ASA in prevenzione primaria nelle donne. Critical Appraisal. Glossario EBM

Alessandro Battaggia[*,**], Lia Battaglia[*,**], Stefano Berardi[*], Fausto Bodini[*], Anna Longobardi[*], Isabella Fracasso[*,**], Giuditta Motta[*,**], Giulio Rigon[*,**], Maddalena Sarti[*,**], Alberto Vaona[*,**]
[*]*Associazione E.Q.M. (Evidenza, Qualità e Metodo in Medicina Generale);* [**]*SIMG Verona*

Nel numero precedente della rivista abbiamo pubblicato un Critical Appraisal *dell'articolo recentemente comparso sul* New England Journal of Medicine *di P.M. Ridker et al.,* A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women *(N Engl J Med 2005;352:1293-304). In questo numero vengono analizzati in ordine alfabetico tutti gli argomenti EBM affrontati in questa analisi.*

LEGENDA EBM IN ORDINE ALFABETICO

Analisi *Intention To Treat* (ITT)

Ogni analisi dei dati ricavati da una sperimentazione dove viene studiato un "evento" parte dal confronto della frequenza dell'evento riscontrata nel braccio di intervento con la frequenza dell'evento riscontrata nel braccio di controllo. Su questo confronto si costruiscono le principali "misure di efficacia". Nell'articolo in oggetto gli Autori hanno per esempio utilizzato il Rischio Relativo: questa misura di efficacia esprime il rapporto matematico tra le frequenze dell'outcome riscontrate nei due bracci (vedi -> RR). In un modello di analisi "*Intention To Treat*" (ITT) i due gruppi di pazienti da mettere a confronto sono effettivamente costituiti dai soggetti assegnati dalla randomizzazione al braccio di intervento e rispettivamente al braccio di controllo. L'analisi ITT prende così in considerazione i due gruppi creati all'inizio della sperimentazione, quando vengono costruiti i due bracci dello studio (vedi -> RCT). In condizioni ideali un paziente assegnato dalla randomizzazione ad un determinato trattamento dovrebbe seguirlo fino alla fine dello studio o per lo meno fino al momento in cui subisce l'outcome. In realtà durante il follow-up in un trial caratterizzato da un grande numero di soggetti seguiti per periodi di tempo molto lunghi insorge una lunga serie di problemi. Alcuni individui vengono per esempio persi al follow-up ossia fuoriescono dalla sperimentazione e non si conosce più il loro destino. Queste perdite al follow-up (*drop-outers*) possono rappresentare un problema molto grave: se queste perdite sono superiori al 10% dei soggetti randomizzati la validità della sperimentazione è irrimediabilmente compromessa. Per perdite minori è possibile eseguire un tipo particolare di *Sensitivity Analysis* (-> vedi). Altri soggetti dopo la randomizzazione non vengono persi al follow-up ma non rispettano più il protocollo o perché non assumono più il trattamento assegnato (*non compliants*) o perché addirittura assumono il trattamento assegnato al braccio opposto (*cross-overs* o *drop-ins*). Queste violazioni del protocollo (di cui però si conosce l'outcome) rappresentano un problema meno grave per la validità della ricerca in quanto possono essere affrontate entro un modello di Analisi ITT. È da sottolineare come sia ininfluente ai fini della analisi ITT il fatto che dopo essere stato assegnato ad uno dei due bracci il paziente abbia rispettato o no il protocollo. Tale modalità di ragionamento può apparire al clinico molto ostica e poco intuitiva in quanto verrebbe spontaneo considerare in un'analisi di efficacia solo gli outcome subiti dai pazienti che effettivamente hanno seguito il protocollo (compliants: vedi -> analisi Per Protocol). L'Analisi ITT rappresenta invece l'unico mezzo per mantenere intatti durante il follow-up tutti vantaggi conferiti all'inizio dalla randomizzazione, ossia la presenza di due gruppi a confronto perfettamente identici nelle loro caratteristiche di base. Una buona randomizzazione (vedi -> *Allocation Concealment*) rappresenta infatti l'elemento più importante di qualità metodologica di una ricerca. È opportuno ripetere e sottolineare che per eseguire una analisi ITT è indispensabile conoscere se il paziente – indipendentemente da eventuali violazioni del protocollo – abbia o no subito l'outcome: nessuna analisi ITT potrebbe essere applicata a pazienti persi al follow-up. Per le modalità di gestione delle perdite al follow-up vedi -> *Sensitivity Analysis*.

