METODOLOGIA STATISTICA

(per gentile concessione del gruppo Linee Guida ISO-SPREAD)

GLOSSARIO

Viene introdotto in queste linee guida un glossario sintetico dei termini di statistica medica utilizzati – o necessari per definire quelli utilizzati – in queste linee guide. Questo glossario non è e non intende essere un "manuale" di statistica: a questo scopo sono disponibili numerosi volumi e, in ogni caso, si raccomanda di seguire un corso specialistico nel caso si fosse interessati. Si intende invece mettere a disposizione del lettore una breve spiegazione dei termini che ricorrono nel testo, in che contesto sono utilizzati negli studi clinici che stanno alla base delle linee guida, indicando anche come hanno influenzato il peso e/o la rilevanza dati agli studi stessi in relazione all'applicazione dei risultati alla pratica, quindi nella costituzione delle raccomandazioni di queste linee guida. Per approfondire gli aspetti più direttamente legati all'analisi della letteratura, si possono suggerire: Primer of Biostatistics 6ª Ed, Glantz SA, McGraw-Hill Professional, 2005; Systematic Reviews in Health Care, Egger M, Smith GD, Altman DG Eds, BMJ Books, 2001; Evaluating drug literature. A statistical approach Slaughter RL, Edwards DJ Eds, McGraw-Hill, 2001; Petitti DB, Meta-analysis, decision analysis and cost-effectiveness analysis. Oxford, 1994; Clinical Epidemiology: A Basic Science for Clinical Medicine 2ª Ed., Sackett DL, Haynes RB, Tugwell P, Guyatt GH, Lippincott Williams & Wilkins, 1991. È anche possibile riferirsi a siti come **http://www.cebm.net/?o=1023**; **http://www.clinicalevidence.com/ceweb/resources/index.jsp**.

Questo glossario è coerente con i termini usati da BMJ Clinical Evidence. Molti esempi sull'utilizzo corretto e scorretto dei termini del glossario sono reperibili nel testo di Egger, Smith e Altman citato all'inizio).

Analisi di regressione	In presenza di dati da una variabile dipendente e da una o più variabili indipendenti, l'analisi di regressione implica la ricerca del migliore modello matematico per descrivere o predire il valore della variabile dipendente in funzione dei valori della o delle variabili indipendenti. Esistono diversi modelli di regressione, adeguati per diverse esigenze. Tra le forme più comunemente usate vi sono la regressione lineare, logistica (nella quale la variabile dipendente è nominale) e del rischio proporzionale (come l'analisi di sopravvivenza secondo Cox).
Analisi di sensibilità	Analisi per valutare se i risultati di una metanalisi sono sensibili all'applicazione di restrizioni ai dati inclusi. Tipiche analisi di sensibilità includono la stima dell'effetto con l'inclusione solo di studi di grandi dimensioni, solo di studi di elevata qualità, solo degli studi più recenti. L'analisi di sensibilità può essere eseguita anche per singoli studi clinici (purché di dimensioni adeguate) stimando l'esito, per esempio, utilizzando diverse tecniche di sostituzione dei dati mancanti nell'analisi ITT (*last value carried forwad, worst-case scenario*). Risultati consistenti tra loro rafforzano l'evidenza e la **generalizzabilità** dei risultati osservati.
Analisi di sottogruppi	Analisi di una parte soltanto della popolazione di uno studio/metanalisi nella quale si ipotizza che l'effetto possa essere differente rispetto all'effetto globale osservato. L'analisi per sottogruppi deve sempre essere indicata come tale e preferibilmente pre-specificata nel protocollo. Nel caso di studi clinici deve anche essere adattata la numerosità campionaria. L'analisi di sottogruppi può essere utile per generare nuove ipotesi da verificare. Solo raramente e soggetti a stretti criteri di validità statistica un'analisi di sottogruppi può essere considerata fonte di evidenza.

