Note statistiche. Significatività statistica e rilevanza clinica: due facce della stessa medaglia?

Ettore Marubini[1], Bruno Mario Cesana[2]

[1]*Istituto di Statistica Medica e Biometria "G.A. Maccacaro", Cascina Rosa, Università degli Studi, Milano,*
[2]*Sezione di Statistica Medica e Biometria, Dipartimento di Scienze Biomediche e Biotecnologie, Facoltà di Medicina, Università degli Studi, Brescia*

Key words:
Clinical relevance; Sample size calculation; Statistical significance.

The aim of this statistical note is to draw the attention of the cardiologists to the aspects pertinent to the clinical relevance of the result of a clinical controlled randomized trial. The difference between clinical relevance and statistical significance is shown by using the results of GISSI and GUSTO III clinical controlled randomized trials.

(G Ital Cardiol 2007; 8 (6): 349-352)

Ricevuto il 12 febbraio 2007; nuova stesura il 13 marzo 2007; accettato il 14 marzo 2007.

Per la corrispondenza:

Prof. Ettore Marubini

Istituto di Statistica Medica e Biometria "G.A. Maccacaro" Università degli Studi Via Venezian, 1 20133 Milano E-mail: ettore.marubini@unimi.it

Nella precedente nota statistica[1] si è visto che l'aggettivo "significativo" è frequentemente utilizzato nelle sperimentazioni cliniche controllate randomizzate (SCCR) per presentare il risultato di test di ipotesi; esso non deve essere confuso con "rilevante" o "importante" dal punto di vista biologico o clinico. Di fatto un risultato "statisticamente significativo" può essere "clinicamente rilevante" o "irrilevante". Per familiarizzare il lettore con il differente significato che si vuole attribuire ai termini "significativo" e "rilevante" ricorriamo ad un esempio paradossale.

Due amici si incontrano dopo un lungo lasso di tempo e vogliono celebrare l'evento bevendo un caffè: poiché sono ambedue avari decidono di lanciare una moneta per stabilire chi pagherà. Il signor A, che è un medico, ne estrae una dalla tasca, ma il signor B, che è uno statistico dilettante e per di più diffidente, non vuole utilizzarla fino a che non si dimostri che la moneta non sia difettosa; in altri termini, vuole che in una lunghissima serie di lanci, "testa" appaia tante volte quante "croce". Quindi i due amici, ai quali il tempo libero non fa difetto, programmano un esperimento: essi lanceranno la moneta 10 miliardi di volte; concluderanno che la moneta è difettosa oppure no in base al risultato del test di ipotesi che eseguiranno al termine dell'esperimento. Essi non possono che assumere per vera l'ipotesi "nulla" (H_0) che la moneta non sia difettosa o detto altrimenti che sia perfettamente bilanciata, cosicché la probabilità (π) di comparsa della testa sia pari a 0.5 ($^1/_2$), formalmente H_0:$\pi = 0.5$.

Se questa ipotesi è vera, in 10 miliardi di lanci ci si attende che la testa compaia 5 miliardi di volte. Peraltro, i due amici sanno che nel lancio di una moneta perfettamente bilanciata la comparsa di testa piuttosto che croce è espressione *solo* della variabilità casuale, componente ineliminabile di ogni esperimento; quindi essi sono pronti ad osservare come risultato dell'esperimento un numero di teste che sia pari o si scosti leggermente da 5 miliardi.

Di fatto il risultato dell'esperimento dà 5 000 100 000 teste cui corrisponde una frequenza relativa (fr) di comparsa di testa: fr = 0.50001 (5 000 100 000/10 000 000 000).

Come ricordato nella sesta di queste note[1] l'esperimento fornisce anche una misura appropriata dell'errore dovuto alla variabilità casuale, e ciò consente di costruire un test atto a saggiare l'ipotesi nulla summenzionata. Dato il numero elevato di lanci, è possibile saggiare l'ipotesi nulla per mezzo del test z che ammette distribuzione gaussiana. I due amici calcolano z = 2 e, dalle tavole della distribuzione gaussiana, ricavano un livello di significatività <5%, pari a p = 0.0455 (test a due code).