Analisi *Per protocol*

Ogni analisi dei dati ricavati da una sperimentazione dove viene studiato un "evento" parte dal confronto della frequenza dell'evento nel braccio di intervento con la frequenza dell'evento nel braccio di controllo. Nel modello di analisi "Per protocol" vengono esclusi in ciascun braccio dal calcolo della frequenza dell'evento i pazienti che per qualche motivo hanno violato il protocollo (*non compliant* e *cross-overs* – vedi -> analisi ITT). L'analisi "Per protocol" si basa pertanto sulla rile-

vazione dell'outcome nei soli pazienti che hanno seguito correttamente il trattamento assegnato. Questo tipo di analisi, anche se appare più ovvia al clinico rispetto al modello ITT, può portare in realtà a grave distorsione (bias) dei risultati del trial. Una volta eliminati dal calcolo i pazienti che hanno violato il protocollo i due bracci a confronto possono infatti essere fortemente sbilanciati nella distribuzione di fattori prognostici in grado di influenzare l'outcome. Per fare un esempio: se alcuni pazienti nel braccio di intervento hanno violato il protocollo a causa di effetti collaterali del farmaco (interrompendone l'assunzione), un'analisi per protocol, che rileva la frequenza dell'outcome solo nei pazienti che hanno assunto il farmaco può portare ad una grave sottostima degli effetti indesiderati del farmaco in quanto è una analisi eseguita sui pazienti più "resistenti" alle azioni negative del farmaco.

Intervalli di confidenza di una misura di efficacia

In una ricerca dove l'oggetto di studio è un "evento" gli intervalli di confidenza al 95% esprimono il grado di imprecisione con cui viene calcolata la misura che esprime la diversità riscontrata tra le frequenze dell'evento nei due bracci. Un intervallo di confidenza è rappresentato da un range di valori. Questo intervallo numerico esprime entro quali limiti cadrebbero tutte le stime della stessa misura di efficacia se la sperimentazione fosse ripetuta per cento volte e nelle stesse condizioni. Se vengono calcolati (come di norma si fa) gli "intervalli di confidenza al 95%" significa che ripetendo la sperimentazione per cento volte in 95 casi su 100 le stime della misura di efficacia cadrebbero entro quel range. Traducendo questi concetti in modo operativo e prendendo come esempio di misura di efficacia il Rischio Relativo (RR) calcolato per l'outcome primario dello studio di Ridker [RR = 0,91 (0,8-1,03)]: il valore di RR stimato dalla ricerca è 0,91 ma con una attendibilità del 95% il valore reale di questa stima (cioè quello della popolazione da cui il campione è stato estratto) è compreso tra il limite inferiore di 0,8 e il limite superiore di 1,03 (intervallo di confidenza al 95%). Se ripetessimo cento volte questa ricerca – ogni volta nelle stesse condizioni – ad ogni calcolo del Rischio relativo ogni stima sarebbe leggermente diversa dalle altre a causa delle differenze casuali esistenti tra i campioni estratti ma in 95 casi su 100 i valori di queste stime cadrebbero tra 0,8 e 1,03. Se l'intervallo di confidenza di un RR contiene – come in questo caso – il valore di 1 la differenza riscontrata tra i due bracci non può essere considerata significativa. Infatti se il valore di un RR è uguale a 1 significa che la frequenza dell'evento nel braccio di intervento è identica alla frequenza dell'evento nel braccio di controllo, fatto questo che soddisfa l'ipotesi nulla, cioè che l'intervento non sia né efficace né dannoso ma assolutamente neutro (vedi -> P). In base a queste note si intuisce come gli intervalli di confidenza oltre a esprimere il grado di imprecisione di una stima possano essere utilizzati per valutare la significatività statistica delle differenze tra i due bracci e in modo molto più intuitivo dei valori di P. Il lavoro di Ridker riporta – correttamente – per ogni calcolo di misura di efficacia sia gli intervalli di confidenza che il valore di P. Per esempio il valore di P riportato accanto al RR dell'outcome primario RR = 0,91 (0,8-1,03) corrisponde a P = 0,13 (significa: "la probabilità che la differenza riscontrata tra i due bracci sia solo dovuta al caso è pari al 13%": ciò ci autorizza ad affermare – coerentemente con l'intervallo di confidenza che contiene il valore di 1 – che il risultato "non è significativo" – vedi -> P).

