GIMBE news®
Vol. 2, n. 3
Marzo 2009
www.gimbenews.it

# PILLOLE DI METODOLOGIA DELLA RICERCA

## Quanti soggetti arruolare in un trial? (I)

### Alchimie statistiche per stimare la dimensione del campione

Sin dalla stesura del protocollo di un trial controllato randomizzato (RCT), una delle principali esigenze dei ricercatori è la stima del numero di soggetti da arruolare nello studio, al fine di ottimizzare investimenti, tempo e impegno di professionisti e pazienti. Idealmente un trial dovrebbe essere adeguatamente dimensionato per rilevare la minima differenza clinicamente significativa tra i due trattamenti a confronto; in realtà, numerosi trial vengono condotti su campioni sottodimensionati (*underpowered*), sollevando ragionevoli dubbi sulla loro eticità.

L'item 7 del CONSORT statement richiede di specificare i metodi utilizzati per stimare la dimensione del campione, oltre alle analisi intermedie (*ad-interim analyses*) e ai criteri di interruzione del trial (*stopping rules*), che saranno approfonditi successivamente.

**1. Ipotesi nulla, errore alfa, errore beta**

Per dimostrare la maggiore efficacia del trattamento A (sperimentale) rispetto al trattamento B (controllo) è necessario rifiutare l'ipotesi nulla, secondo la quale i due trattamenti hanno la stessa efficacia. Se il trial dimostra la superiorità di A rispetto a B (o viceversa), l'ipotesi nulla viene rifiutata; in caso contrario il trial è negativo perchè non rileva alcuna differenza tra i due trattamenti. Tuttavia, i risultati dello studio potrebbero essere falsi perchè influenzati da due errori:

- Errore alfa (o di tipo I): la differenza rilevata tra i due trattamenti in realtà non esiste, per cui il risultato è falsamente positivo.
- Errore beta (o di tipo II): il trial non riesce a rilevare la differenza tra i due trattamenti, per cui il risultato è falsamente negativo.

**2. Quali ingredienti sono necessari?**

Per stimare accuratamente quanti soggetti arruolare in un trial è necessario definire:

- Componenti base: livello di significatività statistica, potenza, incidenza attesa dell'end-point primario nel gruppo di controllo e nel gruppo sperimentale.
- Componenti accessorie: *compliance* attesa, rapporto di allocazione diverso da 1:1.

**2.1. Componenti base**

**Livello di significatività statistica**. Definisce il margine di accettabilità di un risultato falsamente positivo e concide con la soglia del *p value*, generalmente fissato al 5% (molto raramente al 1%). Pertanto, se $p > 0.05$ la differenza di eventi osservata tra i due gruppi viene considerata casuale; se $p < 0.05$ l'efficacia del trattamento viene considerata reale e al diminuire del valore di $p$ si riduce la probabilità che la differenza osservata tra i due trattamenti sia dovuta al caso. Ad esempio, $p < 0.001$ ci informa che la probabilità di un risultato falsamente positivo è inferiore a 1/1000. In sintesi:

- $p < 0.05$ = trial statisticamente significativo = risultato non dovuto al caso = ipotesi nulla rifiutata.
- $p > 0.05$ = trial statisticamente non significativo = risultato verosimilmente dovuto al caso = ipotesi nulla non rifiutata.

**Potenza**. E' la capacità del trial di rilevare l'efficacia terapeutica di uno dei trattamenti in studio. La potenza dello studio - complementare all'errore beta - viene generalmente fissata all'80% che corrisponde ad accettare un errore falso negativo del 20%, perchè il trial non riesce a rilevare l'efficacia del trattamento in studio una volta su cinque. Per ridurre la probabilità di un risultato falsamente negativo al 10%, la potenza deve essere aumentata al 90%, con notevole incremento della dimensione del campione.

**Incidenza dell'end-point primario nel gruppo di controllo**. Se la potenza e il livello di significatività statistica vengono definiti in maniera convenzionale, l'incidenza attesa dell'end-point primario nel gruppo di controllo - *Control Event Rate* (CER) - dipende dal rischio basale dei soggetti arruolati. Il CER viene in genere stimato facendo riferimento a studi osservazionali oppure a trial precedenti, ma non sempre risulta accurato: infatti, può risultare inferiore (aumentando la probabilità di un risultato falsamente negativo) o più elevato (determinando una precoce interruzione del trial per ragioni etiche).

**Incidenza dell'end-point primario nel gruppo sperimentale**. L'*Experimental Event Rate* (EER) si correla direttamente all'efficacia dell'intervento sperimentale, in particolare alla sua capacità di ridurre il rischio dell'end point primario nei soggetti trattati rispetto ai controlli. Per tale ragione alcuni sofware richiedono - invece dell'EER - la riduzione del rischio relativo o del rischio assoluto. L'EER è meno prevedibile del CER perchè non sempre esistono in letteratura dati a cui fare riferimento: se è più semplice stimarlo per le terapie farmacologiche - visto che un RCT solitamente è preceduto dagli studi di fase II - è più arduo per altri interventi sanitari.

