Note statistiche. Cosa significa *"p"* a conclusione di un test d'ipotesi in una sperimentazione clinica controllata di superiorità?

Ettore Marubini[1], Fabio Gallo[2], Sara Pizzamiglio[2], Paolo Verderio[2]

[1]Istituto di Statistica Medica e Biometria, Università degli Studi, [2]U.O. di Statistica Medica e Biometria, Istituto Nazionale dei Tumori, Milano

Key words:
Probability level; Significance level; Test of hypothesis; Test of significance.

The aim of this statistical note, the sixth in the series, is to introduce the rationale of the test of hypothesis suitable for comparing the effect of two treatments in a randomized controlled clinical trial of superiority.

The presentation takes advantage of the analogy with a criminal trial debate based upon circumstantial evidence in an Italian Court. The results of three randomized controlled clinical trials: ISIS-1, AIMS and RESTORE are introduced and proper ways for their interpretation are suggested.

(G Ital Cardiol 2006; 7 (10): 684-686)

Ricevuto il 13 aprile 2006; accettato il 23 maggio 2006.

Per la corrispondenza:

Prof. Ettore Marubini

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

Al lettore cardiologo è noto che le conclusioni di una sperimentazione clinica controllata randomizzata (SCCR) di superiorità sono supportate fondamentalmente dai risultati di test d'ipotesi presentati in termini di *"p"*.

Ad esempio, la SCCR ISIS-1[1] che si proponeva di valutare l'effetto dell'atenololo rispetto ad un trattamento di controllo nell'infarto acuto del miocardio, riferisce così i suoi risultati: "… During the treatment period (day 0-7) there were 313 (3.89%) vascular deaths in the atenolol group compared with 365 (4.57%) in the control group. This 15% lower vascular mortality [*differenza relativa*] in the atenolol group is conventionally significant ($p < 0.04$) … ".

Un ulteriore esempio: nel paragrafo Risultati della SCCR AIMS[2], che si proponeva di valutare l'effetto dell'anistreplase contro placebo nell'infarto acuto del miocardio, si legge: "… 113 (17.8%) patients given placebo and 69 (11.1%) given anistreplase had died at one year, an odds reduction [*relativa*] in mortality of 42.7% ($p = 0.0007$) …". Si ricorda che le misure di effetto: differenza relativa e riduzione relativa degli odds ratio sono state presentate nella prima[3] di queste note statistiche.

Vediamo ora di chiarire l'iter logico che è alla base del test d'ipotesi. Ci avvarremo dell'analogia con il dibattimento in un processo *indiziario* svolto in un tribunale italiano.

In prima istanza il collegio giudicante presuppone l'innocenza dell'imputato; si richiede al Pubblico Ministero di fornire prove indiziarie atte a sostenere l'eventuale accusa di colpevolezza. D'altro canto la difesa controargomenta con l'obiettivo di sminuire o addirittura vanificare le prove del Pubblico Ministero. Spetta quindi al collegio giudicante mettere in un appropriato rapporto le argomentazioni contrastanti di accusa e difesa per giungere finalmente al verdetto.

Nel contesto di un test d'ipotesi per il confronto dell'effetto di due trattamenti (sperimentale vs controllo) in una SCCR di superiorità, l'ipotesi nulla, H_0, di equiattività dei due trattamenti coincide con la presupposta innocenza dell'imputato.

Formalmente la quantità di interesse è la differenza:

$$\mu_s - \mu_c = \delta$$

dove μ_s e μ_c sono misure dell'effetto dei due trattamenti (rispettivamente sperimentale e di controllo) sulle *popolazioni* di pazienti presenti e futuri che hanno la malattia e per i quali si ritiene appropriato il ricorso all'uno o all'altro dei due trattamenti (in queste note statistiche μ è stato usato per indicare la mortalità).

L'ipotesi nulla è ora:

$$H_0: \delta = 0$$

Trattandosi di mortalità, l'ipotesi clinica alternativa (H_a) di maggiore efficacia del

trattamento sperimentale è rappresentata da una riduzione della mortalità stessa e quindi:

$$H_a: \delta < 0$$

e coincide con l'eventuale ipotesi di colpevolezza dell'imputato.

Il ricercatore ha qui il ruolo che nel dibattimento ha il collegio giudicante; egli si basa sui risultati della sperimentazione per confutare ed eventualmente rifiutare H_0.

Di fatto, la sperimentazione non consente di conoscere μ_s e μ_c, ma soltanto le stime di queste quantità, rispettivamente m_s ed m_c, fornite dai campioni di pazienti reclutati; come tali le stime sono suscettibili di errore dovuto alla variabilità casuale propria del processo di campionamento dei pazienti. Esse possono considerarsi pertanto solo come *indizi* dell'effetto dei due trattamenti, di cui μ_s e μ_c sono le *vere* ed *ignote* misure.

La sperimentazione fornisce anche un'appropriata misura dell'errore dovuto alla variabilità casuale; nell'analogia qui adottata l'informazione fornita da tale misura è assimilabile alle controargomentazioni proposte dalla difesa.

È intuitivo che la riduzione osservata $d = m_s - m_c$ tenda ad essere tanto maggiore quanto più gli effetti dei due trattamenti sono differenti, suggerendo il rifiuto dell'ipotesi nulla a favore dell'accettazione dell'ipotesi clinica alternativa. Tuttavia qualunque sia il valore di d, esso non può essere valutato in senso assoluto, ma solo in rapporto alla dimensione del suo errore casuale. La valutazione formale di tale rapporto è eseguita ricorrendo al test d'ipotesi. Esso permette di calcolare la probabilità "p" che si verifichi, per puro effetto del caso, una riduzione di mortalità pari o superiore a quella osservata, condizionatamente all'assunto che l'ipotesi nulla sia vera, ovvero che i due trattamenti siano equiattivi. Quanto più il valore di tale probabilità è basso, tanto più si è portati a dubitare della veridicità di H_0, sino a giungere al rifiuto di H_0 in favore dell'ipotesi clinica alternativa. Tale conclusione non è formulabile in presenza di un alto valore di "p".

