È clinicamente rilevante ritardare il primo evento scheletrico?

Enzo Ballatori
Statistico medico, Spinetoli (AP)

Fausto Roila
SC di Oncologia Medica Azienda Ospedaliera "S. Maria", Terni

RIASSUNTO

Nella valutazione degli studi clinici che hanno portato all'approvazione di denosumab come trattamento contro gli eventi scheletrici prodotti da metastasi ossee, ci si può accorgere che l'endpoint primario non appare quello più importante e che gli strumenti statistici usati sono complessi e inadeguati, e potrebbero nascondere la vera efficacia del farmaco.

Parole chiave. *Studi di non inferiorità, modello di Cox, log-rank test.*

SUMMARY

Is clinically relevant to delay the first scheletal event?

In assessing clinical studies produced for the approval of denosumab as treatment against scheletal events induced by bone metastases, not only the primary endpoint does not appear the most relevant, but also the used statistical methods are complex and inadequate, and may hide the true efficacy of the drug.

Key words. *non inferiority studies, Cox models, log-rank test.*

Negli ultimi 4 anni sono stati pubblicati tre studi fotocopia sull'efficacia differenziale di denosumab (DEN) rispetto all'acido zoledronico (ZOL) nel ritardare il primo evento scheletrico correlato alla presenza di metastasi ossee. Quello presentato nella scheda riguarda pazienti con carcinoma prostatico resistente alla castrazione; gli altri riguardano pazienti con carcinoma della mammella metastatico[1] e pazienti con carcinoma metastatico (esclusi il cancro della prostata e quello della mammella) o con mieloma multiplo[2]. Tali studi, tutti di ampia dimensione campionaria, sono stati accompagnati dalla pubblicazione di almeno altri tre lavori condotti considerando insieme i pazienti arruolati nei tre studi originali e/o valutando l'effetto differenziale di DEN rispetto a ZOL su altre caratteristiche non considerate negli studi originali.

I tre studi adottano la stessa metodologia e pervengono a risultati similari. Così, la discussione è incentrata solo su quello riportato nella scheda, ma quanto sarà esposto si estende anche agli altri due.

Le principali ragioni per discutere di tali studi (peraltro non recentissimi) in una rivista di terapia di supporto sono dovute ai risultati conseguiti ed alla natura degli endpoint considerati. Infatti, nei suddetti studi sono state condotte analisi *esplorative* sulla sopravvivenza globale e sul tempo alla progressione e, in tutti e tre i casi, le curve di sopravvivenza (e del tempo alla progressione, ma questo è secondario) relative ai due bracci di trattamento sono sovrapponibili. Sebbene non sia chiara la ragione per cui tali analisi siano state considerate "esplorative", la uguale sopravvivenza comporta automaticamente che il vantaggio di denosumab in termini di efficacia, ossia di effetto sul paziente, riguarda la sfera della qualità di vita, ma in tal caso non è chiaro il motivo per cui questo aspetto non sia stato direttamente valutato mediante appositi questionari psicometrici.

Endpoint

L'evento scheletrico correlato alla malattia può manifestarsi in vari modi, per cui avendo deciso che l'endpoint principale è il tempo al primo evento scheletrico correlato, la scelta di un endpoint composto sembra ragionevole (v. scheda). Al riguardo, però, resterebbe da chiarire perché siano state escluse fratture incorse in seguito a traumi severi, quando è noto che la presenza di metastasi ossee può determinare punti di minore resistenza, per cui anche in seguito a traumi severi è possibile che si sia verificata una frattura che non si sarebbe prodotta se in quel punto non fosse stata presente una metastasi. Per non parlare poi della definizione di "trauma severo" che, non essendo stata data, è lasciato all'arbitrio del ricercatore decidere quando un trauma è da considerarsi severo, con conseguente introduzione di elementi di soggettività.

Comunque, il punto più controverso è proprio la scelta del "tempo al primo evento scheletrico correlato" come endpoint principale. Infatti, ci si sarebbe atteso che l'endpoint principale fosse stato l'incidenza degli eventi scheletrici e, nel gruppo dei pazienti che hanno avuto almeno un evento scheletrico, una misura della ripetitività del fenomeno avrebbe fornito un rilevante endpoint secondario.