Outcome (*end-point*, esiti)

In uno studio come quello di Ridker gli outcome sono costituiti da "eventi". Una prima classificazione degli outcome è quella che distingue "outcome maggiori" e "outcome surrogati". Un outcome "maggiore" è un *end-point* fortemente correlato allo stato di salute del paziente (es: mortalità). Un outcome "surrogato" si correla solo indirettamente con lo stato di salute ma è ben correlato con un outcome "maggiore". Esempio: l'outcome surrogato "valori di pressione arteriosa" è ben correlato con l'outcome maggiore "*stroke*" ma in sé non è in grado di influenzare direttamente la mortalità, che viene invece condizionata da uno *stroke*. Una seconda classificazione degli outcome distingue outcome "hard" e outcome "soft". Un outcome "hard" può essere misurato in modo facile e inequivocabile (es: mortalità); un outcome "soft" può essere misurato solo con difficoltà e introducendo valutazioni soggettive (esempio: QoL). Una terza classificazione degli outcome distingue outcome "primari" e outcome "secondari". L'outcome "primario" è quello su cui gli Autori della ricerca hanno tarato la potenza statistica dello studio (-> vedi) e quindi la numerosità del campione. Un outcome "secondario" non ha queste caratteristiche e le analisi ad esso correlate dovrebbero essere correttamente considerate solo come valore aggiunto e dovrebbero essere all'estremo utilizzate solo per generare ipotesi da affrontare in studi di numerosità campionaria adeguata. Una quarta classificazione degli outcome distingue "outcome compositi e outcome singoli". Un outcome "composito" è costituito dall'associazione di più outcome singoli. Un esempio di outcome composito è l'outcome primario considerato dalla ricerca di Ridker (mortalità cardiovascolare + infarto non fatale + *stroke* non fatale). Gli Autori dei trial utilizzano spesso come outcome "primario" un outcome composito: ciò per non essere costretti ad arruolare il numero più elevato di soggetti che richiederebbe lo studio di un outcome singolo. Un'analisi di un outcome singolo (a parità di potenza statistica) richiede di arruolare un numero più elevato di soggetti perché la frequenza con cui un outcome singolo compare nella popolazione è giocoforza molto inferiore a quella con cui compare un outcome composito che comprende anche

quell'outcome singolo. La differenza tra le frequenze di un outcome singolo rilevate nei due bracci può quindi essere molto piccola, con necessità di arruolare un numero molto elevato di soggetti per dimostrarla (vedi -> potenza statistica)

P e significatività statistica

Il valore di P esprime la probabilità che il risultato ottenuto dal confronto tra braccio di intervento e braccio di controllo sia solamente un effetto del caso. È solo per convenzione statistica che si attribuisce un significato particolare ai valori di P a seconda che siano inferiori o superiori al famoso "cut off": 0,05. Se la probabilità che il risultato ottenuto dal confronto tra i due bracci sia solo un effetto del caso è inferiore a 1/20, (vale a dire: $P < 0,05$) per convenzione statistica affermiamo che il risultato non è casuale (in quanto consideriamo questa probabilità 1/20 molto bassa). Se invece la probabilità che il risultato ottenuto dal confronto tra i due bracci sia solo casuale è superiore a 1/20 (vale a dire: $P > 0,05$) per convenzione statistica affermiamo che il risultato è, appunto, solamente dovuto al caso. Ogni test di significatività statistica consente il calcolo dei valori di P. Applicando un test statistico occorre procedere con una logica particolare: il confronto deve sempre partire dal presupposto chiamato "ipotesi nulla": ossia che non esista alcuna differenza non casuale tra i due bracci e che le differenze inevitabilmente registrate (è virtualmente impossibile che i due bracci siano assolutamente identici!) rappresentino appunto solo l'effetto della casualità con cui sono stati scelti i campioni. In uno studio di "eventi" l'"ipotesi nulla" identifica quindi una situazione estrema in cui la frequenza dell'evento nel braccio di intervento è identica alla frequenza dell'evento nel braccio di controllo. Un valore di $P < 0,05$ ci autorizza per convenzione statistica a ricusare l'ipotesi nulla: con $P < 0,05$ siamo cioè autorizzati ad affermare che il risultato ottenuto dal confronto tra i due bracci non è casuale e che esiste – cioè – "significatività" di questo risultato. Un valore di $P > 0,05$ ci costringe – all'opposto – ad abbracciare l'ipotesi nulla e ad affermare che la differenza riscontrata tra i due bracci è un fenomeno attribuibile solo alla casualità con cui abbiamo scelto i campioni.