Tale risultato, statisticamente "significativo", implica il rifiuto dell'ipotesi nulla che porta a concludere che la moneta non sia perfettamente bilanciata. Questa conclusione è "rilevante" ai fini della scommessa per il caffè? È ragionevole che il signor B rifiuti l'uso della moneta del signor

A dato che la differenza (tra la risposta fornita dall'esperimento e il suo valore atteso in base all'ipotesi nulla) d = 0.50001 – 0.50000 = 1 x 10^{-5} è estremamente piccola e quindi verosimilmente ininfluente sulla scommessa che è basata su un solo lancio? La risposta non può che essere negativa e l'apparente contraddizione con il risultato del test di ipotesi può divenire un'*impasse* per i due amici. Peraltro essi sanno (anche il medico!) che si può dimostrare statisticamente significativa ogni differenza (d) empirica, indipendentemente da quanto piccola sia tale differenza purché la dimensione dello studio sia sufficientemente grande. Non rimane loro che riconoscere di aver fatto un errore di programmazione dell'esperimento e precisamente di non averne calcolato la dimensione in modo appropriato.

Nell'ambito della sperimentazione clinica al lettore cardiologo è noto come gli articoli che riferiscono i risultati di SCCR riportino, nella sezione "Materiali e metodi", un paragrafo dedicato al calcolo della dimensione dello studio.

In una SCCR di superiorità in cui si voglia saggiare la riduzione della mortalità con un nuovo trattamento rispetto al trattamento di controllo, il calcolo della dimensione dello studio si basa sulla conoscenza delle seguenti quantità:
- stima *affidabile* della mortalità nel gruppo di pazienti sottoposti al trattamento di controllo, derivata dalla esauriente ricerca bibliografica disponibile al momento della programmazione dello studio;
- ipotesi clinica: valore atteso del vantaggio *vero* del nuovo trattamento ossia, ad esempio, riduzione della mortalità che lo sperimentatore si attende somministrando il nuovo trattamento alla *popolazione* dei pazienti. Questa quantità può esprimersi in modo assoluto o relativo, rispetto alla mortalità del gruppo esposto al trattamento di controllo (differenza assoluta o differenza relativa trattate nella prima di queste note statistiche)[2];
- livello di significatività del test statistico (si veda la sesta nota statistica[1]);
- potenza del test statistico (si veda la quarta nota statistica[3]).

Si sottolinea che il punto cruciale è rappresentato dalla specificazione dell'ipotesi clinica: un errore a questo livello, in senso sia ottimistico sia pessimistico, può vanificare lo studio.

L'importanza del calcolo della dimensione della sperimentazione in funzione dell'ipotesi clinica (nel linguaggio statistico indicata come ipotesi alternativa) sta nel fatto che, al termine dello studio, ci si attende (potenza) che il test risulti statisticamente "significativo" (rifiuto dell'ipotesi nulla) se il *vero* vantaggio del nuovo trattamento è pari o maggiore a quanto specificato dall'ipotesi clinica; in tal caso statisticamente "significativo" equivale a "clinicamente rilevante".

Nella sezione "Patients and methods" della SCCR del Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI)[4], che aveva l'obiettivo di valutare l'effetto della streptochinasi (SK) in pazienti con infarto acuto del miocardio verso un gruppo di controllo (c) si legge: "… The sample size of about 12 000 patients to be randomised was estimated according to the following criteria: initial estimate of baseline miocardial infarction mortality 12%, expected reduction in overall mortality as a consequence of SK treatment 20% *[differenza relativa]*, significance level 1% and power 95% …".

Dalla tabella IV a pagina 399[4] si apprende che la mortalità per tutte le cause nel gruppo trattato con SK è pari a m_{sk} = 10.7% e in quello di controllo è pari a m_c = 13.0%; ne consegue che la stima della riduzione relativa (d_R) è:

$$d_R = \frac{m_{sk} - m_c}{m_c} = \frac{10.7\% - 13.0\%}{13.0\%} = -17.7\%$$

e gli autori concludono: "… SK treatment produced a statistically highly significant 18% decrease in the overall mortality".

