

Prevenzione della cardiotoxicità da trastuzumab nelle pazienti con carcinoma mammario HER2 positive

Enzo Ballatori
Statistico medico,
Spinetoli (AP)

Fausto Roila
SC di Oncologia Medica
Azienda Ospedaliera Universitaria
"S. Maria della Misericordia", Perugia

RIASSUNTO

Sono stati discussi metodi e risultati di uno studio in cui un ACE-inibitore e un beta-bloccante sono usati per prevenire la cardiotoxicità del trastuzumab in pazienti con carcinoma mammario operato HER2 positive e le conclusioni degli autori appaiono inaccettabili.

Parole chiave. Dimensione dello studio, potenza dello studio, analisi della varianza, test *t* di Student, studi a tre bracci.

SUMMARY

Prevention of cardiotoxicity in patients with breast cancer over-expressing HER2

We discuss both methods and results of a study where an ACE-inhibitor and a beta-blocker were used to prevent cardiotoxicity induced by trastuzumab in patients with early breast cancer HER2 positive. Authors' conclusions seem to be unacceptable.

Key words. Sample size, study power, ANOVA, Student's *t*-test, three arms studies.

Background

Nelle pazienti affette da carcinoma mammario con sovra-espressione di HER2 l'aggiunta di trastuzumab migliora l'efficacia della chemioterapia, ma tale farmaco è cardiotossico, inducendo in circa un caso su 6 disfunzione cardiaca e in un caso su 25 insufficienza cardiaca. Al manifestarsi della cardiotoxicità, trastuzumab viene sospeso e la paziente viene trattata con ACE-inibitori o con beta-bloccanti. L'idea centrale dello studio è considerare tali trattamenti come profilassi per prevenire la cardiotoxicità indotta dal trastuzumab. Lo scopo dello studio è dunque rilevante, ma ha importanti limitazioni che non consentono di ritenere del tutto affidabili i risultati conseguiti ed anche la sua programmazione presenta alcuni aspetti non chiari. Oltre a discutere di tali shortcomings, nella presente nota si coglie l'occasione di puntualizzare alcuni strumenti usati per l'analisi dei dati.

Endpoint principale dello studio è il rimodellamento cardiaco, valutato con RM cardiaca, espresso come cambiamento del volume cardiaco del ventricolo sinistro alla fine

della diastole, al termine del trattamento con trastuzumab (cioè al ciclo 17, ossia dopo circa 52 settimane dall'inizio del trattamento) rispetto al basale. La frazione di eiezione del ventricolo sinistro (LVEF) è considerata endpoint secondario.

a. Principale difetto dello studio: il basso numero delle pazienti arruolate.

1. Nella pianificazione dello studio. Il calcolo del *sample size* è basato:

- su uno studio pilota in cui la percentuale delle pazienti con cambiamento del volume cardiaco fu stimato essere dell'11% (1 caso su 9). Uno studio pilota di norma è eseguito su un piccolo numero di pazienti e, pertanto, la variabilità della stima è elevata. Una notevole diversità di tale stima dal vero valore del parametro produce grandi variazioni nella determinazione della dimensione dello studio e, pertanto, un atteggiamento più prudente avrebbe imposto di determinare un intervallo di confidenza per quell'11% assumendo come stima del parametro il valore corrispondente alla massima numerosità delle pazienti da arruolare;
- è stato ipotizzato che l'uso di ACE-inibitori o di beta-bloccanti avrebbe prevenuto il 90% di tali cambiamenti. Questa ipotesi ci sembra francamente irrealistica, e tale dubbio è confermato dai risultati dello studio stesso.

Probabilmente gli autori sapevano fin dall'inizio di non poter arruolare un notevole numero di pazienti ed hanno adattato stime ed assunzioni per l'adeguamento della numerosità campionaria alla loro realtà. Ci si può chiedere, però, che senso abbia il calcolo del *sample size* basato su assunzioni forzate; ma si sa, nessuna rivista accetterebbe mai un articolo che non contenesse tali calcoli che, come si è visto, nello studio in oggetto sono puri artefatti. Inoltre, il vantaggio di operare con una elevata numerosità campionaria è quello di valutare efficacia e tollerabilità in un gruppo di pazienti più eterogenei, dando consistenza e rilevanza clinica ai risultati ottenuti.

