

# Pitfalls nella ricerca clinica. Il caso dei nuovi farmaci in Oncologia

Enzo Ballatori<sup>1</sup>, Fausto Roila<sup>2</sup>, Benedetta Ruggeri<sup>3</sup>

<sup>1</sup> Statistico Medico, Spinetoli (AP) - Italy

<sup>2</sup> Oncologia Medica, Ospedale di Terni - Italy

<sup>3</sup> Governo Clinico, AV 5, ASUR Marche - Italy

## Pitfalls in clinical research. The case of the new antineoplastic agents

Pitfalls are often present in clinical research. In this paper, we highlight some shortcomings affecting the comparative studies on the efficacy of the new drugs in Oncology. More precisely, we will focus on the frequent choice of an intermediate endpoint – instead of the final endpoint of effectiveness – and the also frequent practice to allow the patient to crossover to the alternative treatment, when the proof of a greater effectiveness has not yet been reached. The effects of the Bonferroni's inequality on the results of a clinical study, and the choice of the comparator, as well as the use of the placebo, are also discussed. Finally, the importance of an appropriate procedure to determine the appropriate sample size is pointed out. An example from published studies is presented after the discussion of each issue.

**Keywords:** *Ad interim* analysis, Bonferroni's inequality, Effectiveness and activity, Minimum difference clinically relevant, Placebo, Surrogate endpoints, Treatment crossover

## Introduzione

I nuovi farmaci oncologici forniscono la prova tangibile delle discrepanze che esistono tra la teoria, suffragata da prove di laboratorio, e i risultati della ricerca clinica. Idee geniali, come quella di inibire il meccanismo dell'angiogenesi, vengono concretizzate in farmaci che, tuttavia, spesso mostrano una limitata efficacia quando somministrati ai pazienti oncologici. Corollario a tale premessa è che per valutare l'efficacia di un nuovo farmaco l'unica strada percorribile è la sperimentazione clinica.

Lo sviluppo di nuovi farmaci è un processo assai costoso per l'industria farmaceutica. Quindi non c'è da stupirsi che vengano poste in essere procedure tese a valorizzare gli effetti positivi del nuovo farmaco. Talvolta però l'industria farmaceutica, che è anche lo sponsor della ricerca clinica, sia nella programmazione dello studio che nell'analisi e interpretazione dei risultati, applica in modo quantomeno non appropriato le metodologie statistiche allo scopo di enfatizzare i risultati. È pertanto fondamentale che le Autorità regolatorie, che dovrebbero vigilare attentamente su efficacia e sicurezza dei trattamenti, e il decisore di spesa, che dovrebbe mirare a un'allocazione ottimale delle risorse pubbliche, traggano da questi esempi lo stimolo a un più convinto uso delle procedure di *Health Technology Assessment* (HTA).

Il valore di alcuni dei nuovi farmaci è stato dimostrato con studi ineccepibili. In altri casi, invece, le prove di maggiore efficacia rispetto alla terapia standard sono discutibili.

Scopo del presente lavoro è evidenziare alcuni dei trabocchetti (*pitfalls*) che possono celarsi nella ricerca clinica, così che Autorità regolatorie, decisori di spesa, Comitati etici, oncologi prescrittori e utenti finali – i pazienti, rappresentati dalle loro associazioni – possano avere gli strumenti per una valutazione indipendente dell'efficacia e della sicurezza dei nuovi farmaci.

L'esposizione di ciascuno dei trabocchetti considerati è seguita da un solo esempio concreto (raramente più di uno), utile per ancorare i concetti alla realtà, ma occorre avvertire che sono numerosi gli studi pubblicati che contengono i *pitfalls* di seguito descritti.

## Efficacia e attività

Si definiscono:

- "efficacia" l'azione del trattamento sul paziente
- "attività" l'effetto del trattamento sulla malattia.

Mentre diversi possono essere gli *endpoints* di attività, quali ad esempio la scomparsa o la notevole riduzione della massa tumorale (detti, rispettivamente, "Risposta completa" e "Risposta parziale"), solo due sono gli *endpoints* di efficacia: Sopravvivenza globale (*Overall Survival*, OS: tempo trascorso dall'inizio del trattamento alla morte del paziente per qualunque causa) e Qualità di vita (*Quality of Life*, QoL, misurata per mezzo di appositi strumenti, come, ad esempio, i questionari psicometrici).

Per valutare l'efficacia di un nuovo trattamento, tuttavia, spesso sono usati altri *endpoints*, detti **endpoints intermedi**, riservando la dizione di **endpoints surrogati** a quegli *endpoints* intermedi di cui sia stata provata una stretta relazione con l'*endpoint* finale.

Accepted: October 20, 2014

Published online: November 28, 2014

## Indirizzo per la corrispondenza

Enzo Ballatori  
Via Pillino 2  
63078 Spinetoli (AP), Italy  
e.ballatori@alice.it

In Oncologia ce ne sono parecchi (e sono tutti *endpoints* di attività), ma il più usato è la sopravvivenza libera da progressione (*Progression Free Survival*, PFS) misurata dal tempo che intercorre tra l'inizio del trattamento e la progressione di malattia o la morte del paziente per qualunque causa, se questa interviene prima della progressione.

C'è però da chiedersi se la PFS sia un buon *endpoint* surrogato della OS, nel senso che i pazienti che hanno avuto una più lunga PFS abbiano avuto anche un'aumentata sopravvivenza. La risposta non è necessariamente affermativa. Infatti, negli studi comparativi, ci sono casi in cui:

- a) la PFS è non significativamente differente tra i trattamenti, ma la OS è più lunga con il nuovo trattamento;
- b) la PFS è superiore con il nuovo trattamento, mentre la OS non è significativamente differente tra i due gruppi.

Il caso (a), difficilmente spiegabile, è poco frequente; ne riportiamo qui un esempio.

#### **Esempio 1. Regorafenib nel colon-retto metastatico.**

Regorafenib è stato approvato per pazienti con carcinoma del colon-retto metastatico che sono già stati trattati con le terapie disponibili, o che non sono candidati a riceverle.

