Dolore da metastasi ossee: ibandronato o radioterapia?

Enzo Ballatori
Statistico medico, Spinetoli (AP)

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

RIASSUNTO

Vengono discussi alcuni punti importanti del lavoro che impediscono di considerare ibandronato non inferiore alla radioterapia nel controllo del dolore da metastasi ossee nei pazienti con carcinoma prostatico. In particolare, vengono esaminati i problemi relativi alla scarsa potenza dello studio, alle tecniche di randomizzazione adottate, all'uso di analisi ad interim.

Parole chiave. *Studi di non inferiorità, potenza, dimensione del campione, randomizzazione, analisi ad interim.*

SUMMARY

Bone pain: ibandronate or radiotherapy?

Some points of this study are emphasized because they produce serious difficulties in considering ibandronate not inferior to radiotherapy in the control of the bone pain in prostate cancer patients. In particular, low power of the study, randomization, and the use of ad interim analyses are discussed.

Key words. *Non-inferiority studies, power, sample size, randomization, ad interim analysis.*

Lo studio sintetizzato nella scheda sembra abbastanza corretto; in particolare va elogiata la competenza dello statistico che si può evincere da molti elementi, tra cui:

a. l'uso di due criteri di valutazione (ITT e *per-protocol*) per l'analisi dei risultati di uno studio di non inferiorità;
b. aver riscontrato un'asimmetria nella valutazione della risposta in base al criterio EAS, con conseguente ricorso a test non parametrici (test U di Mann-Whitney);
c. aver riscontrato irrealistica l'ipotesi del *proportional hazard*, con conseguente abbandono del modello di Cox;
d. l'uso dell'analisi di covarianza per la valutazione della qualità di vita, eseguita per aggiustare i risultati rispetto ai valori basali.

Ma, come si sa, la Statistica non può aggiustare tutto; infatti, nel lavoro compaiono alcuni elementi su cui vale la pena di riflettere.

Anzitutto, si tratta di uno studio di non inferiorità e riteniamo che, per le finalità dello studio, tale scelta sia appropriata. Tuttavia raccomandiamo vivamente al lettore di riguardare tale argomento nella rubrica "Casi clinici", in CASCO 3, per la piena comprensione del lavoro.

1. **Intervallo di non inferiorità ± 15%**, dove il 15% si riferisce alla differenza tra le percentuali di risposte al trattamento nei due bracci. Formalmente la scelta di esprimere l'ipotesi nulla con un intervallo che va da -15% a + 15% è tipica degli studi di equivalenza; negli studi di non inferiorità l'ipotesi nulla è H_0: $C - T \geq \Delta$, e l'ipotesi alternativa è H1: $C - T < \Delta$, dove, nel nostro caso, C rappresenta la percentuale di risposte nel braccio di controllo (Radioterapia, RT), T la percentuale di risposte nel gruppo sperimentale (Ibandronato, I) e Δ la differenza al di sotto della quale possiamo ritenere il trattamento sperimentale non inferiore a quello di controllo. Quindi, essendo Δ un numero positivo, negli studi di non inferiorità, l'intervallo comprende tutti i valori inferiori a Δ (ad es., essendo fissato Δ al 15%, non solo –16%, ma anche –40% indicherebbe una non inferiorità).

Ma non è sugli aspetti formali che vogliamo richiamare l'attenzione, quanto su quelli sostanziali. Considerando che i calcoli finali per la determinazione della dimensione dello studio sono stati eseguiti ipotizzando un 40% di risposte a RT, prefissare al 15% l'estremo dell'intervallo di non inferiorità ci sembra davvero esagerato. Infatti, se RT producesse una percentuale di risposte del 40% potremmo dichiarare I non inferiore a RT anche quando ottenesse una percentuale di risposte del 25%, cioè poco più della metà del valore prefissato per RT.