Dalla prima nota di questa serie[2] si ricava che l'intervallo di confidenza (IC) al 95% di d_R, *vera* differenza relativa nella popolazione obiettivo è: -28.0%, -10%. È degno di nota il fatto che questo IC includa il valore di riduzione (pari al 20%) atteso in fase di programmazione della SCCR. Ciò significa che si sono realizzate le aspettative dei clinici riguardo all'entità dell'effetto della SK (coerenza interna della SCCR) e quindi si può affermare che nella SCCR GISSI "significatività" statistica e "rilevanza" clinica sono due facce della stessa medaglia.

Il valore 0.1295 ≅ 13% di m_c, stima della vera mortalità per tutte le cause nella popolazione di controllo (μ_c), fornito dalla SCCR GISSI è simile al valore del 12% considerato in fase di programmazione della SCCR GISSI. Ciò permette di asserire che la SCCR GISSI soddisfa anche il criterio di coerenza esterna.

Il risultato della SCCR GISSI permette anche di sottolineare come la quantificazione della rilevanza clinica di un risultato statistico si debba basare, oltre che sull'osservata differenza della misura di effetto considerata (mortalità per tutte le cause pari al 2.3%: 13.0% - 10.7%), anche sulla prevalenza della malattia studiata nella popolazione di riferimento e sulla gravità della stessa; infatti una differenza del 2.3% potrebbe sembrare esigua e quindi clinicamente non rilevante, ma, considerando che si stima che in Italia nell'anno 2000 si siano verificati 78 808 nuovi eventi coronarici (51 874 negli uomini e 26 934 nelle donne di età compresa tra 25 e 84 anni)[4] una riduzione della mortalità del 2.3% equivale ad evitare 1812 morti per infarto miocardico.

I risultati della SCCR GISSI[5] ci permettono di effettuare un esercizio teorico particolarmente istruttivo al fine di comprendere il ruolo della potenza di uno studio. La Figura 1 riporta, in termini di differenza relativa, sei IC al 95% indicati con le lettere da (a) a (f), di cui (a) corrisponde all'IC presentato dagli autori della SCCR GISSI[4]. L'IC indicato con (b) corrisponde a quello che si sarebbe calcolato se nello studio fossero

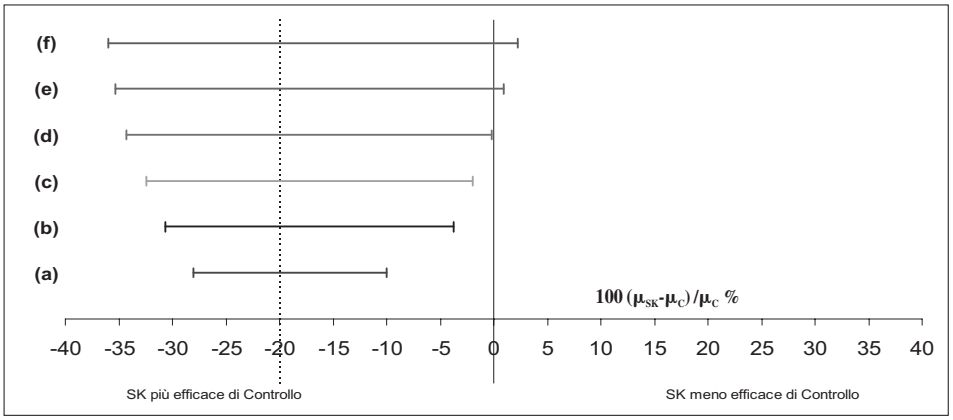

Figura 1. *Ampiezza di 6 intervalli di confidenza al 95% della vera differenza relativa tra la terapia con streptochinasi (μSK) e il controllo (μc) in accordo con i risultati della sperimentazione GISSI. L'intervallo di confidenza (a) è calcolato con la numerosità campionaria di 5250 per una significatività statistica dell'1% e una potenza del 95%. Per il calcolo dei restanti 5 intervalli di confidenza al 95% si veda il testo.*

stati arruolati 5250 pazienti come risultanti da una programmazione basata su un livello di significatività pari al 5% (test a due code) e una potenza dello studio dell'80% (usualmente impiegati nelle sperimentazioni cliniche controllate di superiorità). I rimanenti 4 sono stati calcolati con numerosità campionarie risultanti da pianificazioni basate su un livello di significatività pari al 5% (test a due code) e su valori di potenza decrescenti: (c) potenza 70%; (d) potenza 60%; (e) potenza 55%; (f) potenza 50%. Si può osservare come tutti gli IC includano la differenza relativa del 20%, considerata come obiettivo clinicamente rilevante dello studio e, inoltre, come l'ampiezza dell'intervallo aumenti con il diminuire della potenza.