Regorafenib è stato valutato in uno studio in doppio cieco controllato con placebo in 760 pazienti, randomizzati 2:1 entro 3 mesi dalla progressione dopo qualsiasi precedente terapia. Tutti i pazienti erano stati trattati in precedenza con bevacizumab: 505 ricevettero regorafenib e 255 placebo (1). L'*endpoint* primario era la OS.

La OS mediana era risultata significativamente superiore con regorafenib (6,4 vs 5,0 mesi) evidenziando una riduzione del rischio di morte (*Hazard Ratio*, HR) del 23% rispetto al placebo. La PFS mediana era di 1,9 mesi con regorafenib vs 1,7 mesi con placebo (differenza non significativa). La percentuale di risposte era rispettivamente dell'1,0% e dello 0,4%. Gli eventi avversi, significativamente più frequenti con regorafenib (grado 3 o 4 nel 54% vs il 14% dei pazienti), hanno portato a interrompere il trattamento nel 61% vs il 22% dei pazienti. Gli eventi avversi di grado 3 o 4 più comuni con regorafenib sono stati: sindrome mani-piedi (17% dei pazienti), *fatigue* (10%), diarrea (7%), ipertensione (7%) e rash cutaneo (6%). Il deterioramento della qualità di vita, valutata con il questionario QLQ-C30 dell'EORTC (*European Organisation for Research and Treatment of Cancer*), era simile nei due bracci di trattamento.

Il farmaco presenta una maggiore efficacia marginale (+1,4 mesi di OS) con importante tossicità, ma è difficile capire come esso riesca dopo la sua sospensione a migliorare significativamente la OS, dato che non determina un aumento significativo della PFS.

Il caso (b), invece, è molto comune. La principale ragione di ciò è dovuta al fatto che, per discutibili motivi etici, una volta provato che il nuovo farmaco migliora significativamente la PFS, si consente il crossover del trattamento.

#### **Crossover del trattamento**

Nessuno saprà mai quante analisi *ad interim*<sup>a</sup> non programmate si facciano durante uno studio clinico, violando così il teorema di Bonferroni (vedi sezione Reiterazione dei test statistici), perché la proprietà del dato è dell'industria.

Certamente se ne producono in occasione dei congressi più importanti, veri e propri trampolini di lancio sul mercato del nuovo farmaco.

Quando l'*endpoint* primario è la PFS e ad un'analisi *ad interim* si trova che il nuovo trattamento è significativamente superiore a quello di controllo sorgerà, almeno in apparenza, un problema etico: se si proseguisse lo studio nei termini previsti dal protocollo, i pazienti del braccio di controllo sembrerebbero poco tutelati perché continuerebbero a ricevere il trattamento meno efficace. Allora, in tal caso, spesso si concede al paziente del gruppo di controllo di passare al nuovo farmaco.

Tale procedura è aberrante per diverse ragioni:

- 1) Solo chi non ha esperienza di Statistica può ritenere definitivo un risultato che, per definizione di analisi *ad interim*, è invece solo provvisorio. Quante volte una differenza statisticamente significativa si riassume nel prosieguo dell'arruolamento!
- 2) Così operando, non si saprà mai qual è l'efficacia differenziale dei trattamenti (in termini, quindi, di OS) in quanto molti pazienti del braccio di controllo ricevono anche il nuovo trattamento.

Due conclusioni:

- a) Le analisi *ad interim* non dovrebbero essere eseguite o, se proprio non se ne può/vuole fare a meno, dovrebbero essere programmate, precisando nel protocollo dello studio il loro numero, lo *step* dello studio nel quale eseguirle, le *stopping rules* predefinite sulla base del numero di analisi, l'aggiustamento del livello di significatività per ciascuna analisi sulla base della disuguaglianza di Bonferroni (vedi sezione Reiterazione dei test statistici). Mai dovrebbero essere accettate dalla comunità scientifica analisi *ad interim* non programmate.
- b) Il crossover dei pazienti del braccio di controllo al nuovo trattamento non dovrebbe essere consentito, o almeno dovrebbe poggiare su basi più certe; più precisamente, il crossover potrebbe essere consentito solo quando, all'accumularsi dei dati, si sia potuto provare che la OS (e non la sola PFS) è significativamente superiore con il nuovo trattamento. Consentendo il crossover in conseguenza solo di una significativamente superiore PFS, in pratica, si diminuisce drasticamente il valore dei risultati di uno studio clinico perché non si conoscerà mai se e di quanto il nuovo trattamento prolunghi la sopravvivenza del paziente rispetto alla terapia standard.

#### **Esempio 2. Vemurafenib nel melanoma maligno metastatico.**

A nostra conoscenza lo studio registrativo di vemurafenib è l'unico caso in cui si sia consentito il crossover del trattamento dopo averne provato (con un'analisi *ad interim*) la superiorità, rispetto al trattamento di controllo, in termini di OS.

La registrazione di vemurafenib (nuovo farmaco) è avvenuta sulla base dei risultati di uno studio randomizzato di fase III, di confronto con dacarbazina (trattamento di controllo) come prima linea di terapia in 675 pazienti affetti da melanoma metastatico con mutazione V600E (2). Vemurafenib è stato usato a dosi di 960 mg per via orale 2 volte al giorno (bid). Gli *endpoints* primari erano la PFS e la OS. Il crossover dei pazienti trattati con dacarbazina a vemurafenib era pre-

visto solo dopo che con vemurafenib fosse stato dimostrato un aumento statisticamente significativo della OS (non della sola PFS). Un'analisi *ad interim* ha evidenziato un aumento significativo della OS mediana a 6 mesi con vemurafenib (percentuale di sopravvissuti dal 64% all'84%), della PFS mediana (1,6 vs 5,3 mesi) e della percentuale di risposte obiettive (dal 5% al 48%). I dati aggiornati a 12 mesi hanno confermato la maggiore efficacia del nuovo farmaco.

Vemurafenib ha rappresentato, insieme con ipilimumab, la prima novità terapeutica importante nella cura del melanoma maligno metastatico negli ultimi 20 anni.

La ragione più importante del crossover, quando viene provata la superiorità del nuovo trattamento in termini della sola PFS, risiede nelle pressioni che vengono fatte da più parti (ricercatori coinvolti nello studio, associazioni di pazienti) per fare in modo che anche i pazienti del gruppo di controllo possano beneficiare il prima possibile del nuovo trattamento.