2. In sede di analisi dei risultati. A fronte delle 159 pazienti programmate, in realtà ne sono state analizzate solo 99, con una riduzione di oltre 1/3 rispetto a quanto programmato. La perdita di potenza dello studio è pertanto rilevante e ciò induce a ritenere che alcune differenze trovate non significative (comprese quelle relative all'endpoint primario) sarebbero potute esserlo in presenza di una maggiore numerosità campionaria.

b. Lo studio è a tre bracci: ciò produce una notevole maggiore complessità rispetto ad uno studio a due bracci dovuta ad una serie di fattori.

1. Siano A, B e C i 3 trattamenti. I confronti possibili sono 3; ad esempio, A vs B, A vs C, B vs C. In alcuni casi però uno dei confronti potrebbe essere irrilevante e pertanto i confronti potrebbero essere ridotti a due. Comunque, eseguendo due o più confronti sullo stesso materiale sperimentale occorre sempre tener presente la disuguaglianza di Bonferroni (v. "Statistica per concetti" in CASCO 1). In particolare andrebbe precisato nella pianificazione dello studio quali confronti saranno eseguiti in sede di elaborazione dei dati, cosa che nel lavoro in discussione non è stata fatta. A prima vista sembrerebbero due: bisoprolol (BIS) vs placebo (PL) e perindopril (PER) vs PL. Leggendo però attentamente la scheda, sono stati eseguiti tutti e tre i confronti (nella scheda evidenziati in grassetto), evidentemente anche per valutare se i due trattamenti attivi (ACE-inibitore e beta-bloccante) avessero una diversa efficacia.

2. Aumentata dimensione del campione. La numerosità campionaria per uno studio a tre bracci è più che doppia rispetto a quella richiesta per uno studio a due bracci e

dipende anche dal numero di confronti che saranno eseguiti nella elaborazione dei dati.

c. Uso dell'analisi della varianza. Nelle tabelle 1, 2 e 3 (per brevità non riportate nella scheda) in cui vengono riportati i risultati dello studio, per eseguire i confronti fra i tre trattamenti (BIS, PER, PL) si usa l'analisi della varianza ad una via (indicata dagli autori con ANOVA). Il problema affrontato dall'ANOVA è il confronto tra le medie di 3 (o più) popolazioni: nell'articolo quella delle pazienti trattate con BIS, con PER e con PL. Per eseguire l'ANOVA con un solo fattore di classificazione (ANOVA 1-way) si usa il test F di Fisher-Snedecor. L'ipotesi nulla è che, nelle tre popolazioni, la media sia la stessa. L'ipotesi alternativa è che **almeno una media sia diversa dalle altre**. Quindi se il test F è significativo vuol dire che nelle 3 popolazioni almeno una media è diversa dalle altre; se è non significativo vuol dire che non ci sono elementi per ritenere che le medie non siano tutte e tre uguali.

Quindi se il test F risulta significativo, il passo successivo è quello di vedere quali medie sono diverse e quali, invece, possono essere ritenute uguali. Ad esempio, nella tab. 3 dell'articolo, il cambiamento di LVEF al ciclo 17 ri-

SCHEDA

Pituskin E, Mackey JR, Koshman S, et al. Multidisciplinary approach to novel therapies in cardio-oncology research (manticore 101-breast): a randomized trial for the prevention of trastuzumab-associated cardiotoxicity. J Clin Oncol 2017; 36: 870-8.

Gli HER2 sono sovra-espressi nel 20-25% delle pazienti con carcinoma mammario e sono associati ad una precoce progressione di malattia. In tali pazienti, l'aggiunta di trastuzumab alla chemioterapia adiuvante pur migliorando significativamente gli outcome di sopravvivenza della chemioterapia, induce però disfunzione cardiaca fin nel 18% dei casi, con insufficienza cardiaca fin nel 4% dei casi. Una ridotta frazione di eiezione del ventricolo sinistro (LVEF) porta ad un'interruzione del trastuzumab con conseguente aumento del rischio di morte. La risonanza magnetica cardiaca (MRI) è raccomandata dalle linee guida per la conferma delle disfunzioni cardiache indotte dalla terapia. ACE-inibitori e beta-bloccanti sono raccomandati come terapia per

la cura di disfunzioni cardiache e per prevenire l'insufficienza cardiaca. Scopo del lavoro è provare se tali farmaci possono essere utili per prevenire la disfunzione cardiaca.

Metodi Disegno dello studio

Lo studio doppio cieco, randomizzato a tre bracci (1:1:1), con blocchi di 6 pazienti, volto a verificare se perindopril (PER) o bisoprolol (BIS) siano superiori al placebo (PL) nella prevenzione delle disfunzioni cardiache.