La conseguenza più grave è che, con una scelta di Δ così alta, la dimensione del campione si riduce, così che diventa poco probabile dichiarare il nuovo trattamento inferiore a quello standard, anche se di fatto lo fosse.

2. **Dimensione del campione.** In sede di programmazione dello studio fu ipotizzato un 70% di risposte per RT, pervenendo ad una dimensione campionaria di 580 pazienti. Alla prima analisi ad interim tale percentuale risultò invece del 40%, il che ha indotto gli autori a ricalcolare la dimensione del campione in 470 pazienti. I risultati, invece, mostrano una percentuale di risposte del 50-60% (quindi, ben superiore al 40%), a seconda del criterio di valutazione usato (WHO, EAS). Da quest'ultimo dato emerge che il campione è sottodimensionato, e, quindi, la potenza dello studio (cioè la probabilità di dichiarare inferiore il trattamento sperimentale I, anche se realmente lo fosse) è molto inferiore a quella programmata (90%). Pur sospendendo il giudizio sulla liceità di modificare i parametri definiti in sede di programmazione di uno studio clinico sulla base di quanto emerge da un'analisi

SCHEDA

Hoskin P, Sundar S, Reczko K, et al. A multicenter randomized trial of ibandronate compared with single-dose radiotherapy for localized metastatic bone pain in prostate cancer. J Natl Cancer Inst 2015; 107: 1-9.

Si tratta di uno studio multicentrico, randomizzato, di non inferiorità, *nonblind*, volto a confrontare l'efficacia di ibandronato (I, un difosfonato di terza generazione) rispetto ad una singola dose di radioterapia (RT, trattamento standard) nel controllo del dolore da metastasi ossee in pazienti con carcinoma della prostata.

Metodi

Risposte. Endpoint primario: riduzione del dolore a 4 settimane rispetto al basale.
La valutazione basale comprendeva: un'autovalutazione del dolore alle ossa per mezzo del Wisconsin Brief Pain Inventory (BPI) e il consumo di analgesici, oltre al questionario FACIT-G per la valutazione della qualità di vita (QoL).
Il BPI incorpora un analogo visivo da 0 (nessun dolore) a 10 (peggior dolore immaginabile). Ai pazienti è stato chiesto di registrare il peggior dolore, la media e il minimo dolore avuti negli ultimi 3 giorni. Questi dati sono stati combinati con quelli del consumo degli analgesici sia con la scala predisposta dal WHO, sia con la scala EAS (Effective Analgesic Score).
La valutazione del dolore e il consumo di analgesici sono stati registrati a 4, 8, 12, 26 e 52 settimane. La QoL è stata valutata, oltre che al basale, a 4 (s4) e a 12 (s12) settimane.
Nelle analisi principali è stato usato il punteggio relativo al peggior dolore. Sono state definite:
"Risposta completa" (CR) uno score del peggior dolore pari a 0 con stabile o ridotto consumo di analgesici
"Risposta parziale" (PR) una riduzione dello score del dolore di 2 o più punti con stabile o ridotto consumo di analgesici, ovvero una riduzione del consumo di analgesici di almeno il 25% con un cambiamento dello score del dolore di più di 1 punto.
La misura dell'uso degli analgesici è stata eseguita sia con la scala WHO (che vale 1 se si consumano farmaci non oppioidi, 2 se si usano deboli oppioidi, 3 se si usano forti oppioidi, considerando la medicazione più forte ad ogni tempo di analisi), sia con l'EAS [così come descritto nell'articolo Mercadante S. Scoring the effect of radiotherapy for painful bone metastases. Support Care Cancer 2006; 14: 967-69] che considera tipo e dose di analgesia usata trasformata in morfina equivalente riportando il tutto su una scala da 0 a 150. La risposta sulla scala EAS è stata definita o come uno score di 0 al tempo di interesse, ovvero come riduzione di almeno il 20% rispetto al basale.
Endpoint secondari includevano la risposta al dolore nel lungo periodo, cioè a 26 (s26) e a 52 (s52) settimane, la valutazione della QoL a s4 e a s12, i tassi di passaggio al trattamento alternativo, complicanze ossee, tossicità e sopravvivenza globale (OS).

