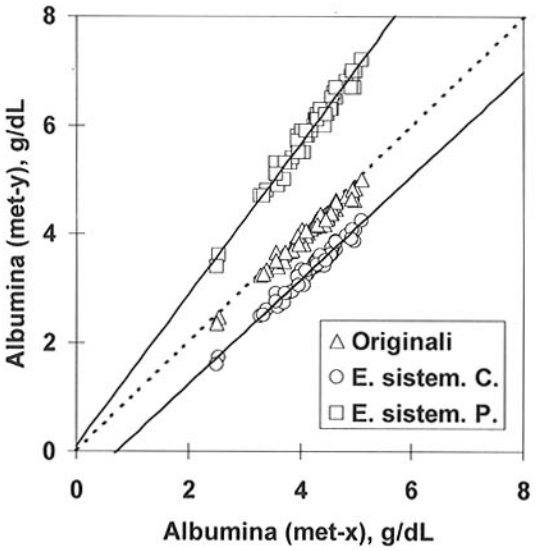

Figura 4
Come figura 1, dopo simulazione di un errore sistematico costante (tondi) e di un errore sistematico proporzionale (quadrati) nel metodo-y. Sono anche riportati i dati originali (triangoli), per confronto

statisticamente differenti da 0 per quanto concerne la intercetta (in presenza di errore sistematico costante) e da 1 per quanto concerne la pendenza (in presenza di errore sistematico proporzionale). Si può quindi conludere che in questi due casi si ha forte associazione ma accordo carente.

In generale si può pertanto confermare che associazione ed accordo variano indipendentemente: di conseguenza, per una descrizione completa dei risultati di confronti tra metodi differenti per la misura della medesima grandezza, si devono riportare misure di accordo e di associazione. In particolare, il solo "r" dà una valutazione assolutamente incompleta dei risultati di comparazione.

In linea generale non è da sottovalutare l'importanza della presentazione grafica. Lo scopo della figura è di consentire di giungere rapidamente ad una conclusione preliminare corretta, prima di indagare il significato dei valori numerici delle differenti analisi statistiche. Credo che chiunque sarebbe soddisfatto dei dati rappresentati dalla figura 1, mentre avrebbe forti perplessità su quelli rappresentati nelle figure 2, 3 e 4. Secondo alcuni il "grafico delle differenze" darebbe una rappresentazione più intuitiva della struttura dei dati e delle loro eventuali differenze. In ogni caso è importante rammentare che anche le figure devono essere costruite correttamente: con piccoli artifici è possibile fare apparire (ad un primo esame) buono l'accordo cattivo e viceversa (C. Franzini. Correct graphic presentation of method comparison data. Clin Chem Lab Med 2001;39:460-4601.).

Cinque brevi "appendici".

1) E' noto che J.M. Bland e D.G.Altman ("Statistical methods for assessing agreement between two methods of clinical measurement". Lancet 1986;I:307-310) hanno proposto il metodo della distribuzione delle differenze (il cui corripondente grafico è il "*difference plot*" su menzionato) come più efficace metodo per evidenziare l'accordo tra due metodi. I vantaggi e gli svantaggi di tale appproccio, nei confronti della regressione lineare, sono spesso discussi nella letteratura recente. Per maggiore sicurezza, o per avere più evidente supporto grafico-statistico alla interpretazione dei dati sperimentali relativi a confronto metodi, con sempre maggiore frequenza si osservano lavori nei quali i risultati vengono presentati nei due modi (regressione/grafico-di-dispersione e distribuzione-delle-differenze/grafico-delle-differenze). In linea di massima sembra si possa ritenere che entrambi i metodi grafico/statistici consentono di arrivare a conclusioni sostanzialmente equivalenti, se applicati ed interpretati correttamente. E' importante comunque rammentare che l'utilizzo del solo "r" non consente conclusioni corrette.

2) In associazione alla analisi della regressione si utilizza spesso "r" come parametro di associazione. Si ritiene che "Syx" sia un parametro migliore: oltre tutto essa offre una misura "tangibile" della dispersione dei singoli valori intorno alla retta di regressione, espressa nelle stesse unità in cui è espressa la grandezza misurata.

3) Benchè qui, a scopo esemplificativo, si sia impiegato il modello "ordinario" di regressione lineare, si deve rammentare che esistono modelli che meglio si adattano alla

struttura dei dati sperimentali, tra i quali la regressione di Deming, la regressione non-parametrica di Passing e Bablok, la componente principale standardizzata.

4) La applicazione anche del più sofisticato modello statistico non migliora dati ricavati dalla applicazione di un disegno sperimentale non corretto. Per quanto concerne il confronto di metodi, ai fini della attendibilità dei calcoli statistici sono fattori importanti del disegno sperimentale la numerosità del campione e l'intervallo di valori che assume la variabile: entrambi, più ampi sono meglio è (K. Linnet. Necessary sample size for method comparison studies based on regression analysis. Clin Chem 1999;45:882-894.).

5) Si è precedentemente menzionato come la decisione sulla applicabilità dei dati ad un contesto medico sia una operazione differente da quella di stabilire la attendibilità analitica dei dati medesimi, per esempio mediante confronto di metodi. E' nota peraltro la attuale tendenza a spostare, giustamente, la attenzione dall' "*analiticamente corretto*" al "*clinicamente utile*". Affrontato in senso tradizionale, il confronto tra metodi comprende due passaggi: a) la descrizione, statisticamente corretta, delle differenze intermetodi; b) la valutazione dell'impatto medico delle eventuali differenze. Recentemente è stata suggerita la possibilità di adottare approcci statistici atti a fornire l'informazione necessaria in un unico passaggio (R. Haeckel, W. Wosnick, I. Puentmann. Discordance rate, a new concept for combining diagnostic decisions with analytical performance characteristics. 1. Application in method or sample system comparisons and in defining decision limits. Clin Chem Lab Med 2003;41:347-355).