Analisi intention to treat (ITT)	Analisi dei dati di tutti i partecipanti a uno studio indipendentemente dal rispetto dei criteri di inclusione/esclusione, dall'aderenza al trattamento e dal completamento dello studio. Esistono diverse definizioni non tutte ugualmente ragionevoli e immediate del principio ITT, che possono originare dataset differenti per le diverse analisi. Tra le più comuni sono: inclusione di tutti i casi randomizzati assegnati al gruppo di trattamento cui erano stati randomizzati indipendentemente dal trattamento effettivamente ricevuto (dataset differenti per efficacia e per sicurezza; in quest'ultimo i soggetti sono assegnati al trattamento effettivamente ricevuto); inclusione di tutti i soggetti randomizzati assegnati al gruppo del trattamento effettivamente ricevuto (i dataset per efficacia e sicurezza sono gli stessi); esclusione dei soggetti che hanno fisicamente rifiutato il trattamento prima della prima somministrazione/assunzione/applicazione. Altre definizioni sono pure usate ma sono meno accettabili. In ogni caso l'esatta definizione di ITT deve essere chiaramente presentata. L'analisi dei dati ITT è in ogni caso la più probante. Tuttavia l'implicazione delle tecniche di sostituzione dei dati mancanti, che si rende indispensabile per l'analisi ITT, può introdurre un *bias* e va considerata caso per caso. È infatti ben diverso trascinare copiare nelle misure successive l'ultimo valore misurato – tecnica "last value carried forward" – quando l'evoluzione spontanea è in miglioramento (tecnica conservativa che non favorisce indebitamente il gruppo con più casi persi all'osservazione) o fare lo stesso quando l'evoluzione spontanea è in peggioramento (in tal caso il gruppo con più casi persi all'osservazione è indebitamente favorito). Le tecniche di sostituzione e le ragioni per la scelta effettuata dovrebbero essere sempre indicate esplicitamente. A volte viene associata anche un'analisi di sensibilità valutando gli esiti stimati utilizzando diverse tecniche di sostituzione.
Analisi *per-protocol*	Analisi dei dati solo per quei partecipanti allo studio che hanno completato lo studio secondo il protocollo (cioè senza deviazioni importanti dai criteri di inclusione e di esclusione e dai criteri di aderenza alla prescrizione). Una forma meno restrittiva dell'analisi *per-protocol* è l'analisi dei dati di tutti i partecipanti rimasti alla fine dello studio ("completers"). Entrambe queste analisi sono fortemente soggette a *bias* e dovrebbero essere evitate se non in casi assolutamente particolari (p.es. studio di fase II di relazione dose-effetto).
Applicabilità	L'applicazione dei risultati da studi clinici a singoli pazienti. Uno studio clinico fornisce evidenza diretta di causalità entro quello specifico studio. È necessario un ulteriore passo logico per applicare tale risultato a uno specifico soggetto. Le caratteristiche individuali influenzeranno certamente l'esito per quel soggetto. Quando i risultati sono espressi come **NNTB** e **NNTH**, è possibile stimare su base clinica lo scostamento del soggetto dal profilo medio dei soggetti esaminati nello studio (o nelle metanalisi). Tale scostamento andrà da 1 (completa sovrapponibilità) a 0 (completo scostamento). Il prodotto di **NNTB** e **NNTH** per il grado di scostamento stimato può permettere di valutare caso per caso l'opportunità dell'intervento. Ciò è solo possibile se le caratteristiche dei soggetti reclutati per uno studio sono ben descritte per ciascuno dei due gruppi di intervento.
Aumento assoluto di rischio (ARI; *absolute risk increase*)	La differenza assoluta tra il rischio nel gruppo sperimentale e quella nel gruppo di controllo in uno studio. Si utilizza quando il rischio nel gruppo sperimentale supera il rischio nel gruppo di controllo e si calcola sottraendo il rischio assoluto nel gruppo di controllo dal rischio assoluto nel gruppo sperimentale. Questo indice non dà nessuna informazione sulla proporzione di aumento del rischio fra i due gruppi. Per questa informazione si utilizza il rischio relativo (RR).