### **Esempio 3.** Bevacizumab nel carcinoma metastatico dell'ovaio.

Lo studio GOG-0218 è uno studio in doppio cieco in cui le pazienti erano randomizzate a ricevere: a) carboplatino + paclitaxel (gruppo di controllo); b) l'aggiunta a carboplatino + paclitaxel di bevacizumab (un farmaco antiangiogenetico) alla dose di 15 mg/kg ogni 3 settimane dal 2° al 6° ciclo di chemioterapia; c) oltre a bevacizumab nei primi 6 cicli una dose di mantenimento ogni 21 giorni fino a un totale di 21 cicli (15 mesi di terapia) (3). L'*endpoint* principale inizialmente era la OS, ma, secondo quanto riferito dagli autori, è stato poi cambiato con la PFS in quanto lo studio non sarebbe stato più fattibile per le pressioni dei medici investigatori e delle pazienti. Pertanto, considerando che solo se fosse stato mantenuto in cieco il trattamento dopo la progressione della malattia si sarebbe potuto evitare il crossover, la valutazione accurata dell'impatto di bevacizumab sulla OS non è stata possibile. È stata ritenuta necessaria una numerosità di 1800 pazienti per evidenziare con il 90% di probabilità una riduzione del rischio di progressione o morte del 23% con l'aggiunta di bevacizumab. Sono entrate nello studio 1873 pazienti. La PFS mediana è stata di 10,3 mesi nel gruppo di controllo, 11,2 mesi con 5 cicli di bevacizumab e 14,1 mesi con 21 cicli (riduzione del rischio di progressione o morte rispetto al gruppo di controllo di circa il 10% e il 28%, rispettivamente). La OS mediana non era significativamente differente (39,3-38,7-39,7 mesi, rispettivamente) così come l'impatto di bevacizumab sulla qualità di vita valutata con il questionario FACT. Come previsto, l'incidenza di eventi avversi era aumentata con l'aggiunta di bevacizumab.

Consideriamo ora il caso (b) quando il crossover del trattamento non sia consentito.

### **Esempio 4.** Axitinib nel carcinoma del rene metastatico.

Axitinib è stato approvato in seconda linea di terapia del carcinoma renale metastatico in seguito ai risultati di uno studio di fase III che ha confrontato axitinib (5 mg bid) con sorafenib (400 mg bid) (4). L'*endpoint* primario dello studio era la PFS. Nei 723 pazienti arruolati, la PFS mediana è risultata significativamente superiore con axitinib (6,7 mesi vs 4,7 mesi). Il beneficio osservato era superiore nei pazienti che, in

prima linea, avevano ricevuto citochine rispetto a quelli che avevano ricevuto sunitinib. Anche la percentuale di risposte era superiore con axitinib (19% vs 9%). La OS mediana, i cui risultati sono stati riportati più recentemente (5), non era tuttavia significativamente differente fra i due trattamenti (20,1 vs 19,2 mesi), malgrado non fosse stato consentito il crossover. Gli eventi avversi più comuni erano ipertensione, diarrea e *fatigue* con axitinib ed eritrodissiestesia palmo-plantare e alopecia con sorafenib. Il tempo mediano al deterioramento dei sintomi era a favore di axitinib, probabilmente perché i pazienti avevano una progressione della malattia più tardiva. In seguito alla progressione di malattia la differenza nel tempo al deterioramento dei sintomi, tra i due bracci, diminuiva verso il fine vita.

Il mancato impatto sulla OS mediana e il modesto miglioramento della PFS mediana permettono di concludere che axitinib aggiunge poco in termini di efficacia a sorafenib. Inoltre, un altro studio, in cui si confrontava sorafenib con axitinib in prima linea di terapia del carcinoma metastatico del rene, non ha evidenziato alcun beneficio del farmaco rispetto al controllo (6).

Se la PFS è superiore con il nuovo trattamento, mentre la OS non è significativamente differente, potrebbe essere accaduto che la sopravvivenza post-progressione (*Survival Post Progression*, SPP - che va dalla ripresa di malattia alla morte per qualunque causa) non fosse la stessa per i pazienti che hanno avuto una PFS "breve" o una PFS "lunga". Se questi ultimi avessero un valore di SPP assai più breve, l'ipotesi che la PFS sia un buon *endpoint* surrogato di OS potrebbe rivelarsi infondata. Inoltre, occorre considerare che:

- a) mentre la sopravvivenza globale è certa (basta guardare la data del decesso nel certificato di morte), la PFS non lo è. Infatti, ci si accorge della progressione di malattia in occasione di una visita di follow-up (ma, in tal caso, la progressione per alcuni pazienti potrebbe essere iniziata da poco, per altri da molto più tempo e comunque la precisione della sua determinazione dipende dalla frequenza delle visite di follow-up) o per l'insorgere di una nuova sintomatologia che l'oncologo riconosce come dovuta alla progressione (ma, al riguardo, il comportamento dei pazienti è assai difforme: all'insorgere della nuova sintomatologia, alcuni si precipitano dall'oncologo, altri soprassedono sperando che il sintomo cessi spontaneamente);
- b) spesso alla progressione i pazienti sono sottoposti a successive linee di terapia che possono avere la loro efficacia e contribuire a prolungare la sopravvivenza.

A queste ultime obiezioni si potrebbe rispondere che, essendo lo studio clinico randomizzato, tutti i fattori prognostici, noti e sconosciuti, sono abbastanza ben bilanciati fra i trattamenti. In altre parole, per il principio di randomizzazione il paziente viene randomizzato al trattamento con tutte le sue caratteristiche note (compresa la frequenza delle visite di follow-up, stabilita nel protocollo dello studio, vedi (a)) e sconosciute (tra cui le terapie successive alla progressione, vedi (b)). Per approfondimenti si rinvia a CA-SCO (7) (*Current Advances in Supportive Care in Oncology* - una rivista on-line ad accesso gratuito, che viene distribuita anche su supporto cartaceo e che contiene la rubrica "Stati-



stica per concetti” ove sono presentati importanti concetti di statistica usata nelle applicazioni mediche, senza quasi introdurre formule).

In conclusione, le obiezioni (a) e (b) sono superate dalla randomizzazione, a condizione, però, che la frequenza delle visite di follow-up prevista dal protocollo sia la stessa per i due bracci di trattamento.