Potenza statistica

La Potenza statistica di uno studio rappresenta la capacità di dimostrare differenze tra i due bracci quando queste effettivamente esistono. La potenza statistica di uno studio è direttamente proporzionale alla numerosità del campione arruolato (*sample size*). Più piccola è la differenza tra i due bracci che lo studio si propone di dimostrare, maggiore deve essere la numerosità dei soggetti da reclutare al fine di garantire un'adeguata potenza statistica. Per programmare *ex ante* una adeguato *sample size* il ricercatore deve quindi avere un'idea preliminare "di massima" della differenza tra un braccio e l'altro che crede di poter arrivare a dimostrare attraverso la propria ricerca. Il complementare a 100 della potenza statistica si chiama "errore beta". Lo studio di Rikter aveva una potenza statistica dell'83%: ne consegue che l'errore beta era (100-83) = 17%. L'errore beta è la probabilità di definire erroneamente "non significativa" una differenza tra i due bracci che invece effettivamente esiste. Si considerano di norma accettabili livelli di errore beta compresi tra il 10% e il 20%.

Qualità esterna

La "qualità esterna" di una ricerca coincide con il concetto di trasferibilità dei risultati della ricerca a popolazioni diverse da quella studiata. Uno studio – anche se metodologicamente impeccabile – è utile solo se i suoi risultati possono essere trasferiti a pazienti reali. I principali elementi in grado di condizionare la trasferibilità sono:

a) la somiglianza dei pazienti arruolati dallo studio con quelli normalmente visibili in condizioni reali (esempio: se i criteri di arruolamento adottati dal trial sono troppo rigidi la popolazione arruolata non è rappresentativa della realtà); a questo proposito un dato molto utile per valutare la somiglianza dei pazienti del trial con quelli reali è rappresentato dal Rischio Assoluto dell'outcome nel braccio di controllo (vedi -> RR), buon indicatore delle condizioni "basali" della popolazione arruolata: può essere confrontato con il Rischio Assoluto dell'outcome riscontrabile nei pazienti "reali";
b) la somiglianza dell'intervento studiato con quelli normalmente somministrati ai pazienti reali (in termini di: natura, dosaggio, modalità di somministrazione);
c) il tipo e la durata del follow-up (la compliance al trattamento dei pazienti reclutati nei trial è molto più alta di quella dei pazienti reali);
d) il tipo di *setting* in cui si svolge la ricerca (malattie affrontate in ospedale possono essere molto diverse da quelle affrontate sul territorio in termini di gravità clinica);
e) il tipo di outcome (un intervento su un outcome surrogato come – ad esempio la pressione arteriosa – non necessariamente è seguito da effetti favorevoli sulla morbilità e la mortalità dei pazienti reali).

Qualità interna

La "qualità interna" coincide con il concetto di validità metodologica. Uno studio di buona qualità è uno studio senza bias ossia i cui risultati non sono stati distorti da errori nella conduzione della ricerca. Gli elementi metodologici più importanti e critici di uno studio controllato come quello di Ridker sono:

a)un'adeguata numerosità del campione (*sample size*);
b) una buona "*allocation concealment*" ossia l'assegnazione dei pazienti ai due bracci deve essere eseguita utilizzando una procedura di randomizzazione validata e sicuramente in cieco (vedi);

c) un follow-up di lunghezza adeguata a rilevare l'outcome e caratterizzato da poche o pochissime perdite al follow-up;
d) la cecità nella somministrazione dell'intervento e nella rilevazione dei dati;
e) la scelta di outcome importanti e ben misurabili ossia "maggiori" e "hard" (vedi -> outcome).