Aumento relativo di rischio (RRI; *relative risk increase*)	L'aumento proporzionale di rischio tra i partecipanti assegnati al gruppo di sperimentale e quelli assegnati al gruppo di controllo in uno studio.
***Bias* (errore sistematico)**	Deviazione sistematica dei risultati di uno studio dai risultati veri a causa del modo nel quale è stato condotto lo studio. In ogni studio o revisione sistematica dovrebbero essere discusse le potenziali fonti di *bias* e le misure prese per evitarne/limitarne l'effetto.
***Bias* di pubblicazione**	Probabilità che uno studio sia pubblicato in funzione dei risultati trovati. Studi con risultati statisticamente significativi, finanziati dall'esterno o su campioni più grandi hanno una maggiore probabilità di essere pubblicati rispetto a studi spontanei o con risultati non statisticamente significativi o su campioni più piccoli. Il *bias* di pubblicazione porta a sovrastimare l'effetto di un intervento/trattamento (e sottostimarne i rischi).

Blocchi di randomizzazione (o randomizzazione in blocchi)	Procedura di randomizzazione tramite la quale ogni dato numero di soggetti sono presenti trattati e controlli nella proporzione prevista dal protocollo di studio (generalmente 1:1). Serve a evitare che i gruppi di trattamento siano sbilanciati negli studi multicentrici. Può introdurre un *bias* di randomizzazione se chi recluta i pazienti è a conoscenza della dimensione del blocco, che quindi non dovrebbe essere indicata in protocollo. Per ridurre il rischio di *bias* è anche preferibile che i blocchi siano quanto più ampi possibile compatibilmente con le necessità pratiche (comunque mai meno di 4 elementi) e siano variabili (cioè non tutti della stessa lunghezza) nell'ambito dello stesso studio.
Chi quadrato (χ^2)	Tecnica di valutazione statistica della distribuzione di esiti in forma nominale fra trattamenti. L'obiettivo è stimare quanto è probabile che la distribuzione osservata possa verificarsi per puro caso se la distribuzione degli esiti è indipendente dalla distribuzione dei trattamenti. Se la probabilità è piccola (<5%) si può respingere l'ipotesi di indipendenza tra esiti e trattamenti. Il test chi quadrato è alla base di tutte le stime di rischio. Respingere l'ipotesi di indipendenza non significa, tuttavia, affermare un rapporto causa-effetto.
Cieco/cecità	Procedura tramite la quale i partecipanti ad uno studio clinico non sono a conoscenza del trattamento/procedura cui i singoli soggetti sono stati allocati. I partecipanti includono i soggetti dello studio, i ricercatori, i responsabili della raccolta ed elaborazione dei dati. La cecità è di particolare rilevanza per la determinazione dell'esito del trattamento/procedura e per la definizione dei soggetti da escludere dall'analisi statistica. Esistono termini convenzionali come singolo cieco e doppio cieco, tuttavia non sono utilizzati sempre in maniera coerente. È preferibile che sia indicato specificamente chi non è a conoscenza del trattamento/procedura assegnato.
Cecità rispetto all'allocazione	Metodo utilizzato per evitare il *bias* di selezione. Consiste nel rendere inaccessibile la sequenza di allocazione dei trattamenti a chi assegna i partecipanti ai gruppo di interventi. La cecità rispetto all'allocazione evita che i ricercatori possano influenzare (inconsciamente o volutamente) a quale gruppo di intervento venga allocato ciascun partecipante. Gli aspetti che possono rendere inutile tale tecnica includono tutte quelle deviazioni che possono permettere al ricercatore di prevedere a che gruppo di intervento debba essere assegnato il paziente al momento dell'allocazione (sequenze di allocazione prevedibili, buste non opache, assegnazione in aperto).