Inoltre, considerando l’obiezione (b) ci si potrebbe domandare se il trattamento superiore in termini di PFS sia veramente utile, dato che, con le successive linee di terapia, si ottiene poi una sopravvivenza globale (e, quindi, un’efficacia) simile nei due bracci di trattamento.

Nel 2009 è stato pubblicato un articolo sul *Journal of the National Cancer Institute* – una delle più importanti riviste di Oncologia – in cui sono esposti i risultati di un rigoroso studio sulla relazione tra PFS e OS (8). Gli autori hanno mostrato che la PFS è un buon *endpoint* surrogato di OS solo quando la SPP è lunga. In altre parole, se la SPP è breve, la relazione tra PFS e OS è debole. Nel 2012, sull’*European Journal of Cancer* è apparso un altro articolo sull’argomento (9), in cui gli autori, usando metodologie differenti, hanno confermato tali risultati, proponendo l’introduzione di un *endpoint* composto tra PFS e qualità di vita da usare come *endpoint* intermedio, in luogo della semplice PFS. Tali risultati suggeriscono dunque di usare la PFS solo quando la prevista SPP sia lunga; d’altronde, se la SPP è breve, non c’è ragione di ricorrere a un *endpoint* intermedio in quanto è fattibile l’adozione di quello finale (OS).

A volte l’efficacia di un trattamento può essere valutata solo molto tempo dopo l’inizio di uno studio clinico, con la conseguenza di privare i pazienti di un trattamento potenzialmente utile ovvero di poterne accertare l’efficacia quando, a causa di nuove scoperte, esso è già obsoleto. Ad esempio, ci sono alcune neoplasie che consentono a molti pazienti una lunga sopravvivenza – come ad esempio i tumori della mammella e dell’ovaio – per cui sembra lecito avvalersi in questi casi della PFS anziché dell’OS (d’altronde, in tali casi, la SPP è spesso lunga, il che legittima di per sé l’uso della PFS).

Sono trascorsi cinque anni dalla pubblicazione del primo dei due articoli citati (8), eppure si usa ancora indiscriminatamente la PFS come *endpoint* surrogato di OS.

## La reiterazione di test statistici

Fissare il livello di significatività di un test statistico al 5% vuol dire che, se l’ipotesi nulla di uguale efficacia di due trattamenti è vera, c’è il 5% di probabilità di sbagliare nel respingerla, ma c’è anche il 95% di probabilità di indovinare accettandola.

Siano A e B i trattamenti e C1 e C2 due sottogruppi della stessa popolazione identificati dalla presenza o meno nei pazienti di un determinato fattore; ripetendo lo stesso test due volte, in ciascuno dei due sottocollettivi, nell’ipotesi che i due trattamenti abbiano la stessa efficacia, la probabilità di accettare in entrambi i casi l’ipotesi nulla per ciascun confronto (o contrasto) è pari a  $0,95 \times 0,95 = 0,9025$ . Quindi, la probabilità di respingere l’ipotesi nulla almeno una volta (sbagliando, perché A e B sono ugualmente efficaci) è pari a  $1 - 0,9025 = 0,0975$ , quasi il doppio del 5% fissato convenzionalmente. Un evento con

una tale probabilità non può essere considerato “raro” e, pertanto, cade il ragionamento alla base del test statistico<sup>b</sup>.

Occorre quindi modificare il livello di significatività per ciascun confronto in modo che sia pari al 5% il livello di significatività complessivo. Tale modifica – che chiameremo **correzione di Bonferroni** – può essere apportata in vari modi, ad esempio dividendo (approssimativamente) il livello di significatività per il numero dei confronti eseguiti. Nell’esempio considerato, il livello di significatività di ciascun confronto potrebbe esser posto al 2,5% per mantenere al 5% il livello di significatività complessivo. Se i confronti fossero quattro, per mantenere al 5% la probabilità di sbagliare almeno una volta dichiarando diversamente efficaci i due trattamenti, si potrebbe fissare (all’incirca) all’1,25% il livello di significatività per ciascun confronto.

La situazione diventa intrattabile quando i confronti fra i trattamenti sono ripetuti numerose volte; in tali casi, a meno che non ci siano grandi differenze di efficacia, nessun confronto risulterà mai significativo (vedi CASCO, “Statistica per concetti” (10), e, per ulteriori approfondimenti, si veda (11)).

La correzione di Bonferroni si applica ogni qualvolta si eseguano più volte test statistici sullo stesso materiale sperimentale. Pertanto, nella ricerca clinica, essa non solo è importante nelle analisi dei sottocollettivi, ma anche nelle analisi *ad interim*, così frequenti nello sviluppo di un farmaco. Diventa essenziale, quindi, che tali analisi siano programmate anche per modificare adeguatamente il livello di significatività.

### Esempio 5. La saga dei RAS.

Gli studi che avevano portato alla registrazione di cetuximab e di panitumumab avevano ottenuto risultati statisticamente significativi, ma clinicamente poco rilevanti. Furono allora eseguite analisi per sottogruppi, tutte retrospettive, relative allo stato KRAS (*wild type* vs mutato) accertato in una percentuale variabile di pazienti (non su tutti i pazienti), talora ripetute più volte nello stesso studio sulla base delle determinazioni dello stato KRAS al momento disponibili. Tali analisi hanno evidenziato che, nei pazienti con KRAS *wild type* le risposte cliniche miglioravano sostanzialmente, mentre nei pazienti con KRAS mutato erano simili se non addirittura peggiori. Sulla base di tali risultati, gli Enti regolatori rapidamente mutavano l’approvazione di cetuximab e di panitumumab riservandola ai soli pazienti con KRAS *wild type*. Si osservi che, malgrado i molteplici confronti eseguiti (spesso non programmati), a nessuno è venuto in mente di apportare la correzione di Bonferroni al livello di significatività. È comprensibile che non sia venuto in mente all’industria, lo è molto meno che questa lacuna non sia stata considerata dagli Enti regolatori che, a quanto ci consta, per la prima volta hanno dato l’approvazione a un farmaco sulla base dei risultati di analisi per sottogruppi e non di studi prospettici randomizzati. A nostro avviso, i risultati di analisi per sottogruppi debbono essere considerati non come prove ultime, ma suggestivi di ipotesi da testare con studi pianificati “ad hoc”. In realtà, due studi indipendenti furono successivamente eseguiti aggiungendo o meno cetuximab alla chemioterapia: il primo (MCR COIN *trial*) in Inghilterra, Irlanda e Belgio valutando 715 pazienti (tutti con KRAS *wild type*) (12), il secondo (NORDIC-VII) arruolando 194 pazienti con KRAS *wild type* e 130 pazienti con KRAS mutato (13). Entrambi gli studi hanno mostrato risultati

