Abiraterone: che fatigue?

Enzo Ballatori
Docente di Statistica Medica, Facoltà di Medicina e Chirurgia, Università di L'Aquila

Fausto Roila
SC di Oncologia Medica Azienda Ospedaliera di Terni

Questa rubrica si occupa di fisiopatologia degli studi clinici. In questa occasione, siamo lieti di essere nel campo di una buona fisiologia, in quanto, a nostro avviso, lo studio riportato sinteticamente nella scheda 2 (studio A) è pressoché ineccepibile. Resta, però, da spiegare l'apparente discrepanza di risultati circa l'effetto dell'abiraterone sulla fatigue (affaticamento) rispetto allo studio che è stato presentato in forma di extended abstract all'ECCO 2011 (studio B, v. scheda riepilogativa). Infatti, nessun effetto sulla fatigue era stato evidenziato nello studio A, mentre nell'altro (oggetto del presente commento) emerge un deciso beneficio di abiraterone anche su questo aspetto. Del resto, non sembrano sussistere nemmeno diretti benefici per il marketing, dato che i risultati dello studio A sono di per sé sufficienti a cambiare la pratica clinica, malgrado gli eventi avversi correlati ai mineralcorticoidi fossero più frequenti nel gruppo di trattamento che in quello di controllo: ritenzione di fluidi ed edema (31 vs 22%), ipertensione (10 vs 8%), ipocalcemia (17 vs 8%).

Il primo punto da osservare è che, come è uso comune negli studi di Good Clinical Practice (GCP), l'effetto di abiraterone sulla fatigue è stato valutato nello studio A con gli ***NCI-Common Toxicity Criteria (NCI-CTC)***, mentre nello studio B con il ***Brief Fatigue Inventory (BFI)***, somministrato ai pazienti nel corso dello studio A con finalità dichiaratamente esplorative.

L'NCI-CTC valuta la fatigue come:
1. leggera, cioè compatibile con le attività della vita quotidiana,
2. moderata, che produce difficoltà nel fare alcune attività quotidiane,
3. grave, che interferisce pesantemente nelle attività quotidiane,
4. disabilitante.

Invece, il BFI è composto da 10 item: il primo a risposta binaria, mentre per gli altri, la risposta va collocata su una scala a 11 punti, da 0 a 10. Il primo item mira ad accertare se il soggetto ha fatigue o meno, il secondo il livello della fatigue al momento della rilevazione, il terzo il livello "usuale" di fatigue nelle ultime 24 ore, il quarto il peggior livello nelle ultime 24 ore, dal quinto al decimo se, nelle ultime 24 ore, la fatigue ha interferito con 6 aspetti della qualità di vita (v. scheda riepilogativa)[1].

Non sarebbe da meravigliarsi se due strumenti volti a valutare i livelli dello stesso concetto danno luogo a diversi risultati: ciò è accaduto perfino con i due più usati questionari per la misura della qualità di vita (il QLQ-C30 dell'EORTC e il FACT-G), che, somministrati agli stessi pazienti, hanno fornito risultati che presentavano deboli correlazioni nella misura dei livelli delle stesse dimensioni[2].

Nel nostro caso, però, la determinante più importante della diversità dei risultati è la metodologia della valutazione: nello studio A si è semplicemente misurata l'incidenza della fatigue e della fatigue severa nell'intero periodo di osservazione. Nello studio B, invece, sono stati selezionati due gruppi di pazienti che avevano al basale un livello non inferiore a 5 (su un massimo di 10)
a. dell'intensità massima di fatigue nel giorno precedente (**maxF**),
b. dell'interferenza media della fatigue con i 6 aspetti della vita quotidiana considerati dal BFI, **intF**).

I pazienti sono stati seguiti nel tempo per valutare l'evoluzione differenziale della fatigue nel gruppo trattato con abiraterone rispetto al gruppo di controllo.

Vi sono critiche ai risultati dello studio B che concernono la scelta e l'uso dello strumento adoperato per valutare la fatigue, il BFI.