Concordanza, accordo (κ)	La misura κ di concordanza secondo Cohen è una misura generalmente robusta per stimare l'accordo tra diversi stimatori di esiti dicotomici. Il test "depura" la concordanza osservata dalla quota di concordanza comunque prevedibile per effetto casuale (che si ottiene, come per il chi quadrato, dalla stima della distribuzione indipendente delle frequenze marginali della tabella). Normalmente il valore di k varia tra 0 e 1 (può anche assumere valori negativi fino a -1). Un accordo <0,60 (quota di accordo residuo non casuale giustificata dalla concordanza vera <60%) non è più che moderato. Si considera buono un accordo con valori di k almeno superiori al 60%~75%.
Consistenza (talvolta: coerenza)	Risultati ottenuti da studi diversi e indipendenti che affrontano la stessa domanda clinica. I livelli di evidenza sono tanto più elevati quanto maggiore è la consistenza dei risultati. Non costituisce consistenza la ripetizione dello stesso studio in campioni non indipendenti.
Coefficiente di correlazione	Una misura di associazione che indica in che grado due variabili variano insieme in relazione lineare (o che può essere resa lineare con trasformazioni matematiche anche complesse). Viene rappresentato con la lettera r e varia tra -1 e +1. Quando vale +1 indica relazione positiva perfetta (quando una variabile aumenta, aumenta anche l'altra e la differenza rimane proporzionalmente costante). Quando vale -1 indica relazione negativa perfetta (quando una variabile aumenta, l'altra diminuisce e la differenza rimane proporzionalmente costante). Quando vale 0 indica assenza di correlazione, ma ciò non esclude una correlazione, esclude esclusivamente che tale correlazione abbia la forma considerata dall'equazione in studio.
Controlli	In uno studio randomizzato e controllato, il termine "controlli" si riferisce ai partecipanti allogati nel gruppo di riferimento. Il gruppo di riferimento è allocato al placebo, a nessun trattamento specifico, o a un trattamento considerato di riferimento (*standard of care, gold standard*).

Dimensione dell'effetto (differenza media standardizzata)	Il termine "dimensione dell'effetto" nella letteratura medica viene utilizzato con diverse accezioni. Una delle definizioni più classiche è quella di Cohen – utilizzata soprattutto per il calcolo del campione – ed è costituita da Δ/σ dove Δ indica l'entità delle differenza attesa tra i due gruppi in relazione alla misura dell'*endpoint* primario dello studio (misurato su variabile continua o assimilabile) e σ indica la deviazione standard cumulata di tale stima. Cohen definisce "piccolo" un effetto con Δ/σ = 0,2; "medio" se pari a 0,5; "grande" se pari o superiore a 0,8. Per estensione si può utilizzare tale definizione anche per combinare variabili analoghe misurate in studi diversi su scale diverse, utilizzando la procedura di "standardizzazione": Con tale procedura le differenza media di effetto osservata in ciascuno studio viene divisa per la deviazione standard della differenza stessa. Ciò permette di valutare una entità di effetto complessiva su tutti gli studi. Tuttavia, anche se tale procedura è corretta sul piano statistico, il risultato è di difficile interpretazione per chi non sia statistico di professione e migliora l'accuratezza statistica a spese dell'intelligibilità clinica. Pertanto è preferibile evitare di trarre conclusioni basate su tale procedura.
Effetti casuali	Modello di metanalisi. Il modello per effetti casuali assume che l'effetto in ogni studio sia diverso e prende in considerazione questa variabilità come fonte aggiuntiva di variazione dell'esito. Ciò genera intervalli di confidenza intorno all'effetto stimato un po' più ampi rispetto al modello per effetti fissi in quando gli effetti sono ipotizzati come distribuiti casualmente intorno all'effetto globale. Il vantaggio del modello per effetti casuali è che riduce l'impatto dell'eterogeneità fra studi (in assenza di eterogeneità i risultati dei due modelli sono praticamente coincidenti). Tuttavia il modello per effetti casuali, se pur li riduce, non rimuove gli effetti dell'eterogeneità.