Randomizzazione

In un RCT (vedi) l'assegnazione casuale dei pazienti ai due bracci prende il nome di "*allocation concealment*". Gli elementi qualificanti di una buona "allocation" sono:
a) la generazione dei numeri random attraverso metodi validati (tabelle; programmi informatici);
b) l'implementazione della randomizzazione (ossia: il metodo materiale con cui il singolo paziente viene allocato ad un braccio o ad un altro);
c) il mascheramento della randomizzazione, che in condizioni ideali deve essere in "doppio cieco".

La cecità identifica la "non conoscenza del braccio a cui il paziente è stato allocato". La randomizzazione è l'unico metodo in grado di distribuire equamente tra i due bracci tutti i fattori di rischio noti e non noti, rendendoli perfettamente confrontabili nelle condizioni di partenza. Se i due gruppi di pazienti a confronto sono uguali, una diversità nella frequenza dell'outcome riscontrata tra i due bracci potrà essere ragionevolmente attribuita ad un effetto dell'intervento studiato.

RCT = studio randomizzato e controllato (studio "sperimentale" propriamente detto)

Rappresenta il *golden standard* quando l'obiettivo della ricerca è verificare l'efficacia di un intervento sanitario. Nel modello più semplice esistono due gruppi di pazienti a confronto (braccio di intervento, braccio di controllo), composti da individui pressoché identici nelle caratteristiche di base. I pazienti del braccio di intervento ricevono l'intervento medico o chirurgico studiato dalla ricerca (esempio: un farmaco); i pazienti del braccio di controllo ricevono placebo o un intervento alternativo. In queste condizioni se dopo la somministrazione dell'intervento si apprezzano differenze tra i due bracci per l'outcome considerato dalla ricerca appare ragionevole attribuire queste differenze ad un effetto dell'intervento. La conditio *sine qua non* è che i pazienti dei due bracci siano quasi identici nelle condizioni di base: questa "identicità" è garantita dalla randomizzazione, che rappresenta l'unico metodo per distribuire in modo assolutamente uniforme tra i due bracci tutti i fattori prognostici noti e ignoti (vedi). In un disegno "fattoriale" come quello dello studio di Ridker i bracci a confronto sono più di uno (vedi testo).

RR = rischio relativo

Il Rischio Relativo esprime la frazione di rischio basale osservata dopo l'intervento. Ogni analisi dei dati ricavati da una sperimentazione dove viene studiato un "evento" parte dal confronto della frequenza dell'evento nel braccio di intervento con la frequenza dell'evento nel braccio di controllo. La frequenza di un evento in un braccio, espressa dal rapporto (numero di eventi osservati / totale dei pazienti appartenenti a quel braccio) si definisce "Rischio assoluto" dell'evento per quel braccio (AR). Pertanto il confronto tra i due bracci di uno studio è espresso dal confronto tra i due Rischi assoluti dell'outcome nei due bracci. Una modalità per esprimere matematicamente tale confronto è il calcolo del rapporto tra i due Rischi assoluti, che si chiama Rischio Relativo (RR). Quindi: RR = (AR intervento / AR controllo). Il Rischio Relativo esprime come detto la frazione di rischio basale osservata dopo l'intervento. Quale è allora il Rischio basale? Il Rischio basale, ossia quello che si osserverebbe se non venisse applicato l'intervento, coincide con il Rischio assoluto del braccio di controllo (AR controlli). Per fare un esempio il Rischio Relativo per l'outcome primario nel lavoro di Ridker corrisponde a RR = 0,91. Significa che nei pazienti trattati con ASA è stato osservato un rischio pari al 91% del rischio che si sarebbe osservato non trattandoli con ASA. Quest'ultimo (= Rischio basale) coincide come detto con il Rischio dei controlli, a cui appunto non è stato somministrato ASA. I due Rischi Assoluti di outcome primario nei due bracci dello studio si calcolano così:

- per il braccio di intervento: 477 eventi in 19.934 pazienti = 477/19934 = 0,02393;
- per il braccio di controllo: 522 eventi in 19.942 pazienti = 522/19942 = 0,02618.