Omissione. Quando si misura il livello di un concetto, le più importanti dimensioni da valutare sono **incidenza**, **intensità**, **durata** e **rilevanza** (cioè, impatto sulla qualità di vita). Nel BFI manca la valutazione della durata: un conto è soffrire di fatigue costantemente tutto il giorno, un conto è soffrirne a tratti, peraltro, con intensità differenti. In altre parole, ben diversa è la condizione dei pazienti in cui l'intensità massima di fatigue registrata nel questionario si riferisce all'intera giornata, da quella dei pazienti che ne hanno sofferto, anche molto, ma per brevi periodi della giornata.

Validazione. Poche sono le versioni linguistiche validate del BFI: filippino, giapponese, cinese (Cina Popolare e Taiwan), coreano, tedesco, oltre che inglese, mentre 13 sono i paesi partecipanti allo studio. Quindi, il BFI non è stato validato in tutti i paesi in cui è stato condotto lo studio. La carenza più importante, però, è la insufficiente validazione cui è stato sottoposto il BFI.

Com'è noto, un questionario psicometrico va validato per affidabilità, validità, responsività. L'**affidabilità** si articola in coerenza interna e riproducibilità. La **coerenza interna** consiste nel verificare che le due parti in cui può essere suddiviso il questionario, comunque scelte, sono in accordo nel valutare lo stesso dominio; di norma la si

valuta con l'indice α di Cronbach. La **riproducibilità** è l'attitudine dello strumento a fornire all'incirca la stessa misura in caso di stabilità del contesto: la si valuta con tecniche di tipo test-retest, ossia facendo compilare lo stesso questionario due (o più volte) allo stesso paziente, a distanza di tempo, e misurando il grado di accordo tra le risposte fornite. La **validità** è la capacità dello strumento di misurare ciò per cui è stato progettato; in assenza di gold standard (e questa è la norma), esiste tutta una gamma di prove che attesta da vari punti di vista la validità di un questionario (validità di contenuto, v. discriminante, v. legata al criterio, e così via). La **responsività** (al cambiamento) è l'attitudine di un questionario di percepire (attraverso gli score che fornisce) i mutamenti delle condizioni del paziente nel tempo[4].

Il BFI non è stato mai validato né per riproducibilità, né per responsività[3]. Sono due lacune importanti, ma la seconda è particolarmente grave, dato che nel lavoro sintetizzato nella scheda di riferimento, è stato proprio usato per seguire l'andamento della fatigue dei pazienti nel tempo.

Nello studio A, è stato altresì usato il FACT-P (un altro strumento della serie FACIT, ideata da David Cella, v. scheda 2) per la misura della qualità di vita nei pazienti con carcinoma prostatico, di cui, però, non sono stati riportati i risultati (o almeno, non ancora). Non è chiaro perché la scelta sia ricaduta sul BFI quando esiste il questionario FACT-F (una sottoscala di 13 item, quindi semplice quasi quanto il BFI) specifico per la valutazione della fatigue che, non solo è più omogeneo al FACT-P, ma soprattutto ha subito un ben più robusto processo di validazione. Ad esempio, la valutazione della sua responsività lo ha messo in grado di individuare il cambiamento minimo clinicamente significativo[3].

Definizione. Nell'abstract dello studio B si parla di miglioramento e di peggioramento della fatigue, senza precisare come tali mutamenti siano stati determinati. È sufficiente un decremento anche di lievissima entità nel punteggio, o, invece, è necessario che lo score si abbassi oltre una prestabilita quantità? Ovviamente, i risultati dell'analisi dipendono da tale scelta, che, peraltro, è del tutto soggettiva.

Tutto sommato, tali critiche hanno un moderato impatto sull'attendibilità dei risultati per via della randomizzazione.

SCHEDA RIEPILOGATIVA

Sternberg CN, Scher HI, Molina A, et al. Fatigue Improvement/Reduction with Abiraterone Acetate in Patients with Metastatic Castration-Resistant Prostate Cancer (mCRPC) Post-Docetaxel: Results From the COU-AA-301 Phase 3 Study. Eur J Cancer 2011; 47 (Suppl. 3): 488-9, abstr. 7015

Basandosi sui dati dello studio sintetizzato nella successiva scheda 2, in cui ai pazienti era stato anche somministrato il questionario Brief Fatigue Inventory (BFI), al basale (cioè prima dell'inizio del trattamento), e dopo ogni ciclo di terapia di 28 giorni, è stato retrospettivamente valutato l'effetto dell'abiraterone sulla fatigue. Sono stati investigati due aspetti della fatigue: l'intensità massima (item 3: peggior fatigue nelle precedenti 24 ore) e la sua interferenza con aspetti della vita quotidiana (item 4.a: attività in generale, 4.b: umore, 4.c: capacità di camminare, 4.d: lavori usuali, 4.e: relazioni, 4.f: apprezzamento della vita), misurata calcolando la media della interferenza della fatigue sui 6 aspetti. Sono stati considerati eleggibili, separatamente, i pazienti che, al basale, presentavano uno score ≥5 (su un massimo di 10) della peggior fatigue o dell'interferenza media della fatigue sulla vita quotidiana.