Disegno dello studio

Lo studio sembra programmato come studio di non inferiorità, ma, una volta provata la non inferiorità di DEN rispetto a ZOL (in relazione all'endpoint considerato), si passa a testare la superiorità del farmaco in studio.

Cosa pensiamo degli studi di non inferiorità è stato ampiamente esposto nel n. 3 di CASCO: salvo rari e motivati casi, gli studi di non inferiorità non dovrebbero essere accettati dalla comunità scientifica. Qui siamo di fronte ad uno studio che è al contempo di non inferiorità e di superiorità condi-

zionata e ci chiediamo fino a che punto sia corretto un tale modo di procedere, anche perché nell'articolo non sono state date informazioni dettagliate su tale scelta.

Per utilità del Lettore, riportiamo una sintesi brevissima (e quindi necessariamente approssimativa) della storia dell'uso della statistica nelle applicazioni.

L'uso dei test statistici nella ricerca scientifica si può far risalire alle scuole di Biometria (seconda metà del XIX secolo); l'ipotesi nulla (negli esperimenti comparativi, quella di uguale efficacia dei trattamenti) era sottoposta ad un test statistico per valutare se la differenza riscontrata tra i due gruppi poteva essere attribuita al caso o, invece, era così rilevante da indurre a propendere per la diversa efficacia dei trattamenti. In tal caso, il segno della differenza era dirimente per individuare il trattamento più efficace (non c'era bisogno di ipotesi alternativa).

Gradualmente, però, nuovi campi di applicazione si aprono per la Statistica, ad esempio nell'industria, dove necessitano strumenti non tanto orientati ad investigare la natura, quanto finalizzati a prendere decisioni. Poiché nell'industria ogni decisione sbagliata ha un costo, si pone l'attenzione non solo alla probabilità di commettere un errore respingendo l'ipotesi nulla quando invece è vera (livello di significatività), ma anche alla probabilità di commettere un errore accettando l'ipotesi nulla quando invece i trattamenti sono diversamente efficaci. Il complemento a 1 di quest'ultima probabilità si chiama potenza del test che, a parità delle altre assunzioni (cioè: differenza minima rilevante, test statistico prescelto, livello di significatività), è funzione matematica della dimensione campionaria, nel senso che, fissata la potenza, è univocamente determinata la numerosità e, viceversa, fissata la numerosità, è univocamente determinata la potenza. Il Lettore avrà certamente notato, nelle applicazioni, l'asimmetria nel controllo dei due tipi di errore: per convenzioni internazionali ormai consolidate, il livello di significatività si pone (quasi) sempre al 5%, mentre la probabilità dell'errore che si commette accettando l'ipotesi nulla quando i trattamenti sono diversamente efficaci è (spesso) del 20% (= 1 – 0,8; potenza = 80%). In quest'ultimo caso, se si accetta l'ipotesi nulla si può avere comunque un 20% di probabilità di sbagliare e tale probabilità non può essere considerata bassa. Però, se volessimo aumentare la potenza, la dimensione campionaria crescerebbe rapidamente e se la portassimo al 95% molti studi sarebbero infattibili a causa dell'enorme numero di unità da osservare. Successivamente, dall'industria farmaceutica, furono introdotti studi di non inferiorità e di equivalenza, unicamente per rispondere ad istanze del marketing.

Nella ricerca clinica, andrebbero distinti due tipi di studi: quelli con finalità regolatorie e quelli volti ad approfondire la relazione tra fenomeni (ad es., studi spontanei). Solo per i primi la Statistica come strumento decisionale appare giustificata: si tratta di decidere se autorizzare o meno il farmaco per la rimborsabilità e il protocollo garantisce che le regole andranno rispettate. I secondi dovrebbero essere liberati da ogni struttura che li ingessi, come ad esempio prefissare (nel protocollo) gli strumenti statistici che saranno impiegati, in quanto l'analisi dei dati consiste proprio in una continua interazione tra ricercatore e risultati delle elaborazioni che via via vengono eseguite.