Effetti fissi	Il modello di metanalisi ad "effetti fissi" assume, non sempre obiettivamente, che la variabilità tra studi dipende esclusivamente da variazione casuale di campionamento intorno a un effetto fisso. Si contrappone a "effetti casuali".
Esito (*outcome*)	Il termine "esito" si presta ad alcune ambiguità. La definizione stretta di "esito" lo indica come l'evento clinico misurabile tramite il quale si valuta se l'obiettivo dello studio è stato raggiunto o meno. Quindi qualsiasi evento clinico di qualsiasi natura può costituire esito di uno studio clinico. Tuttavia non tutti gli esiti sono ugualmente esenti da *bias* né sono tutti espressione diretta della domanda clinica posta. Un esito diretto è sempre preferibile a un esito surrogato (p.e., la riduzione di pressione arteriosa è surrogato della riduzione di rischio di eventi cardiovascolari, ma l'esito diretto è il verificarsi degli eventi cardiovascolari). Esiti direttamente rilevanti per i pazienti costituiscono maggiore evidenza rispetto a esiti surrogati. Esiti espressi come eventi poco soggetti a *bias* costituiscono evidenza di grado superiore rispetto a esiti presumibilmente soggetti a *bias* o con *bias* difficilmente valutabile (p.e., vivo/morto rispetto a pressione alta/ pressione normale).
Eterogeneità	Nel contesto di una metanalisi, il termine "eterogeneità" indica dissimilarità fra studi. Può essere dovuta a cause di tipo matematico-statistico (eterogeneità statistica) o alla valutazione di soggetti con diverse caratteristiche o diversi trattamenti o diversi esiti o esiti valutati in modo diverso (eterogeneità clinica). Una elevata eterogeneità può rendere poco appropriata o affidabile la metanalisi di dati accorpati in una revisione sistematica. L'assenza di evidenza di eterogeneità (statistica) non è la stessa cosa che trovare evidenza che non vi è eterogeneità. L'impiego dell'indice I^2, che stima l'entità dell'eterogeneità fra studi indipendentemente dalla significatività statistica dell'eterogeneità stessa, può aiutare a discriminare tra eterogeneità che possono influenzare le conclusioni sull'esito ed eterogeneità di scarso rilievo.
Evento	Il verificarsi di un esito dicotomico cercato nello studio (p.es. il verificarsi di decesso, o di ictus, o un peggioramento di almeno 4 punti alla scala NIHSS).
Falso negativo	Un soggetto che presenta la condizione bersaglio secondo il gold standard ma fornisce un risultato negativo al test sotto studio.
Falso positivo	Un soggetto che non presenta la condizione bersaglio secondo il gold standard ma fornisce un risultato positivo al test sotto studio.
Fattoriale (disegno)	Un disegno fattoriale ha lo scopo di valutare più di un intervento in confronto con un controllo in un unico studio, attraverso tecniche di randomizzazione multipla.
Incidenza	Numero di nuovi casi di una condizione specificata che si verificano in una popolazione su un arco di tempo specificato.

Inclusione/esclusione (criteri)	Criteri in base al quale viene ammessa una unità di valutazione in una ricerca. Se riferito a un singolo studio indica i criteri demografici e clinici ai quali devono sottostare i soggetti che possono essere trattati nello studio. Se riferito a metanalisi o a revisione sistematiche indica i criteri minimi di accettabilità per considerare correttamente valutabile uno studio clinico. Non necessariamente i criteri di inclusione ed esclusione sono tutti basati su motivazioni obiettive (validi). Criteri non validi possono ridurre l'**applicabilità** e la **generalizzabilità** degli studi e delle metanalisi.