non significativi. Ciò ha scatenato una ridda di ipotesi (es., che in presenza degli altri farmaci usati nel *trial* cetuximab perdesse efficacia) sulle ragioni del mancato successo dei nuovi farmaci, ipotesi volte a far passare il messaggio che i risultati di analisi per sottogruppi avrebbero più valore di quelli ottenuti da studi clinici programmati *ad hoc*.

## Il “comparator”

Nelle sperimentazioni cliniche longitudinali randomizzate è necessario che il trattamento di controllo (*comparator*), cioè quello con cui confrontare il nuovo trattamento, sia la migliore terapia esistente, per motivi sia etici che metodologici.

Infatti, da un lato, sarebbe non etico privare la metà dei pazienti in studio del migliore trattamento esistente e, dall'altro, la scelta di una terapia sub-ottimale come controllo condurrebbe a risultati inutili ai fini della pratica clinica, perché, nel caso che il nuovo trattamento risulti superiore al *comparator*, non sapremmo comunque se sia superiore, equivalente o inferiore alla terapia più efficace disponibile.

### Esempio 6. Afatinib nel carcinoma polmonare metastatico.

Afatinib è un inibitore irreversibile dell'EGFR (*Endothelial Growth Factor Receptor*), approvato per la prima linea di terapia nei pazienti con adenocarcinoma del polmone con mutazione dell'EGFR. Lo studio registrativo era uno studio di fase III in prima linea di terapia (studio LUX-lung 3), che ha confrontato afatinib *versus* la combinazione di cisplatino e pemetrexed in pazienti prevalentemente asiatici con adenocarcinoma del polmone localmente avanzato o disseminato. Sono entrati nello studio 345 pazienti: 230 hanno ricevuto afatinib 40 mg os/die e 115 la chemioterapia. L'*endpoint* primario era la PFS. La PFS mediana era significativamente aumentata da 6,9 mesi con la chemioterapia a 11,1 mesi con afatinib (riduzione dell'HR del 42%). Anche la percentuale di risposte ottenute era significativamente superiore con afatinib (56% *vs* 23%). La OS mediana non era invece significativamente differente (16,6 mesi *vs* 14,8 mesi) (14), probabilmente perché, alla progressione, i pazienti potevano scegliere il crossover al trattamento sperimentale. In tale studio è stato anche valutato l'impatto dei trattamenti sulla qualità di vita, usando il questionario EORTC QLQ-C30 +LC13 (modulo per le neoplasie polmonari). Il tempo al deterioramento dei sintomi, specie tosse e dispnea, ma non il dolore, era significativamente più lungo con afatinib (15). L'incidenza degli effetti collaterali di grado 3 o 4 non era significativamente differente tra i due gruppi di pazienti (49% *vs* 48%). Ma diarrea, rash cutaneo, secchezza e irritazione della cute e delle mucose erano più frequenti con afatinib, mentre diminuzione di appetito, *fatigue*, nausea e vomito e mielosoppressione si presentavano più spesso nel braccio della chemioterapia.

Questi risultati sono stati recentemente confermati da un altro studio, eseguito in pazienti asiatici con mutazione EGFR, che ha confrontato afatinib *versus* la combinazione di cisplatino e gemcitabina in prima linea di trattamento (16). I pazienti erano randomizzati 2:1 a ricevere afatinib (242 pazienti) o cisplatino più gemcitabina (122 pazienti). Anche in questo studio la PFS era l'*endpoint* primario. La PFS mediana era significativamente aumentata da 5,6 mesi a 11,0 mesi. Anche

la percentuale di risposte era superiore con afatinib (66,9% *vs* 23%). La OS mediana era rispettivamente di 22,1 mesi *versus* 22,2 mesi, ma anche in questo caso, come nello studio registrativo, vi era stato un crossover di circa il 48% dei pazienti.

I risultati ottenuti con afatinib sono simili a quelli di gefitinib ed erlotinib (altri due inibitori di EGFR), la cui efficacia era stata trovata in precedenza superiore a quella della sola chemioterapia. Vi è un'apparente maggiore tossicità di afatinib rispetto a gefitinib ed erlotinib, specie in termini di rash cutanei e diarrea, ma questa impressione deve essere verificata con studi clinici di confronto attualmente in corso i cui risultati non sono ancora noti. Pertanto, al momento, la scelta tra i tre farmaci in prima linea di chemioterapia nei pazienti con adenocarcinoma del polmone EGFR-mutato dovrebbe essere basata sul costo del trattamento.

Si osservi che, nel carcinoma polmonare metastatico EGFR-mutato, afatinib non è stato confrontato con il trattamento standard disponibile (basato su erlotinib o su gefitinib) ma con la chemioterapia che già in precedenza era stata dimostrata essere meno efficace sia di erlotinib che di gefitinib. Gli studi menzionati, pertanto, debbono essere ritenuti non etici, in quanto i pazienti del gruppo di controllo non hanno ricevuto il miglior trattamento disponibile, e metodologicamente scorretti perché i loro risultati sono inutili per la pratica clinica in quanto non ci consentono di sapere se afatinib sia più o meno efficace di erlotinib o di gefitinib.

## Il placebo

Il placebo è ogni terapia sprovvista di attività specifiche sulla patologia che si deve trattare, impiegata per il suo effetto psicologico o psicofisiologico senza svelare al paziente la sua inerzia.

La prova dell'esistenza dell'effetto placebo fu fornita nel 1955 da Beecher (17), che stimò in oltre un terzo la frazione di pazienti, affetti da differenti patologie, che potrebbero migliorare o guarire con l'uso del placebo.