Pazienti arruolate

Pazienti con carcinoma mammario operato (Early Breast Cancer, EBC) HER2 positive furono arruolate consecutivamente da due programmi canadesi. Criteri di eleggibilità: età > 18 anni, con EBC (stadio da I a IIIA), con sovra-espressione di HER2 e con trattamento pianificato con trastuzumab. Criteri di esclusione: LVEF < 50% al basale, storia di insufficienza cardiaca, cardiomiopatia e infarto del miocardio, con controindicazioni

all'uso di beta-bloccanti o di ACE-inibitori, con ipertensione incontrollata, e controindicazioni alla MRI.

Outcome

Obiettivo primario è il rimodellamento cardiaco, valutato con MRI, ed espresso come cambiamento del volume cardiaco (del ventricolo sinistro alla fine della diastole) dal basale al termine del trattamento con trastuzumab (al ciclo 17, circa 52 settimane dall'inizio della terapia). Endpoint secondario: cambiamento della LVEF dal baseline alla fine del trattamento con trastuzumab. La "cardiotoxicità clinica" fu definita come interruzione di trastuzumab per più di 7 giorni a causa di una peggiorata funzionalità cardiaca.

Analisi statistica

Sample size. Basandosi su un precedente studio pilota fu stimata pari all'11% la frazione di pazienti con cambiamento del volume cardiaco indotto da trastuzumab, e fu ipotizzato che l'uso di ACE-inibitori o di beta-bloccanti

spetto al basale è stato nei tre bracci (media \pm deviazione standard)

PL: -5 ± 5 ; PER: -3 ± 4 ; BIS: -1 ± 5 ;

$P < 0,001$ rispetto al test F

La significatività di F attesta che almeno una media è diversa dalle altre; ad es., potrebbe essere che la media della popolazione PL sia diversa sia da quella della popolazione BIS sia da quella della popolazione PER, oppure che la media di PL sia diversa da quella di BIS, ma non da quella di PER, e così via. Per trovare quali siano le medie diverse dalle altre esistono diversi test per confronti multipli, il più affidabile dei quali è il test di Scheffé, ma non si possono eseguire test t di Student, come invece hanno fatto gli autori, pena incorrere nella disuguaglianza di Bonferroni, nemmeno accennata dagli autori.

Una seconda obiezione all'uso dell'ANOVA deriva dal fatto che F è un test parametrico, nel senso che la sua validità è condizionata al verificarsi di due assunzioni: a) uguaglianza delle varianze nelle tre popolazioni, b) distribuzione normale del carattere nelle tre popolazioni. Esistono test per verificare tali assunzioni, ma in questo caso la loro potenza è limitata dalla bassa dimensione

campionaria; in altre parole, con circa 30 pazienti per gruppo, è assai difficile che possano risultare significativi: è probabile che diano un risultato non significativo, ma data la loro scarsa potenza, tale risultato potrebbe essere dovuto al basso numero di pazienti arruolate e non al fatto che nelle tre popolazioni siano effettivamente verificate le due assunzioni a) e b).

In generale, quando la numerosità campionaria è scarsa, è impossibile verificare se le assunzioni alla base dei test parametrici possano ritenersi valide; è allora opportuno procedere con test non parametrici: il test di Kruskal-Wallis è il test non parametrico corrispondente al test F dell'ANOVA, mentre il test non parametrico corrispondente al t-test è il test di Wilcoxon per la somma dei ranghi (o il suo equivalente test U di Mann-Whitney); anche quest'ultimo, però, deve tener conto della disuguaglianza di Bonferroni.

d. Difetti di minore rilevanza.

1. Randomizzazione a blocchi di 6. La più rilevante proprietà della randomizzazione è la sua imprevedibilità, nel senso che non dovrebbe mai essere prevedibile a quale trattamento verrà allocato il successivo paziente arruo-

avrebbero prevenuto il 90% di tali casi. Fissando una deviazione standard all'interno dei gruppi di 20 ml, 47 pazienti per braccio sarebbero stati sufficienti per provare l'efficacia di ciascuno dei due farmaci vs PL ad un livello di significatività per un test bidirezionale del 5%, con una probabilità dell'80%. Attendendosi un tasso di drop-out del 10% e un tasso di mortalità del 3% fu deciso di arruolare complessivamente 159 pazienti.

L'analisi degli endpoint clinici fu condotta in accordo al criterio di intenzione a trattare (ITT) e non furono considerati le pazienti con missing data. t-test bidirezionali o un'analisi della varianza ad una via furono usati per determinare la significatività delle differenze tra i gruppi per le variabili continue. Il chi-quadro fu utilizzato per il confronto tra variabili categoriche. Un $P < 0,05$ fu usato come misura di significatività per tutti i confronti.