Il rischio relativo RR è allora espresso da 0,02393 / 0,02618 = 0,91416.
Significa che applicando l'intervento osserviamo un rischio di outcome primario (arrotondando: 2,3% in 10 anni) pari al 91% del rischio che avremmo osservato non applicando l'intervento (arrotondando: 2,6% in 10 anni).

RRR = riduzione di rischio relativo

È il valore complementare a 1 del Rischio Relativo. Per esempio se RR = 0,8 RRR = (1-0,8) = 0,2. Clinicamente esprime la frazione di rischio basale abbattuta dall'intervento. Tornando allo studio sull'ASA abbiamo visto che il RR per l'outcome primario è pari a 0,91. Allora la Riduzione di Rischio Relativo per l'outcome primario sarà pari a RRR = (1-0,91) = 0,09. Significa che la somministrazione di ASA abbatte il rischio basale del 9%.

Sensitivity Analysis

Ogni analisi dei dati ricavati da una ricerca dove viene studiato un "evento" parte dal confronto delle frequenze dell'evento calcolate in due gruppi di pazienti. Una *Sensitivity analysis* rappresenta un modello generale di analisi dove questi calcoli vengono rifatti dopo aver escluso o incluso gruppi pazienti con ben determinate caratteristiche: tutto ciò allo scopo di saggiare la robustezza delle precedenti conclusioni attraverso il

confronto di scenari (immaginari) estremi. Un modello di analisi del genere viene spesso sfruttato in presenza di perdite al follow-up, ossia di pazienti di cui non si conosce l'esito. In questo caso vengono rifatti i calcoli per quattro volte immaginando quattro scenari estremi:

a) primo scenario: tutti i pazienti persi non hanno avuto l'evento (sia nel braccio di intervento che nel braccio di controllo);
b) secondo scenario: tutti i pazienti persi hanno avuto l'evento (sia nel braccio di intervento che nel braccio di controllo);
c) terzo scenario: i pazienti persi nel braccio di intervento hanno avuto l'evento e i pazienti persi nel braccio di controllo non hanno avuto l'evento;
d) quarto scenario: i pazienti persi nel braccio di controllo hanno avuto l'evento e i pazienti persi nel braccio di intervento non hanno avuto l'evento.

Queste analisi possono essere eseguite entro l'ambito di una analisi ITT o entro l'ambito di un'analisi per Protocol. Questo modello generale di analisi viene utilizzato nel lavoro di Ridker non per considerare le perdite al follow-up (modeste e trascurabili) ma per escludere dei calcoli i pazienti *non compliant*: come spiegato nel testo ciò equivale ad eseguire un'analisi "Per protocol"(-> vedi).

Subgroups analysis

Una analisi per sottogruppi prevede confronti tra braccio di intervento e braccio di controllo eseguiti non sull'intera casistica ma su singole sottopopolazioni di pazienti con caratteristiche particolari. Le insidie di una analisi per sottogruppi sono:

a) maggiore è il numero di sottogruppi studiati, maggiore è il rischio che i risultati di questi confronti siano solo un effetto del caso;
b) un sottogruppo di solito presenta dimensioni campionarie insufficienti ad una adeguata potenza statistica e questo espone l'analisi ad un grande rischio di errore beta (vedi Potenza statistica).

L'affidabilità di questo tipo di analisi aumenta se:

a) esiste grande plausibilità biologica dei suoi risultati;
b) la differenza tra braccio di intervento e braccio di controllo è grande;
c) i risultati sono statisticamente significativi;
d) gli Autori non l'hanno eseguita *ex post* (cioè dopo aver preso visione dei principali risultati dello studio) ma l'hanno prevista *ex ante* e ciò risulta specificato nel protocollo della ricerca pubblicato prima dell'inizio dello studio;
e) i sottogruppi e gli outcome considerati sono pochi.

Un'analisi per sottogruppi dovrebbe essere utilizzata, a rigore, solo per generare ipotesi di lavoro.