Sotto la verosimile assunzione che ogni terapia abbia almeno un effetto placebo, uno studio longitudinale randomizzato *versus* placebo dimostra l'attività farmacologica del trattamento, nel senso che, se il test statistico risulta significativo, si è mostrato che il trattamento sperimentale aggiunge qualcosa all'effetto placebo.

Ovviamente in ogni studio comparativo di un trattamento *versus* placebo vanno adottate tecniche di cecità, cioè lo studio deve essere almeno in cieco semplice (*single-blind*), in quanto se il paziente sa di ricevere un placebo, questa conoscenza di per sé vanifica l'effetto placebo. Se nella valutazione di efficacia (o di attività) si può anche solo sospettare che ci sia una componente di soggettività del medico sperimentatore, allora lo studio deve essere condotto in doppio cieco (*double-blind*), cioè nemmeno il medico sperimentatore deve conoscere il trattamento cui è assoggettato il paziente.

Un'importante avvertenza sull'interpretazione dei risultati di uno studio *versus* placebo deriva dalla constatazione che, talvolta, nell'interpretazione dei risultati si considera il gruppo dei pazienti che ricevono il placebo come pazienti non trattati, il che accade soprattutto negli studi sulle terapie di supporto. Questo è un errore, in quanto i pazienti trattati con placebo ricevono comunque un trattamento: uno studio

in cui si confronta il trattamento sperimentale *versus* nessun trattamento probabilmente fornirebbe risultati diversi. Ma quasi mai il trattamento sperimentale viene confrontato con “nessun trattamento” perché, in caso di esito positivo, si potrebbe ipotizzare che l’efficacia del trattamento sperimentale consista solo nell’effetto placebo.

Il placebo può considerarsi come il meno efficace fra i trattamenti sub-ottimali. Quindi, per ragioni etiche e metodologiche, uno studio *versus* placebo può essere condotto solo quando, per la patologia considerata, non sia disponibile un trattamento utile. Infatti, sarebbe non etico somministrare un placebo ai pazienti del gruppo di controllo quando esiste una terapia utile. Inoltre, dal punto di vista della pratica clinica, uno studio *versus* placebo non dà alcuna informazione in quanto i suoi risultati non possono dire nulla sull’efficacia differenziale del nuovo trattamento rispetto alla terapia esistente di già provata efficacia.

Un importante caso particolare è quello in cui il nuovo farmaco sia da somministrare in aggiunta a una terapia già dimostratasi efficace. Ad esempio, se per un tipo di tumore una chemioterapia C ha dimostrato di essere il miglior trattamento, l’efficacia di un nuovo farmaco biologico B può essere testata con uno studio il cui disegno preveda di trattare tutti i pazienti con C randomizzandoli a ricevere, in aggiunta a C, B o un placebo.

Come al solito, è l’abuso che genera problemi. In Oncologia, per la cura di una certa forma neoplastica, sono talvolta pubblicati risultati di studi *versus* placebo di farmaci somministrati in aggiunta alla terapia più efficace (es., nel caso dei farmaci biologici con una presumibilmente simile efficacia). In tal modo, vi sono diversi farmaci che si sono mostrati efficaci, senza che però il clinico ne conosca il valore differenziale. In tali condizioni, non è dato sapere su quale base avviene la scelta nella pratica clinica (pressione del marketing delle diverse case farmaceutiche?). A nostro avviso, una volta che un farmaco si è dimostrato efficace nella cura di una certa patologia, per concedere la registrazione ai successivi farmaci che via via vengono introdotti, le Autorità regolatorie dovrebbero imporre un confronto comparativo con il farmaco in quel momento trovato più efficace, e non richiedere semplicemente uno studio *versus* placebo.

La sperimentazione *versus* placebo di nuovi farmaci in una stessa patologia, insieme alla grande diffusione che stanno avendo gli studi di non-inferiorità, costituiscono i mezzi più importanti per lo sviluppo dei cosiddetti farmaci “*me-too*” che non apportano alcun beneficio aggiuntivo alla salute del paziente.

#### **Esempio 7. Pazopanib nel carcinoma renale metastatico.**

Pazopanib, un antiangiogenetico orale, è stato sperimentato *versus* placebo in uno studio di fase III, in doppio cieco, randomizzato 2:1, in pazienti con carcinoma renale metastatico in prima linea di trattamento (54%) o che avevano già ricevuto citochine (46%). L’*endpoint* primario era la PFS che fu significativamente prolungata con pazopanib (9,2 vs 4,2 mesi). Le principali reazioni avverse furono diarrea, anoressia, nausea, ipertensione, mutamento del colore dei capelli. Non vi fu alcuna evidenza di differenze clinicamente rilevanti nella qualità di vita tra i due gruppi (18).

È uno studio ben fatto, solo che sunitinib si era già dimo-

strato efficace in pazienti dello stesso tipo nelle stesse condizioni. Quindi si tratta di uno studio non etico, perché i pazienti trattati con placebo avrebbero potuto ricevere un trattamento efficace (sunitinib) e, dal punto di vista della pratica clinica, non aggiunge conoscenze utili perché dai suoi risultati non si può decidere se pazopanib sia o meno più efficace di sunitinib.

È interessante osservare la strategia spesso usata per lo sviluppo dei farmaci “*me-too*”: anzitutto si prova che anche il nuovo farmaco abbia un’efficacia (sperimentando *versus* placebo). Successivamente, si va a vedere se sia “non inferiore” al farmaco che ha già dato prove di efficacia. Infatti, pazopanib è stato confrontato, in uno studio di non-inferiorità, con sunitinib in 1110 pazienti e ha dimostrato una PFS non inferiore (19). Nell’interesse del paziente, il confronto diretto (*head-to-head*) si sarebbe dovuto fare prima evitando il primo studio; solo che, se si fosse dimostrato inferiore, pazopanib sarebbe subito uscito dal mercato. Nel consentire di anteporre le esigenze del marketing a quelle del paziente, vi sono precise responsabilità da parte non solo delle Autorità regolatorie, ma anche delle riviste che pubblicano i risultati degli studi *versus* placebo, quando siano già registrati farmaci di sicura efficacia.

#### **Dimensione dello studio**

La determinazione della numerosità campionaria è un momento assai importante e delicato nella programmazione di uno studio clinico. La procedura per pervenirvi è la seguente.