Un modello di analisi della regressione multipla fu utilizzato per individuare i predittori del cambiamento del volume cardiaco.

Risultati

Furono valutate per eleggibilità 402 pazienti e ne furono escluse 303 soprattutto perché o le loro caratteristiche non erano in accordo con i criteri di inclusione (161) o hanno rifiutato di partecipare (120). Furono dunque randomizzate 99 pazienti ai tre gruppi: 32 (PL), 34 (PER), 33 (BIS). Escludendo quelle che sospesero il trattamento, le pazienti analizzate furono 30 (PL), 33 (PER), 31 (BIS), in totale 94 che, nella maggioranza dei casi (77%) furono sottoposte ad una chemioterapia non contenente antracicline e furono osservate per un periodo medio di 350 ± 18 giorni. Il massimo livello di dose del farmaco nei tre bracci fu del 90% (PL), 75% (PER, dose media: $6,8 \pm 2$ mg), 65% (BIS, dose media: $7,7 \pm 3$ mg, $P < 0,06$, n.s.). Non ci fu alcun evento avverso grave.

I cambiamenti medi del volume cardiaco dopo la somministrazione di trastuzumab furono non significativamente diversi tra i gruppi (PL: $+4 \pm 11$ ml/mq; PER: $+7 \pm 14$ ml/mq; BIS: $+8 \pm 9$ ml/mq; $P < 0,36$). Comunque, **il cambiamento assoluto in LVEF al termine della terapia fu**

più basso con BIS confrontato sia con PL che con PER: $-1 \pm 5\%$ vs $-3 \pm 4\%$ vs $-5 \pm 5\%$, rispettivamente; $P < 0,001$).

Dopo il 4° ciclo, l'incidenza di disfunzione cardiaca fu meno comune tra le pazienti trattate con PER (1/33) e con BIS (1/31) rispetto a PL (6/30; $P < 0,02$), ma non ci furono differenze significative dopo il ciclo 17.

Le interruzioni di trastuzumab per abbassamento di LVEF furono meno frequenti nel gruppo PER (3/33 pazienti) e nel gruppo BIS (3/31 pazienti) rispetto al gruppo PL (9/30 pazienti; $P < 0,03$).

L'analisi multivariata identificò solo il valore basale come predittivo del cambiamento nel volume cardiaco, mentre i predittori di una diminuzione di LVEF furono il valore basale di LVEF e l'uso profilattico di PER ($P < 0,016$) e di BIS ($P < 0,001$).

Conclusioni

Sia peridopril (PER) che bisoprolol (BIS) attenuano il declino di LVEF dovuto al trattamento con trastuzumab, ma non prevengono il rimodellamento del ventricolo sinistro che era l'outcome primario dello studio. •

lato. Usando per 3 trattamenti blocchi di randomizzazione di 6 unità, dopo che al più 5 di esse sono state assegnate ai trattamenti, è prevedibile a quale trattamento sarà assegnata la sesta. Forse blocchi più numerosi (ad es., di 12 unità) sarebbero stati preferibili.

2. Analisi multifattoriale. È stata condotta un'analisi multifattoriale per tentare di identificare i predittori dell'endpoint primario (rimodellamento del volume cardiaco) e di quello secondario (LVEF). Per quanto riguarda l'endpoint primario, solo il valore del basale è stato trovato significativo, mentre per l'endpoint secondario sono risultati significativi anche i trattamenti (v. scheda). Probabilmente se la dimensione dello studio fosse stata considerevolmente più alta sarebbero risultati significativi (anche) altri predittori dei due endpoint.

3. Livello di significatività. Nel lavoro commentato il livello di significatività è indicato con " $P < 0,05$ ". È un'imprecisione in quanto anche $P = 0,05$ può ritenersi significativo. Quindi, gli autori avrebbero dovuto scrivere " $P \leq 0,05$ ".

In conclusione, lo studio ci sembra inadeguato soprattutto per dimensione per poter ritenere affidabili i risultati conseguiti. Gli autori concludono che le due profilassi attive non hanno effetto sull'endpoint primario ma hanno efficacia nel ridurre la diminuzione di LVEF dal basale al ciclo 17. Per quanto esposto sopra, riteniamo che entrambe le conclusioni non siano accettabili: la prima (endpoint primario) per insufficienza di numerosità del campione, la seconda (endpoint secondario) per l'uso di tecniche statistiche inadeguate. •