Intervallo di confidenza (IC; CI)	L'intervallo di confidenza è costituito da due limiti intorno al risultato osservato, generalmente definiti con una probabilità del 95% (a volte del 90%). L'intervallo di confidenza al 95% – o i limiti fiduciali al 95% – include il 95% dei risultati degli studi della stessa dimensione e con lo stesso disegno sperimentale condotti su campioni indipendenti estratti dalla stessa popolazione. È una dizione analoga, ma non identica, rispetto a dire che l'entità vera dell'effetto (che non è mai conosciuta esattamente) ha il 95% di probabilità di cadere entro l'intervallo di confidenza al 95%. Se l'intervallo di confidenza al 95% di un rischio relativo (RR) o di una *odds ratio* (OR) contiene il valore 1, si considera che non vi sia evidenza di effetto. Il vantaggio pratico dell'intervallo di confidenza rispetto al valore di P è che tale intervallo fornisce l'ampiezza degli effetti probabili. Un effetto espresso senza i suoi intervalli di confidenza ha un livello di evidenza praticamente nullo, indipendentemente dall'eventuale valore di P associato.
Metanalisi	Tecnica statistica che riassume i risultati di diversi studi in una singola stima pesata, nella quale il peso maggiore viene attribuito ai risultati di studi con più eventi e talvolta a studi di migliore qualità. L'esecuzione di una metanalisi presuppone l'estrazione dalla letteratura di tutti gli studi pertinenti alla domanda clinica posta (revisione sistematica) e la loro accettazione per la metanalisi secondo criteri di inclusione ed esclusione predefiniti. Sebbene metanalisi e revisione sistematica non siano sovrapponibili – infatti l'esecuzione di una revisione sistematica non implica necessariamente una metanalisi dei risultati – una buona revisione sistematica è presupposto indispensabile per una metanalisi valida e generalizzabile. Normalmente i risultati sono espressi come OR o RR o ARR basate su eventi di natura dicotomica; a volte è possibile valutare in metanalisi anche esiti di tipo quantitativo ma i risultati sono meno affidabili e di più difficile interpretazione. Si noti che non implica l'accorpamento dei dati singoli che si basa esclusivamente sui dati grezzi e non prevede pesature. L'analisi accorpata di dati singoli da studi diversi è poco affidabile per trarre conclusioni generali. È in discussione ma senza una conclusione universalmente accettata la "metanalisi di dati singoli". In tale tecnica si analizzano i dati grezzi provenienti da vari studi, separatamente per ciascuno studio, e questo risultato intermedio – appropriatamente pesato – viene utilizzato per la metanalisi finale. Questa tecnica è utile ad affrontare domande cliniche non affrontate o non adeguatamente esaminate negli studi originali ma cui si può rispondere a partire dai dati raccolti. Si pone però il problema della effettiva generalizzabilità poiché i dati singoli possono essere disponibili per solo pochi degli studi condotti sul tema d'interesse.
Morbosità (detta anche "morbilità")	Tasso di una patologia non letale.
Mortalità	Tasso di morte.
Non significativo/non statisticamente significativo (NS)	Convenzionalmente si definisce come non significativo (più precisamente "non statisticamente significativo") un effetto osservato quando un effetto della stessa entità o di entità maggiore può verificarsi in più di 1 studio su 20 (5%) per puro effetto del caso, assumendo che non vi siano differenze vere di esito fra i trattamenti/interventi esaminati. Non è la stessa cosa che affermare che non c'è un effetto. Il termine "non significativo" indica semplicemente che lo studio non ha fornito una evidenza convincente dell'esistenza di un effetto. Ciò può avvenire perché lo studio non era dimensionato in modo tale da rilevare un effetto che in realtà esiste, o perché davvero non esiste nessun effetto, o per effetto del caso. Una differenza non significativa non è lo stesso che indicare una tendenza ("trend") non significativa. Tendenza ed effetto non sono sinonimi. Possono comunque verificarsi casi in cui una differenza potenzialmente clinicamente rilevante non è allo stesso tempo statisticamente significativa. Tale evenienza è comunque meritevole di essere evidenziata, sempre indicando chiaramente l'assenza di significatività statistica.
***Number needed to harm* (NNTH; NNH)**	Una misura dei potenziali danni di un trattamento/intervento. Corrisponde al numero medio di soggetti appartenenti a una popolazione ben definita che dovrebbero essere trattati con uno specifico intervento e per uno specifico periodo di tempo per causare un esito sfavorevole in più rispetto ai controlli. Si può calcolare da 1/ARI (aumento assoluto di rischio).