Analisi dei risultati

Essendo l'endpoint principale una variabile di tipo "time to failure", gli autori hanno pensato bene di valutarla ricorrendo al modello di Cox, aggiustando i risultati relativi all'endpoint principale per tre fattori: precedenti eventi scheletrici (sì, no), PSA (inferiore o meno a 10ng/ml), trattamento chemioterapico nelle 6 settimane precedenti l'arruolamento (sì, no). I fattori con cui è stato aggiustato l'Hazard Ratio (HR, ricavato in base al modello di Cox) tra i due trattamenti sono gli stessi per cui è stata stratificata la randomizzazione. Qui si apre lo scenario per una discussione articolata.

1. La randomizzazione stratificata ha un'unica finalità: garantire un perfetto bilanciamento tra i due bracci sperimentali dei fattori di stratificazione (quelli ritenuti importantissimi nella valutazione della risposta al trattamento), in modo tale che la risposta (sintetizzata da una media) non risenta più del loro effetto. Ora, se i bracci sono bilanciati, ci si chiede quale sia la necessità di aggiustare per tali fattori. A nostro avviso sarebbe stato sufficiente un semplice log-rank test che, essendo un test non parametrico, contrariamente al modello semiparametrico di Cox, non ha bisogno di alcuna assunzione per essere valido. Nel lavoro sono riportati i tempi mediani al primo evento scheletrico correlato (20,7 mesi con DEN vs 17,1 con ZOL); quindi il ritardo mediano della comparsa del primo evento scheletrico correlato è di circa 3,6 mesi con l'uso di DEN rispetto a ZOL; in base alle analisi condotte, tale differenza risulta significativa. Però, gli intervalli di confidenza di tali stime (v. scheda) presentano regioni di sovrapposizione, il che lascerebbe pensare che il log-rank test non sarebbe risultato significativo. Si noti che, negli altri due lavori, gli intervalli di confidenza per il tempo mediano al primo evento scheletrico correlato non sono stati riportati.
2. L'assunzione alla base del modello di Cox è quella del rischio proporzionale (*proportional hazard*), per cui in tutti i sottogruppi definiti dalle combinazioni di modalità dei fattori di aggiustamento (nel nostro caso 2 x 2 x 2 = 8), il rapporto tra i rischi (HR) relativo ai due trattamenti deve rimanere costante. Esistono metodi statistici per la verifica del proportional hazard, ma nel lavoro in discussione non si fa cenno del loro impiego. Se il proportional hazard non è rispettato, il modello di Cox si riduce ad un puro esercizio matematico: formalmente lo si può sempre costruire (qualunque software statistico è in grado di farlo), ma i risultati che fornisce sono inattendibili. Le cose peggiorano quando si passa alle analisi "esplorative": sopravvivenza globale e tempo alla progressione valutato dal medico sperimentatore. Qui l'aggiustamento è stato fatto, oltre che per i tre fattori di stratificazione, anche per altre 6 variabili valutate al basale (peraltro non è precisato

SCHEDA

Fizozi K, Carducci M, Smith M, et al. Denosumab versus zoledronic acid for treatment of bone metastases in men with castration-resistant prostate cancer: a randomised, double blind study. Lancet 2011; 377: 813-23.

Nel periodo Maggio 2006 - Ottobre 2009, 2516 pazienti furono screenati in 342 centri di 39 paesi; ne furono arruolati 1904 e valutati 1901 che vennero randomizzati a ricevere acido zoledronico (951, ZOL) e denosumab (950, DEN). La randomizzazione fu stratificata per precedenti eventi scheletrici (sì, no), PSA (< 10 vs ≥ 10) e trattamento chemioterapico nelle 6 settimane prima della randomizzazione (sì, no).
Endpoint primario: tempo al primo evento scheletrico (correlato alla malattia) valutato per non inferiorità. Se si dimostrasse non inferiore, allora lo stesso endpoint sarebbe successivamente valutato per superiorità come **endpoint secondario** insieme ai tempi degli eventi scheletrici successivi al primo. Un'analisi esplorativa fu programmata per la valutazione della sopravvivenza globale e per il tempo alla progressione valutato dal medico sperimentatore.
Un evento scheletrico correlato fu definito come

a. frattura patologica (escluse quelle incorse in seguito a traumi severi)
b. radioterapia dell'osso (incluso l'uso di radioisotopi)
c. chirurgia dell'osso
d. compressione della corda spinale.