Risultati

Pazienti	A + P	PL + P	P <
• randomizzati	797	398	
BFI: Intensità			
• eleggibili	384	186	ns
• migliorati, n. (%)	221 (58)	75 (40)	0,0001
• tempo al miglioramento* (Me§, gg)	59	194	0,012
• tempo alla progressione* (P25&, gg)	232	139	0,002
BFI: Interferenza			
• eleggibili	189	92	ns
• migliorati	103 (55)	35 (38)	0,010
• tempo al miglioramento* (Me§, gg)	57	113	ns
• tempo alla progressione* (P25&, gg)	281	139	0,001

A = abiraterone, PL = placebo, P = prednisone; ns = non significativo; *della fatigue; §mediana; &25° percentile.

Conclusioni

Abiraterone ritarda la progressione della fatigue e produce miglioramenti negli score della fatigue rispetto al basale; inoltre, migliora la fatigue più rapidamente del placebo.

Sebbene non sufficientemente validato, il BFI misura comunque qualcosa di attinente alla fatigue e, anche se manca la valutazione della durata della fatigue nelle 24 ore precedenti e non è chiaro come siano stati definiti il miglioramento e il peggioramento, pur tuttavia i risultati si riferiscono a due gruppi randomizzati e concludono che, nel controllare la fatigue, almeno in alcune sue dimensioni, l'abiraterone è più efficace del placebo.

Ciò che invece rende inaccettabili i risultati dello studio B sono le limitazioni inerenti alla metodologia usata. In particolare:

Arbitrarietà. La scelta di selezionare i pazienti che, al basale, avevano score della fatigue (maxF, intF) non inferiori a 0,5 è puramente arbitraria. Infatti, si sarebbero potuti considerare altri cut-off, come ad esempio un cut-off di 0,2 o 0,3 (considerando così i pazienti che soffrivano un po' di affaticamento), ovvero di 0,7 o 0,8 (individuando i pazienti con una forte fatigue che interferiva pesantemente sulla loro vita quotidiana). Per ogni scelta si sarebbero ottenuti risultati diversi.

La motivazione della scelta degli autori, che manca nell'abstract, diventa così cruciale non solo per chiarire le finalità dello studio, ma anche ai fini dell'interpretazione dei risultati.

Una seconda scelta arbitraria ci sembra quella di considerare cicli di chemioterapia di 28 giorni (al termine di ciascuno dei quali è stato compilato il BFI), quando in realtà abiraterone e prednisone sono farmaci che si assumono ogni giorno, in modo continuativo nel tempo. Un effetto importante di tale scelta è l'imprecisione delle valutazioni del tempo al miglioramento e del tempo alla progressione della fatigue (nell'abstract non è stata definita la progressione della fatigue): se fossero stati fissati cicli di terapia di 21 o di 14 giorni, l'imprecisione sarebbe diminuita.

Entrambi questi elementi di arbitrarietà probabilmente rispondono ad istanze pratiche e non hanno nessun supporto teorico (questo aspetto è decisamente negativo per lo studio): nel primo caso la scelta di un cut-off superiore a 0,5 non avrebbe probabilmente consentito l'analisi perché sarebbe stato individuato un sottogruppo troppo esiguo di pa-

SCHEDA 2

de Bono JS, Logothetis CJ, Molina A, et al. Abiraterone and Increased Survival in Metastatic Prostatic Cancer. N Engl J Med 2011; 364: 1995-2005

Studio randomizzato, doppio cieco, controllato con placebo, condotto in 147 centri di 13 paesi, volto a valutare l'effetto sulla sopravvivenza di abiraterone acetato (AA) + prednisone (P) vs placebo (PL) + P in pazienti affetti da carcinoma della prostata metastatico, in progressione di malattia, dopo trattamento con docetaxel e con ECOG performance status (PS) ≤ 2. Criteri di esclusione: livelli anormali di transaminasi, metastasi epatiche, gravi malattie concomitanti.