Anzitutto si stabilisce una **differenza minima clinicamente rilevante**, ossia, rispetto all’*endpoint* primario, si fissa qual è la differenza al di sopra della quale si è disposti a introdurre il nuovo trattamento nella pratica clinica.

Si individua il test statistico che sarà utilizzato per il confronto e si fissa il livello di significatività al quale si è disposti a dichiarare i trattamenti diversamente efficaci (di norma, il 5%).

A questo punto ci sono due variabili, funzione matematica l’una dell’altra, la potenza dello studio e la numerosità dei pazienti da arruolare. In altre parole, fissata la potenza dello studio<sup>c</sup> (generalmente l’80% per studi di superiorità), resta univocamente determinato il numero dei pazienti da arruolare.

La scelta del test non è in discussione, né tantomeno il livello di significatività e la potenza dello studio, entrambi questi ultimi due parametri dettati da standard internazionali. Invece, la differenza minima clinicamente rilevante è il punto delicato perché risente di una componente di soggettività. Ad esempio, per un ricercatore, il 10% dei successi in più ottenuti con il nuovo trattamento rispetto alla terapia standard potrebbe essere una differenza consistente per sostituire nella pratica clinica la terapia standard con il nuovo trattamento, anche se il profilo di tossicità di quest’ultimo non è ancora perfettamente delineato e il suo costo è assai superiore. Un altro ricercatore, invece, potrebbe valutare insufficiente il 10% di successi in più, propendendo per il 15% o il 20%.

A parità delle altre condizioni, quanto più grande è la differenza minima clinicamente rilevante prescelta – D – tanto minore risulterà la dimensione dello studio e, viceversa, quanto minore è D, tanto maggiore risulterà il numero dei pazienti da arruolare. La scelta di D si ripercuote sulla potenza

dello studio: se scegliamo D “troppo piccolo” lo studio necessiterà di molti pazienti e avrà così una potenza molto grande; se lo scegliamo “troppo grande” il *trial* sarà condotto su pochi pazienti e avrà una potenza molto bassa.

I rischi sono:

- scegliendo D troppo grande diventa probabile non riconoscere che il nuovo trattamento è più efficace della terapia standard (es., se il nuovo trattamento producesse il 10% di successi in più rispetto alla terapia standard, ma nella programmazione dello studio si è ipotizzata una differenza minima clinicamente rilevante del 20% - su cui è stata calcolata la dimensione dello studio - è probabile che la differenza osservata non risulti significativa, inducendo ad abbandonare il nuovo trattamento che, invece, è il più efficace);
- scegliendo D troppo piccolo si corre il rischio di dichiarare significativa una piccola differenza che, magari, non è clinicamente rilevante (es., un aumento della PFS mediana di due settimane, a parità di OS, potrebbe anche risultare significativa se lo studio è grande, ma potrebbe essere considerata clinicamente irrilevante).

Gli studi clinici sui nuovi farmaci oncologici sono spesso condotti arruolando moltissimi pazienti e, da quanto esposto, se ne può comprendere la ragione. In tal modo, però, si possono ottenere risultati che, seppure statisticamente significativi, non sono clinicamente rilevanti.

## Discussione e conclusioni

Nel corso di una review degli studi sui nuovi farmaci oncologici (20), abbiamo anzitutto riscontrato che buona parte di tali studi sono di almeno discreta qualità, non paragonabile a quella mediamente osservata in passato. Vi sono però parecchi studi che presentano importanti *pitfalls*, quelli discussi nel presente lavoro. Il loro elenco non ha alcuna pretesa di esaustività, ma quelli considerati sono così rilevanti da rendere non affidabili i risultati ottenuti, almeno nei casi in cui la differenza di efficacia fra i trattamenti a confronto, pur significativa, non sia molto accentuata.

Oltre ad avviare una discussione sulla rilevanza clinica dell'efficacia dei nuovi farmaci oncologici, l'utilità del presente lavoro risiede soprattutto in due aspetti:

- a) indurre a una riflessione sui trabocchetti esaminati che, peraltro, sono generalizzabili a tutti gli altri settori della ricerca clinica;
- b) la possibilità di eseguire valutazioni indipendenti per gli stakeholder della ricerca clinica: Autorità regolatorie, Comitati etici locali, decisori di spesa, oncologi prescrittori, farmacisti ospedalieri, associazioni di pazienti.

Il paziente può essere riguardato come un sistema aperto, dove per “sistema” si intende un insieme di parti in interazione tra loro, e “aperto” indica che il sistema (e le sue componenti) sono in interazione con l'ambiente in cui il paziente è immerso. Tale impressionante complessità spiega la variabilità dei risultati che si ottengono tra pazienti diversi, variabilità che si riduce con l'acquisizione di nuove conoscenze di base, ottenute con la biochimica, con la genetica e con la biologia molecolare, ma che è destinata a non poter essere eliminata completamente. Da ciò scaturisce la necessità della ricerca clinica che, unica, può fornire conoscenze ultime da trasferire nella pratica clinica.

La finalità principale della ricerca clinica è il benessere del paziente, inteso come guarigione dalla malattia, o come prolungamento della sopravvivenza o, ancora, come miglioramento della qualità di vita. Nel conseguimento di tale scopo, tuttavia, si intersecano altri obiettivi (come, ad esempio, la necessità dell'industria farmaceutica – in genere sponsor della ricerca – di produrre utili), che danno luogo ai conflitti di interesse generati dalla contiguità con l'industria sia dei componenti dei Comitati preposti all'autorizzazione dei farmaci, sia degli estensori delle linee guida, sia dei ricercatori, sia dei prescrittori (per approfondimenti sui conflitti di interesse si rinvia all'articolo (21), pubblicato sulla rivista *Ars et Scientia*, di libero accesso on line, che si trova nel sito di “Informazioni sui farmaci”, curato dalle Farmacie Comunali Riunite di Reggio Emilia).

In genere, sia un articolo che riporta i risultati di uno studio clinico, sia una presentazione in sede congressuale si concludono con una frase ad effetto (*take-home*) che offre un'estrema sintesi della qualità dei risultati conseguiti. Il *take-home*, strumento di marketing più che strumento scientifico, talora è in contraddizione con i risultati della ricerca stessa, esagerando le virtù e minimizzando gli aspetti negativi del nuovo farmaco, ma è ciò che resta più impresso a chi non abbia eseguito una lettura criticamente attiva del lavoro pubblicato.