Randomizzazione. L'assegnazione casuale dei pazienti ai trattamenti fu stratificata per centro, con un'allocazione 1:1 usando blocchi di 4.

Dimensione del campione. Ipotizzando un tasso di risposta del 70% nel braccio di controllo alla quarta settimana, ed una massima differenza del 15%, con il 90% di potenza, la dimensione del campione fu calcolata in 580 pazienti, prevedendo 3 analisi ad interim ed un 20% di pazienti non valutabile. Tale dimensione è stata ridotta a 470 pazienti (12% di non valutabili, 90% di potenza, e un livello di significatività del 5% per un test unidirezionale) perché alla prima analisi ad interim il tasso di risposta fu molto più basso di quello ipotizzato (40% anziché 70%). Nelle analisi ad interim non fu utilizzato alcun test statistico.

Analisi statistica. Le analisi sono state condotte in base ad entrambi i criteri di intenzione a trattare (ITT) e *per-protocol*, escludendo, in quest'ultima valutazione, sia i pazienti che avessero ricevuto un trattamento diverso da quello cui erano stati allocati, sia i pazienti ineleggibili, sia quelli in cui il sito del dolore al basale non coincideva con quello dichiarato al momento in cui era stata condotta l'analisi. I pazienti con omessa risposta alla quarta (s4) ed alla 12-esima (s12) settimana furono esclusi dalla valutazione per ITT.
L'effetto del trattamento è stato diverso a seconda che fosse stato adottato il criterio WHO o l'EAS (tab. 2, disponibile online) a s4 e a s12.
Considerando il criterio EAS, le differenze tra s4 e s12 e il basale furono trovate avere una distribuzione asimmetrica, per cui è stata determinata la mediana e, per il confronto, è stato calcolato il test U di Mann-Whitney. Per aggiustare per il valore basale sono state usate tecniche di regressione sulla mediana.
Per la scelta di un test unidirezionale, gli intervalli di confidenza per la differenza delle risposte sono stati costruiti con un coefficiente del 90%.
La sopravvivenza globale (OS) è stata valutata usando le stime di Kaplan-Meyer e il modello di Cox sia per il confronto tra i gruppi di trattamento, sia per i 4 gruppi (non randomizzati) in cui fu eseguito il test-retest.
L'ipotesi del *proportional hazard* era violata in questi ultimi gruppi, e così fu riportata la sopravvivenza mediana.
L'effetto del trattamento sulla QoL a s4 e a s12 è stato valutato aggiustando per il basale mediante l'analisi della covarianza.

→

→

Risultati

Tra aprile 2003 e novembre 2009, 470 pazienti furono arruolati in 58 centri: 27 pazienti fallirono la risposta a s4 e potrebbero essere stati ri-trattati tra s4 e s8; di questi 14 (54%) ricevettero l'altro trattamento. Altri 114 pazienti furono ri-trattati a discrezione del clinico. Non ci fu differenza sostanziale nella percentuale di crossover tra coloro che erano partiti con ibandronato (I; 72, pari al 31%) e quelli che avevano iniziato con la radioterapia (RT; 56, pari al 24%). Come atteso, la percentuale di risposte a s4 fu più bassa tra coloro che erano passati all'altro trattamento: I e poi RT (33,8%) e RT e dopo I (32,6%). Parecchi *non responders* non furono trasferiti all'altro trattamento, mentre alcuni *responders* sì, denotando che la decisione di crossare non fu interamente basata sulla risposta (tab. 3, online).