Un esame dello scheletro fu eseguito ogni 12 settimane ed incluse le radiografie del cranio, colonna vertebrale, torace, pelvi, braccia e gambe. Ogni radiografia fu valutata, indipendentemente ed in cieco, da due lettori ed un terzo valutò le immagini dove si erano riscontrate discordanze di risposta.

Analisi statistica. L'analisi statistica degli endpoint primario e secondario fu condotta, in accordo al principio di intenzione a trattare, mediante l'Hazard Ratio (HR) di denosumab vs acido zoledronico, stimato mediante il modello di Cox, aggiustando per i fattori usati nella stratificazione della randomizzazione (precedenti eventi scheletrici, PSA, chemioterapia nelle precedenti 6 settimane). Nessuna analisi ad interim fu pianificata.
Le analisi esplorative sulla sopravvivenza globale e sul tempo alla progressione di malattia furono condotte mediante il modello di Cox, aggiustando non solo per le variabili di stratificazione, ma anche per le caratteristiche basali del paziente (età, tempo trascorso dalla prima diagnosi di carcinoma prostatico alla malattia metastatica, tempo dalla diagnosi alle metastasi ossee, presenza di metastasi nelle viscere, score di Gleason, Ecog performance status).

Risultati. La quasi totalità dei pazienti del gruppo DEN fu esposta al denosumab per una mediana di 11,9 mesi, mentre la quasi totalità di quelli del gruppo ZOL fu esposta all'acido zoledronico per una mediana di 10,2 mesi.
Denosumab mostrò una maggiore efficacia dell'acido zoledronico nel ritardare la comparsa del 1° evento scheletrico correlato alla malattia; tempo mediano alla comparsa del 1° evento (95% CI)

– con denosumab: 20,7 mesi (18,8-24,9)
– con acido zoledronico: 17,1 mesi (15,0-19,4),

con una differenza tra le mediane di 3,6 mesi ed un Hazard Ratio HR = 0,82 (CI 0,71-0,95), risultati significativi allo 0,001 per non inferiorità e allo 0,008 per superiorità.
Il numero totale degli eventi confermati fu di 386 (41%) nel gruppo ZOL e di 341 (36%) nel gruppo DEN. Scendendo nel dettaglio, si osservarono i seguenti eventi scheletrici, rispettivamente nel braccio ZOL e nel braccio DEN

– Radioterapia all'osso: 21% e 19%
– Frattura patologica: 15% e 14%
– Compressione spinale: 4% e 3%
– Chirurgia dell'osso: < 1% e < 1%.

Sia la sopravvivenza globale che il tempo alla progressione (analisi esplorative) risultarono sovrapponibili tra i due gruppi.
L'incidenza di eventi avversi risultò paragonabile tra i due bracci.

Ruolo della fonte di finanziamento.
L'autore corrispondente ha collaborato con lo sponsor nel disegno dello studio. La raccolta e l'analisi dei dati fu eseguita dallo sponsor. Tutti gli autori hanno partecipato alla stesura dell'articolo con l'assistenza di un *medical writer* fornito dallo sponsor. L'autore corrispondente fu responsabile della decisione di sottoporre il lavoro per la pubblicazione. •

se tali variabili siano state considerate in quanto tali o, invece, trasformate in fattori, v. scheda), per un totale di 9 tra fattori e variabili considerati. Non solo in tale situazione il proportional hazard diventa arduo persino da ipotizzare, ma sarebbe stato sufficiente dare un'occhiata alle corrispondenti curve relative ai due trattamenti per vedere una quasi perfetta sovrapposizione che avrebbe esonerato da qualsiasi altra analisi.

Conclusioni

Nella programmazione degli studi clinici e nell'analisi dei risultati, talvolta la Statistica viene utilizzata in maniera esagerata, quasi per nascondere, con la complessità degli strumenti adottati, la povertà dei risultati conseguiti. Il lavoro sintetizzato nella scheda e gli altri due sullo stesso argomento ci sembrano di questo tipo.

C'è una lunga catena che va dai medici ricercatori che accettano di firmare i protocolli proposti dall'industria, ai comitati etici che approvano i protocolli, alle riviste scientifiche che pubblicano i risultati, alle autorità regolatorie che approvano i farmaci, alle Consensus Conference che introducono i nuovi farmaci nelle linee guida, al medico che li prescrive; tutti i punti nodali di tale catena presentano problemi di varia natura. Dati gli scopi di questa rubrica, ci soffermeremo brevemente solo sul sistema di referaggio delle riviste.