Il presente lavoro vuole essere uno strumento per individuare almeno alcuni dei trabocchetti che si possono celare in uno studio apparentemente ben fatto. Se uno o più di questi fossero presenti, a meno di una grande dimensione dell'effetto a favore della nuova terapia, si potrebbe sospendere il giudizio, attendendo i risultati di una ricerca sugli esiti (*outcome research*) che potrebbe essere programmata *ad hoc* per dissipare i dubbi sull'affidabilità dei risultati pubblicati. Ma questo, almeno per l'Italia, è il futuro.

## Note

<sup>a</sup>Ricordando che la traduzione della locuzione latina “ad interim” è “per ora”, l'analisi ad interim è un'analisi provvisoria condotta sui dati dello studio clinico accumulati fino a un certo istante, precedente quello della chiusura dello studio.

<sup>b</sup>Quando, in seguito all'applicazione di un test statistico, si respinge l'ipotesi nulla di uguale efficacia dei trattamenti, possono essere accadute due cose: a) i due trattamenti hanno realmente una diversa efficacia; b) i due trattamenti hanno la stessa efficacia, ma si è presentato un evento raro (uno degli eventi che complessivamente avevano una bassa probabilità di presentarsi, diciamo del 5%). Tra tali due situazioni optiamo per la prima (dichiarando così diversamente efficaci i due trattamenti) in quanto un evento raro con “pratica certezza” non si presenta. Quindi, la logica del test è incentrata sulla rarità dell'evento la cui probabilità è misurata dal livello di significatività.

<sup>c</sup>La potenza di un test statistico, o anche di uno studio clinico, è la probabilità che, se i trattamenti hanno una diversa efficacia, il test risulti significativo. Semplificando, se i trattamenti hanno una diversa efficacia, la potenza dello studio è la probabilità che ce se ne accorga, risultando il test significativo.

## Disclosures

Financial support: This article was supported by the resources of the Authors.

Conflict of interest: The Authors declare that they have no conflicts of interest.

## Bibliografia

1. Grothey A, Van Cutsem E, Sobrero A, et al. Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised,



- placebo-controlled, phase 3 trial. *Lancet*. 2013;381:303-12.
2. Chapman PB, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med*. 2011;364:2507-16.
  3. Burger RA, Brady MF, Bookman MA, et al. Incorporation of bevacizumab in the primary treatment of ovarian cancer. *N Engl J Med*. 2011;365:2473-81.
  4. Rini BI, Escudier B, Tornczak P, et al. Comparative effectiveness of axitinib versus sorafenib in advanced renal cancer carcinoma (AXIS): a randomized phase 3 trial. *Lancet*. 2011;378:1931-9.
  5. Motzer RJ, Escudier B, Tornczak P, et al. Axitinib versus sorafenib as second-line treatment for advanced renal cell carcinoma: overall survival analysis and updated results from a randomized phase 3 trial. *Lancet Oncol*. 2013;14:552-62.
  6. Hutson TE, Lesovay V, Al-Shukri S, et al. Axitinib versus sorafenib as first-line therapy in patients with metastatic renal-cell carcinoma: a randomised open-label phase 3 trial. *Lancet Oncol*. 2013;14:1287-94.
  7. Ballatori E. Statistica per concetti. *Current Advances in Supportive Care in Cancer (CASCO)* 2011;2:60-1; 2013;7:25-7.
  8. Broglio KR, Berry DA. Detecting an overall survival benefit that is derived from progression free survival. *J Natl Cancer Inst*. 2009;101:1642-9.
  9. Amir E, Seruga B, Kwong R, et al. Poor correlation between progression-free survival and overall survival in modern clinical trials: Are composite endpoints the answer? *Eur J Cancer*. 2012;48:385-8.
  10. Ballatori E. Il problema dei confronti multipli. *Current Advances in Supportive Care in Cancer (CASCO)*, 2011;1:26-7.
  11. Ballatori E. *I Fondamenti della Medicina Scientifica*. Perugia: Margiacchi-Galeno Editrice, 2006.
  12. Maughan TS, Adams R, Smith CG, et al. Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet*. 2011;377:2103-14.
  13. Tveit KM, Guren T, Glimelius B, et al. Phase III trial of cetuximab with continuous or intermittent fluorouracil, leucovorin, and oxaliplatin (Nordic FLOX) versus FLOX alone in first-line treatment of metastatic colorectal cancer: the NORDIC-VII study. *J Clin Oncol*. 2012;30:1755-62.
  14. Sequist LV, Jang JC-H, Yamamoto N, et al. Phase III study of afatinib or cisplatin plus pemetrexed in patients with metastatic lung adenocarcinoma with EGFR mutation. *J Clin Oncol*. 2013;31:3327-34.
  15. Yang JC-H, Hirsh V, Schuler M, et al. Symptom control and quality of life in LUX-lung3: a phase III study of afatinib or cisplatin/pemetrexed in patients with advanced lung adenocarcinoma with EGFR mutations. *J Clin Oncol*. 2013;31:3342-50.
  16. Wu Y-L, Zhou C, Hu C-P, et al. Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-lung 6): an open label, randomized phase 3 trial. *Lancet Oncol*. 2014;15:213-22.
  17. Beecher HK. The powerful placebo. *JAMA*. 1955;159:1602-6.
  18. Sternbeg CN, Davis ID, Mardiak J, et al. Pazopanib in locally advanced or metastatic renal cell carcinoma: results of a randomized phase III trial. *J Clin Oncol*. 2010;28:1061-8.
  19. Motzer RJ, Hutson TE, Cella D, et al. Pazopanib versus sunitinib in metastatic renal-cell carcinoma. *N Engl J Med*. 2013;369:722-31.
  20. Roila F, Ballatori E. *Nuovi Farmaci in Oncologia*. Informazioni sui Farmaci. 2014;38(3):60-78.
  21. Ballatori E, Roila F. *Sui conflitti di interesse*. *Ars et Scientia (A&S)* 2010, n. 2.