Risposta al dolore

1. *Criterio* WHO. Né a s4, né a s12 ci furono differenze significative tra i due bracci per il peggior dolore. A s4 le percentuali di risposte furono: per I il 49,5% e per RT il 53,1%; differenza: - 3,6%; 90%CI: -12,4% --- 5%, P = 0,49); si può osservare che l'intervallo di confidenza cade all'interno del ± 15% programmato. A s12 le risposte furono: per I, il 56,1%, per RT, il 49,4%; differenza: 6,7%; 90%CI: -2,6% --- 16%, P = 0,24).
Risultati simili si ebbero a s26 e a s52 (tab. 4, online). Ad analoghe conclusioni si pervenne per la media del dolore e per il minimo dolore.

2. *Criterio* EAS. Considerando il peggior dolore, la risposta non fu significativamente diversa fra i due trattamenti sia a s4 (I: 52,7%, RT: 60,2; differenza: -7,5%; 90%CI: -15,7% --- 0,7%) che a s12 (I: 56,7%, RT: 60,2%; differenza: -3,5%; 90%CI: -12,3% --- 5,3%). Quest'ultimo risultato è compreso nel preventivato ± 15%, mentre il primo è un poco al di fuori. Simili conclusioni furono raggiunte per s26 e per s52.
Anche per la media ed il minimo del peggior dolore si ebbero risultati analoghi.
Il test U di Mann-Whitney fornì un'evidenza a favore di RT alla settimana 4 (P < 0,04), ma non alle settimane successive (s12, s26, s52).
Anche la media del peggiore dolore fu significativamente diversa tra i due gruppi a s4, ma non alle successive settimane.

Analisi per-protocol

Furono ripetute le analisi escludendo 25 pazienti (7 ineleggibili, 12 che non ricevettero il trattamento cui erano stati assegnati, 6 con un sito diverso del dolore).
I risultati furono simili a quelli ottenuti in base al criterio ITT (tab. 8, online).

Qualità di vita

Si ottennero risultati simili tra i due gruppi, per ogni dimensione ed in totale.

Tossicità

L'incidenza di eventi avversi fu simile tra i due gruppi, con l'eccezione di diarrea (RT: 12%, I: 6%, P < 0,014) e nausea (RT: 25%, I: 18%; P < 0,058, ai limiti della significatività) più frequenti con la RT.
Complessivamente, le altre tossicità furono più frequenti con ibandronato (I: 19%, RT: 9%; P = 0,001).

Sopravvivenza globale (OS)

Dopo un follow up mediano di 11,7 mesi, 395 pazienti (84%) erano morti (I: 200, RT: 195). La sopravvivenza mediana fu simile nei due bracci: in RT: 12,2 mesi, in I: 12,9 mesi. •

ad interim, ci si rende conto che gli autori in questo studio hanno avuto seri problemi di arruolamento (che però non sono stati dichiarati). Infatti, il periodo di reclutamento dei pazienti è stato di circa 6,5 anni; considerando che i centri partecipanti sono stati 58, essendo 470 il numero dei pazienti arruolati, vuol dire che sono stati reclutati in media 1,25 pazienti/anno in ciascun centro. Infatti, correttamente, gli autori non parlano di consecutività dell'arruolamento, ma se tale assunzione viene a cadere, sorgono seri dubbi circa la rappresentatività del campione.

3. Randomizzazione a blocchi. Per assegnare casualmente i pazienti ai trattamenti, in ciascun centro, sono stati usati blocchi di 4 pazienti, cioè, in ogni centro, la randomizzazione ripartiva dopo ogni quarto paziente arruolato.

I centri sono 58, per cui l'arruolamento medio è stato di 8,1 pazienti per centro (due blocchi, circa). Questo è un dato medio; quindi, probabilmente, ci saranno centri che hanno arruolato decine di pazienti e centri che hanno reclutato solo pochissime unità. In questi ultimi verosimilmente non sarà stato esaurito nemmeno un blocco, per cui il beneficio di un bilanciamento in ciascun centro è vanificato.