### 2.2. Componenti accessorie

***Compliance* attesa**. La potenziale *non-compliance* dei pazienti, proporzionale alla durata del follow-up, può sottostimare la dimensione del campione. Infatti, la riduzione della *compliance* influenza l'entità del beneficio terapeutico e condiziona il campione stimato: ad esempio, se in un trial con una *compliance* del 100%,viene stimato un campione di 100 pazienti per braccio, bisognerà arruolarne 280 per braccio se la *compliance* è dell'80%.

***Allocation ratio*.** In rari casi - quando il trattamento sperimentale è molto rischioso, costoso, complesso da somministrare - viene utilizzato un rapporto di allocazione dei pazienti diverso da 1:1. In questi trial, la dimensione del campione è maggiore a parità di potenza dello studio.

### 3. Dalla teoria alla pratica

La stima della dimensione del campione viene calcolata con specifici software previo input delle componenti descritte. Ecco alcuni "trucchi" per ricodare meglio:

• Mantenendo la potenza dello studio e il livello di significatività statistica convenzionali (rispettivamente 80% e 5%), la dimensione del campione:
- aumenta al diminuire del CER e/o dell'efficacia del trattamento sperimentale;
- diminuisce all'aumentare del CER e/o dell'efficacia del trattamento sperimentale.

• Aumentando la potenza dello studio (ad es. al 90%) e/o riducendo il livello di significatività statistica (ad es. all'1%), la dimensione del campione aumenta indipendentemente dal CER e dall'EER.

Pertanto, all'aumentare del benessere della popolazione, diminuisce la possibilità di documentare la superiorità di nuovi trattamenti rispetto a quelli disponibili. Considerato che i mega-trial sono difficilmente sostenibili - specialmente se il ritorno di marketing è dubbio - negli ultimi anni si sono progressivamente fatti largo nel panorama della ricerca sperimentale i trial di equivalenza.

### 4. Dal protocollo del trial alla sua pubblicazione

Il protocollo di un trial e la successiva pubblicazione dovrebbero descrivere tutte le variabili utilizzate per stimare la dimensione del campione: potenza, livello di significatività, CER, EER, percentuale di *non-compliance*. In realtà, lo studio di Chan et coll. rileva un basso livello di trasparenza e un'elevata la probabilità di "acrobazie statistiche" prima della pubblicazione. Solo 11/62 trial analizzati descrivono i dettagli sulla stima della dimensione del campione in maniera completa e consistente sia nel protocollo, sia nel trial pubblicato; solo 37/62 protocolli e 21/62 trial pubblicati descrivono tutte le componenti necessarie per la stima del campione; in 18/34 casi vengono identificate inspiegabili discrepanze tra quanto dichiarato nel protocollo e quanto pubblicato nel trial.

(*continua nel prossimo numero*)

## KEY POINTS

**CHECKLIST PER CALCOLARE IL SAMPLE SIZE**

- **Stimare il CER facendo riferimento a studi precedenti condotti in popolazioni simili**
- **Stabilire per l'end-point primario la differenza minima clinicamente rilevante tra i due gruppi**
- **Definire la potenza dello studio, in relazione alle risorse disponibili**
- **Determinare un livello di significatività statistica accettabile**
- **Stimare l'entità della *non-compliance***

## Per saperne di più

**Letture introduttive**

- Glasziou P, Doll H. Was the study big enough? Two "café" rules. ACP J Club 2007;147(3):A8-9.
- Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348-53.
- Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. Med J Aust 2002;177:256-7.
- Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ 1995;311:1145-8.
- Florey CD. Sample size for beginners. BMJ 1993;306:1181-4.

**Approfondimenti**

- Guyatt GH, Mills EJ, Elbourne D. In the era of systematic reviews, does the size of an individual trial still matter. PLoS Med 2008;5:e4.
- Chan AW, Hróbjartsson A, Jørgensen KJ, et al. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. BMJ 2008;337:a2299.
- Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002; 288:358–62.
- Stenning SP, Parmar MK. Designing randomised trials: both large and small trials are needed. Ann Oncol 2002;13(Suppl 4):131-8.
- Edwards SJ, Lilford RJ, Braunholtz D, et al. Why "underpowered" trials are not necessarily unethical. Lancet 1997;350:804-7.
- Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-4.

**Software**

- PS Power and Sample Size Calculations. Version 3.0, January 2009. Disponibile a: http://biostat.mc.vanderbilt.edu/PowerSampleSize

**Corso avanzato**

- **GIMBE®**. Metodologia della ricerca clinica. Bologna, ottobre-dicembre 2009.