Randomizzazione

I pazienti prima di essere randomizzati in rapporto 2:1 a ricevere, rispettivamente, AA + P o PL + P, furono stratificati rispetto ai seguenti caratteri:
- ECOG PS: 0-1 vs 2;
- livello del peggior dolore nelle precedenti 24 ore, valutato con il Brief Pain Inventory-short form 0-3 (dolore non clinicamente rilevante) vs 4-10 (dolore grave);
- numero di precedenti chemioterapie: 1 vs 2;
- tipo di evidenza di progressione: solo incremento di PSA vs evidenza radiografica.

Endopint

Primario: Overall Survival (OS)
Secondari: risposta obiettiva (lesioni tessuti molli), risposta sul PSA (decremento ≥ 50%, confermato dopo 4 settimane), tempo alla progressione del PSA (almeno il 25% di incremento rispetto al basale).
Altri endpoint usati a fini di una valutazione esplorativa:
- valutazione dell'impatto della malattia/trattamento sulla vita quotidiana del paziente eseguita con il Functional Assessment of Cancer Therapy-Prostate (FACT-P);
- **punteggio di fatigue, valutato con il Brief Fatigue Inventory (BFI)**;
- conta di cellule tumorali circolanti;
- informazioni sulle risorse consumate.

Dimensione del campione

Assumendo un hazard ratio di morte dell'80% per il gruppo di trattamento rispetto a quello di controllo, fissato un livello di significatività del 5% (test bidirezionale), 1158 avrebbero consentito di individuare con l'85% di probabilità una differenza significativa. Fu pianificata un'analisi ad interim.

Risultati di efficacia

Furono reclutati 1195 pazienti, 797 nel gruppo di trattamento (AA + P) e 398 in quello di controllo (PL + P).
All'analisi ad interim risultò che la riduzione di rischio nel gruppo di trattamento fu del 35,4% (hazard ratio: 0,65, 95%CI: 0,54-0,77; P < 0,001): sopravvivenza mediana di 14,8 mesi nel gruppo di trattamento e di 10,9 mesi in quello di controllo. Anche per la buona tollerabilità del trattamento, fu allora consentito il passaggio dei pazienti del gruppo di controllo al trattamento con abiraterone (cross over).

Safety

Tutti gli eventi avversi vennero valutati applicando gli NCI-Common Toxicity Criteria.
Il più comune evento avverso fu la fatigue, che ebbe un'analoga incidenza e severità nei due gruppi. •

zienti, mentre, nel secondo caso, fissare cicli di terapia di durata inferiore a 28 giorni avrebbe richiesto una più frequente somministrazione dei questionari.

Confronto. Mediante gli NCI-CTC, nello studio A è stata considerata la presenza (o meno) di fatigue (di qualunque livello) e di fatigue severa in tutti i pazienti randomizzati ai due trattamenti, nell'intero periodo di tempo dello studio, e si è visto che non vi è alcuna differenza.

Nello studio B, invece, sono stati selezionati i pazienti con maxF e intF non inferiori a 0,5 e solo per questi è stato valutato l'andamento delle fatigue nel tempo: **non sono stati** pertanto **considerati gli altri pazienti**. Nel confronto tra i risultati ottenuti occorre tener conto che potrebbero essersi presentati vari inconvenienti, tra cui uno particolarmente grave: i pazienti non considerati (quelli che presentavano un cut-off inferiore a 0,5) potrebbero aver avuto un incremento di fatigue, nei periodi successivi al basale, in misura più marcata nel gruppo di trattamento con abiraterone rispetto al gruppo di controllo; se così fosse accaduto, tale evento sarebbe sfuggito all'analisi riportata nello studio B. D'altronde, se assumiamo che i due metodi di valutazioni (NCI-CTC e BFI) diano luogo a risultati abbastanza concordanti, la situazione descritta potrebbe essere vicina a quella reale, in quanto, in tal modo, si ricomporrebbe il quadro esposto nello studio A (nessuna differenza tra abiraterone e placebo): a compensare la maggiore efficacia di abiraterone nei pazienti che avevano una almeno discreta fatigue, ci sarebbe la presenza di una maggiore insorgenza di fatigue o di un suo peggioramento, rispettivamente in coloro che non l'avevano o la percepivano lieve nel gruppo dei soggetti trattati con abiraterone. Le conclusioni dello studio B, pertanto, potrebbero essere parziali ed irrealistiche.