Essendo i tre lavori considerati tutti pubblicati su riviste assai importanti (*Lancet* e *JCO*), sapendo che ciascuna di loro fa valutare ogni manoscritto da tre referee, di cui uno statistico, desta meraviglia il fatto che di nove (o poco meno) referee diversi che hanno analizzato i tre studi nessuno abbia avanzato obiezioni sostanziali, o che comunque queste non siano state considerate dall'Editor nel prendere la decisione finale di pubblicare il lavoro. Il lettore non solo dovrebbe sapere chi ha approvato la pubblicazione (in particolare se è stata un'iniziativa del solo *editor in chief* o se anche tutti o in parte i referee erano d'accordo), ma soprattutto dovrebbe conoscere i contenuti delle revisioni dei referee per comprendere meglio i possibili limiti dello studio (e forse anche dei referee!), onde poterne dare una propria valutazione critica. Ma la pubblicazione di tali revisioni (magari solo on-line) non è la sola cosa che andrebbe cambiata nell'attuale politica editoriale delle riviste scientifiche, anche dato il ruolo, sempre più pesante, rivestito dallo sponsor nella ricerca clinica (v. scheda, ultime righe), per non parlare degli abbondanti conflitti di interessi degli autori (indicati nell'articolo, ma non riportati nella scheda). •

Bibliografia

1. *Stopek AT, Lipton A, Body JJ, et al. Denosumab compared with zoledronic acid for the treatment of bone metastases in patients with advanced breast cancer: a randomized, double-blind study. JCO 2010; 28: 5132-9.*
2. *Henry DH, Costa L, Goldwasser F, et al. Randomized, double-blind study of denosumab versus zoledronic acid in the treatment of bone metastases in patients with advanced cancer (excluding breast and prostate cancer) or multiple myeloma. JCO 2011; 29: 1125-32.*

Statistica per concetti

Test statistici per dati appaiati

Riassunto

I test statistici per dati appaiati sono esposti in relazione alla scala in cui si colloca la risposta al trattamento. In particolare, sono descritti il test di Mc Nemar (scala nominale), il test di Wilcoxon per dati appaiati (scala ordinale) e il t-test per il confronto tra due medie nel caso di dati appaiati (scale di rapporti).

Parole chiave. *Disegno cross-over, dati appaiati, test di Mc Nemar, ranghi, test di Wilcoxon per dati appaiati, t-test per dati appaiati.*

Summary

Statistical tests for paired data

Statistical tests for paired data are described in relationship with the scale used to evaluate the response to the treatment. More precisely, we described Mc Nemar test (nominal scale), Wilcoxon matched pairs, signed ranks test (ordinal scale), t-test for paired data (ratio scale).

Key words. *Cross-over design, paired data, Mc Nemar test, Wilcoxon matched pairs signed ranks test, t-test for paired data.*

Si generano dati appaiati (*paired data*) quando uno stesso carattere (qualitativo o quantitativo) è rilevato due (o più) volte sulle stesse unità.

Il vantaggio di operare con dati appaiati consiste nel tenere costanti tutti i fattori (prognostici, predittivi, di rischio) che possono interferire con l'effetto del trattamento sulla risposta.

Nella ricerca clinica si ha spesso a che fare con dati appaiati, quando, ad esempio, si valuta la riproducibilità di un nuovo test diagnostico, per cui gli stessi pazienti sono esaminati da due osservatori indipendenti usando lo stesso strumento, o quando un test psicometrico viene somministrato due volte agli stessi soggetti per valutarne la riproducibilità. Ma l'esempio di dati appaiati più comune è in relazione al disegno cross-over, che pertanto è assunto come applicazione di riferimento nell'esposizione che segue.

Disegno cross-over

Siano A e B i due trattamenti a confronto (ma quanto esposto è generalizzabile al caso di più di due trattamenti). Il disegno cross-over consiste nel somministrare, in tempi diversi, entrambi i trattamenti agli stessi pazienti, valutando così due risposte in ogni paziente: una dopo la somministrazione di A, l'altra dopo quella di B.