Gli IC (e) ed (f) sono così ampi che il limite superiore degli stessi supera lo 0, valore corrispondente alla equiattività dei due trattamenti. In questi due casi lo studio risulterebbe non conclusivo, in quanto i valori tra 0 e il limite superiore porterebbero a concludere a favore del trattamento con controllo, mentre i valori dallo 0 al limite inferiore dell'IC porterebbero a concludere a favore del trattamento con SK. Peraltro, la constatazione che l'IC si posiziona in modo marcatamente asimmetrico rispetto allo 0 (la maggior parte giace nella regione di preferenza per SK) rende sostenibile l'ipotesi che lo studio sia risultato non conclusivo a seguito di un errore nella sua pianificazione consistente in un sottodimensionamento dello stesso (Figura 1).

Purtroppo non tutti gli articoli che riferiscono i risultati di SCCR forniscono le quantità specificate nei quattro punti citati in precedenza. Così, ad esempio, gli autori della SCCR GUSTO III[6] che aveva l'obiettivo di confrontare l'effetto di reteplase con alteplase nel trattamento dell'infarto acuto del miocardio, nella sezione "Statistical analysis" asseriscono: "… the study design required the enrolment of 15 000 patients in order to have at least 85% power to detect a 20% relative reduction in mortality with reteplase as compared with alteplase …". Nessuna informazione è data al lettore circa la stima iniziale di mortalità del gruppo di controllo trattato con alteplase e circa il livello di significatività del test che si intende eseguire; ciò impedisce da un lato la verifica della correttezza del calcolo della dimensione della SCCR e dall'altro la verifica della coerenza esterna dei risultati conseguiti. Inoltre, l'attesa riduzione relativa del 20% sembra essere soltanto giustificata dall'osservazione che: "… Reteplase, a mutant of alteplase tissue plasminogen activator, has a longer half-life than its parent molecule and produced superior angiographic results in pilot studies of acute myocardial infarction …". Alla luce dei risultati, l'attesa riduzione del 20% si è rivelata più che ottimistica. Infatti, come emerge dalla sezione "Results": "… The mortality rate at 30 days was 7.47% in the reteplase group and 7.24% in the alteplase group (odds ratio 1.03; 95% confidence interval: 0.91 to 1.8; p = 0.61)… ". Il non rifiuto dell'ipotesi nulla rende pertanto questo studio non conclusivo sul piano clinico (si vedano, a tale proposito, le conclusioni della sesta di queste note[1]).

Riassunto

Lo scopo di questa nota statistica è quello di far rivolgere l'attenzione dei cardiologi alla rilevanza clinca del risultato di una sperimentazione clinica controllata e randomizzata. La differenza tra la "rilevanza clinica" e la "significatività statistica" è illustrata mediante i risultati della sperimentazione clinica controllata GISSI e della sperimentazione clinica controllata GUSTO III.

Parole chiave: Calcolo della numerosità campionaria; Rilevanza clinica; Significatività statistica.

Bibliografia

1. Marubini E, Gallo F, Pizzamiglio S, Verderio P. Note statistiche. Cosa significa "p" a conclusione di un test d'ipotesi di una sperimentazione clinica controllata di superiorità? G Ital Cardiol 2006; 7: 684-6.
2. Marubini E, Reina G. Note statistiche. Misure di effetto assolute e relative. Ital Heart J Suppl 2004; 5: 466-71.
3. Marubini E, Rebora P, Reina G. Note statistiche. Sperimen-

tazioni cliniche controllate randomizzate sia di superiorità che di non inferiorità: considerazioni critiche. Ital Heart J Suppl 2005; 6: 361-4.
4. www.cuore.iss.it/malattie/incidenza.asp
5. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Lancet 1986; 1: 397-402.
6. The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO III) Investigators. A comparison of reteplase with alteplase for acute myocardial infarction. N Engl J Med 1997; 337: 1118-23.