In secondo luogo, data l'arbitrarietà della scelta del cut-off, nessuno può garantire che i pazienti con livelli di fatigue immediatamente inferiori a quelli stabiliti dal cut-off (ad es., 0,45) si comportino in modo analogo a quello dei pazienti osservati.

In conclusione, a nostro avviso, l'effetto specifico di abiraterone sulla fatigue non appare sufficientemente provato e, quindi, non resta che accogliere il verdetto di parità tra abiraterone e placebo sancito nello studio A. •

Bibliografia

1. *Mendoza TR, Wang XS, Cleeland CS, et al. The rapid assessment of fatigue severity in cancer patients: use of the Brief Fatigue Inventory. Cancer 1999; 85: 1186-96.*
2. *Holzner B, Kemmler G, Sperner-Unterweger B, et al. Quality of life measurement in oncology: a matter of the assessment instrument? Eur J Cancer 2001; 37: 2349-56.*
3. *Minton O, Stone P. A systematic review of the scales used for the measurement of cancer-related fatigue (CRF). Ann Oncol 2009; 20: 17-25.*
4. *Apolone G, Ballatori E, Mosconi P, Roila F. Misurare la qualità di vita in Oncologia. Roma: Il Pensiero Scientifico Editore, 1997; pp. XII + 112.*

Statistica per concetti

Randomizzazione

Nel lavoro sintetizzato nella scheda 2 compare un uso per certi versi non comune della randomizzazione. Pertanto, si è deciso di dedicare a questo argomento la presente rubrica, discutendo, alla fine, le peculiarità della randomizzazione adoperata.

La randomizzazione, operazione fondamentale nelle scienze sperimentali, può definirsi come l'allocazione **rigorosamente casuale** delle unità sperimentali ai gruppi di trattamento. L'avverbio "rigorosamente" indica che l'assegnazione casuale delle unità sperimentali ai trattamenti è eseguita in base ad un metodo ben preciso, e non con il significato di "senza un criterio", che l'aggettivo "casuale" ha spesso nel linguaggio comune.

Ad esempio, dovendo randomizzare 16 pazienti a due terapie, A e B, si può ricorrere alle tavole di numeri aleatori, costruite con una procedura di estrazione casuale di numeri (gioco del lotto): è sufficiente immaginare una fila di 16 pazienti, leggere 8 numeri di due cifre ciascuno compresi tra 01 e 16, e decidere di assegnare i corrispondenti pazienti al trattamento A; per differenza gli altri verranno assegnati a B. Ad es., se i primi 8 numeri letti sono 09, 04, 12, 01, 05, 14, 16, 08, i pazienti che, nell'ordine della fila occuperanno i posti 1, 4, 5, 8, 9, 12, 14, 16 verranno assegnati ad A, gli altri a B.

Qualunque sia il metodo seguito (tavole aleatorie, generazione di numeri pseudo-casuali mediante computer), la randomizzazione deve essere sempre

- **riproducibile**, nel senso che, una volta descritto il procedimento, chiunque lo applichi ottenga sempre lo stesso risultato (quindi, il lancio di una moneta non può essere utilizzato per randomizzare);
- **imprevedibile**, cioè che non sia quasi mai possibile prevedere a quale trattamento verrà assegnata la prossima unità. Infatti, se il medico sperimentatore sapesse a quale trattamento verrà assegnato il successivo paziente, potrebbe essere condizionato nella decisione di reclutarlo (ad es., se sa che il paziente sarà randomizzato al trattamento più tossico, potrebbe non proporre l'ingresso nello studio ad un paziente defedato, malgrado egli soddisfi i criteri di selezione).

Gli **scopi** più importanti della randomizzazione sono:

a. **produrre un accettabile bilanciamento dei gruppi** di trattamento **rispetto ai fattori prognostici noti e sconosciuti**;
b. **fornire una** seconda **base logica al test statistico** che verrà usato per la valutazione di efficacia/tollerabilità differenziali.