I due aggettivi "alto" e "basso" usati a proposito del valore della summenzionata probabilità veicolano una sorta di indeterminatezza che può lasciare insoddisfatto il lettore.

Nella letteratura medica è quindi invalsa la prassi di usare due valori di probabilità, o più propriamente due livelli di significatività, quali soglie per costruire una regola di decisione relativa al rifiuto di H_0. Precisamente, il risultato del test è giudicato:

"altamente significativo" quando "p" è inferiore,
o uguale a 0.01 ($p \leq 0.01$) e
"significativo" quando "p" è compreso tra
0.01 e 0.05 ($0.01 < p \leq 0.05$)

ambedue i casi portano al rifiuto di H_0. Il risultato del test d'ipotesi è giudicato statisticamente "non significativo" quando "p" è maggiore di 0.05 ($p > 0.05$); ciò porta a concludere che "l'esperimento non ha fornito pro e sufficienti per rifiutare H_0".

È opinione degli statistici medici che anziché riferire le conclusioni in termini di "non significativo" o "significativo" e "altamente significativo" sia preferibile riportare nelle pubblicazioni l'esatto livello di significatività calcolato. In tal modo si fa un'affermazione più precisa circa i risultati ottenuti e, ricordando che la scelta dei livelli di significatività è comunque arbitraria, si consente al lettore di valutare i risultati proposti in modo criticamente più esauriente.

Con riferimento agli esempi precedentemente citati appare evidente che la SCCR ISIS-1 nella presentazione dei propri risultati segue la prima delle modalità qui illustrate, mentre la SCCR AIMS si avvale della seconda riportando l'esatto valore di "p" calcolato (p = 0.0007). Comunque, ambedue le SCCR portano al rifiuto dell'ipotesi nulla e quindi all'accettazione dell'ipotesi clinica alternativa (H_a) di maggiore efficacia dei trattamenti sperimentali rispetto a quelli di controllo.

Considerando come terzo esempio i risultati forniti dalla SCCR RESTORE[4], nell'abstract si legge: "... The end points of the study were death from any cause, myocardial infarction … and insertion of the stent due to actual or threatened abrupt closure of the dilated artery, and the primary end point was a composite representing the occurrence of any of these events. The prespecified primary hypothesis of the study was that tirofiban, ... would result in a reduction in the 30-day composite end point compared with placebo. Patients (n = 2139) who were already receiving treatment with aspirin and heparin were randomized to receive tirofiban or placebo. The primary composite end point at 30 days was reduced from 12.2% in the placebo group to 10.3% in the tirofiban group, a 16% relative reduction (p = 0.16)". Poiché "p" è relativamente alta si può asserire che RESTORE non fornisce evidenze sufficienti per il rifiuto dell'ipotesi di equiattività dei due trattamenti posti a confronto.

Concludendo l'analogia con il dibattimento processuale, l'affermazione del ricercatore che il trattamento sperimentale è più attivo del controllo è assimilabile al giudizio di colpevolezza dell'imputato. Così come tale giudizio potrebbe rivelarsi infondato, nel contesto della sperimentazione nulla ci assicura che la conclusione sia corretta; solo analoghi risultati di studi successivi potranno provarlo. Parimenti il non rifiuto dell'ipotesi nulla da parte del ricercatore è assimilabile ad un giudizio di assoluzione "per insufficienza di prove". In altri termini ciò non permette di affermare che il trattamento sperimentale abbia la stessa efficacia del trattamento di controllo, ma semplicemente che l'esperimento non ha avuto la capacità di porre in evidenza la differenza di efficacia ipotizzata dal clinico.

Riassunto

In questa nota si introduce il procedimento logico alla base di un test di significatività atto a confrontare l'effetto di due trat-

tamenti in una sperimentazione clinica controllata randomizzata (SCCR) di superiorità. La presentazione si avvale dell'analogia con il dibattimento in un processo indiziario presso un tribunale italiano. Si riportano i risultati di tre SCCR: ISIS-1, AIMS e RESTORE e si propongono pertinenti modalità di interpretazione.

Parole chiave: Livello di probabilità; Livello di significatività; Test d'ipotesi; Test di significatività.

Glossario

• Test d'ipotesi (test di significatività): procedura formale utilizzata al fine di rifiutare l'ipotesi nulla di equiattività di due trattamenti.

Bibliografia

1. ISIS-1 (First International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous atenolol among 16 027 cases of suspected acute myocardial infarction: ISIS-1. Lancet 1986; 2: 57-66.
2. AIMS Trial Study Group. Long-term effects of intravenous anistreplase in acute myocardial infarction: final report of the AIMS study. Lancet 1990; 335: 427-31.
3. Marubini E, Reina G. Note statistiche. Misure di effetto assolute e relative. Ital Heart J Suppl 2004; 5: 466-71.
4. The RESTORE Investigators. Randomized Efficacy Study Tirofiban for Outcomes and Restenosis. Effects of platelet glycoprotein IIb/IIIa blockade with tirofiban on adverse cardiac events in patients with unstable angina or acute myocardial infarction undergoing coronary angioplasty. Circulation 1997; 96: 1445-53.