Un altro problema scaturisce dal fatto che i blocchi sono solo di 4 pazienti, per cui la proprietà di imprevedibilità della randomizzazione sembra in parte non rispettata; infatti, una volta randomizzati 3 pazienti, è certo il trattamento cui verrà assegnato il quarto (sulla randomizzazione nella ricerca clinica si consiglia di riguardare la rubrica "Statistica per concetti", in CASCO 2).

4. Analisi ad interim. Nell'articolo si dichiara di aver programmato 3 analisi ad interim e che in ciascuna di queste non furono eseguiti test statistici. Probabilmente la mancata esecuzione del test è dovuta alla successiva necessità, in caso contrario, di aggiustare il livello di significatività in accordo con la disuguaglianza di Bonferroni, il che avrebbe accresciuto ul-

teriormente il numero dei pazienti da arruolare. Ma il test statistico è uno strumento decisionale, e ciò fa sorgere un quesito: che decisione avrebbero preso gli autori se alla prima analisi ad interim fosse risultata una percentuale di risposte con RT del 40% e con I del solo il 3%. Anche senza eseguire un formale, quanto inutile, test statistico è verosimile che gli autori avrebbero preso la decisione di interrompere lo studio: la conoscenza dei risultati, in sé, può condurre ad una decisione e, quindi, la mancata esecuzione del test comunque non evita il problema dei confronti multipli (v. "Statistica per concetti", in CASCO 1).

5. Risultati e conclusioni. Nei risultati e nella discussione, gli autori sembrano voler concludere per la non inferiorità di ibandronato rispetto alla radioterapia ma purtroppo le evidenze non depongono in tal senso. Infatti, la scarsa numerosità dei pazienti arruolati, come sopra evidenziato, diminuendo la potenza del test rende poco probabile che, seppure la radioterapia fosse più efficace, gli autori se ne sarebbero accorti. Inoltre, la risposta alla quarta settimana è stata scelta come endpoint principale: nel caso che tale risposta sia definita con il criterio WHO, in effetti il risultato sarebbe interpretabile come non inferiorità, ma nel caso della scelta del criterio EAS la differenza rispetto al test U di Mann-Whitney è significativa, il che induce a ritenere la radioterapia più efficace. Infine, sembra che la tossicità sia a sfavore di I; infatti, gli eventi avversi relativi all'apparato gastro-intestinale (diarrea, nausea) sono più frequenti con la radioterapia, ma tutti gli altri, complessivamente, hanno un'incidenza significativamente maggiore con ibandronato ($P < 0,001$). Nella discussione, però, si ammette che la radioterapia resta il trattamento di scelta. Inoltre, dato che RT e I hanno un differente meccanismo di azione sul dolore, è verosimile che la loro combinazione sia più efficace dei singoli trattamenti: su questo punto siamo d'accordo con gli autori ed auspichiamo che presto si avvii uno studio in tal senso.

6. Minor point. Ormai va di moda riportare molte tabelle e molte procedure, anche importanti, in un'appendice che viene pubblicata solo online. Così è accaduto anche con questo lavoro (v. scheda). Vogliamo stigmatizzare questa abitudine che ostacola il lettore comune che spesso non è abbonato alla rivista in cui il lavoro è pubblicato (non si può essere abbonati a tutte le riviste!) e che, quindi, ha difficoltà a rintracciare tale appendice che pur agevola la valutazione critica del lavoro (peraltro, talvolta le informazioni riportate in appendice non sono affatto secondarie). Se tale moda si rafforzasse, potremmo giungere ad un lavoro pubblicato a stampa, condensato in un paio di pagine per poi riportare tutte le informazioni – anche quelle importanti – nell'appendice digitale. Se questa *malpractice* fosse dovuta solo ad un problema di contenimento della spesa per il costo della carta e della spedizione, tutte le riviste dovrebbero optare per la sola pubblicazione online, lasciando ampi spazi a tutti i dettagli che gli autori desiderano fornire e, naturalmente, abbassare drasticamente i costi dell'abbonamento. •