***Number needed to treat* (NNT)/*Number needed to benefit* (NNTB)**	Una misura dell'efficacia di un trattamento/intervento. Corrisponde al numero medio di soggetti che devono essere trattati con uno specifico intervento per uno specifico periodo di tempo per prevenire un evento sfavorevole in più, o per ottenere un esito favorevole in più, rispetto al gruppo di controllo. Può essere calcolato come 1/ARR (riduzione assoluta di rischio). Lo NNTB è facile da interpretare e facile da confrontare fra interventi diversi, tuttavia fa riferimento un uno specifico livello di rischio basale. Lo stesso intervento con lo stesso esito in popolazioni con rischio basale differente produce NNTB differenti. Se è stata eseguita una metanalisi (con attribuzione dei pesi agli studi considerati) il calcolo di NNTB è più accurato se eseguito calcolando il rischio assoluto per il gruppo in esame non direttamente come AR, ma come OR (considerata con il suo intervallo di confidenza al 95%) moltiplicata per il rischio assoluto nel gruppo di controllo. In tal caso si tiene conto del peso (anche per la qualità degli studi) assegnato agli studi. È ragionevole e utile esprimere i risultati di un confronto in termini di NNTB solo se l'effetto è statisticamente significativo (vale a dire, se entrambi i limiti dell'intervallo di confidenza sono valori positivi). L'interpretatone di un intervallo di confidenza per un NNTB non significativo diventa piuttosto difficile. In tal caso infatti uno dei due limiti dell'intervallo ha valore negativo. Ciò equivale a dire che si passa da un NNTB a un **NNTH** ma il punto di passaggio dall'uno all'altro corrisponde – per il modo in cui si calcola il valore – a infinito, non a zero.
Odds	L'*odds* di un evento viene definita come la probabilità che avvenga un evento definito, espressa come proporzione della probabilità che l'evento non avvenga. Si calcola come rapporto tra soggetti con evento e soggetti senza evento.
***Odds ratio* (OR)**	Una misura di efficacia di un trattamento/intervento. È il rapporto tra la *odds* che avvenga un evento nel gruppo sperimentale e la *odds* che avvenga lo stesso evento nel gruppo di controllo. Si ottiene dividendo il rapporto tra eventi e non eventi nel gruppo in esame per il rapporto tra eventi e non eventi nel gruppo di controllo. Quanto più prossima al valore 1, tanto più piccola la differenza tra l'effetto del trattamento/ intervento in esame e l'effetto del trattamento/intervento di controllo. Una OR maggiore di 1 indica che l'effetto del trattamento/intervento in esame è maggiore rispetto al trattamento/intervento di controllo; se minore di 1 significa il contrario. La OR ha significato solo se espressa con il suo intervallo di confidenza al 95%; se questo comprende il valore 1 non vi è evidenza di differenze di effetto. Per pochi eventi OR e rischio relativo (RR) tendono a coincidere; per molti eventi i due indici divergono rapidamente. L'uso della OR dovrebbe essere limitato a studi retrospettivi o caso-controllo e a metanalisi. In studi prospettici è più accurato esaminare la differenza assoluta di rischio o, al più, il rischio relativo. A volte la *odds ratio* viene citata nei documenti di origine USA come *rate ratio*. Quest'ultimo termine viene evitato in ambito europeo perché facilmente fonte di confusione.
Omogeneità	Similarità. Termine poco usato perché impreciso. In metanalisi, si considerano omogenei (in realtà poco eterogenei) studi con indice I^2 <30%. Quando riferito a gruppi di soggetti (tipicamente nella dizione "gruppi omogenei" in uno studio randomizzato e controllato) indica semplicemente che non vi è evidenza di scostamenti statisticamente significativi tra i gruppi a confronto. Se si riferisce anche all'assenza di evidenti scostamenti sotto il profilo clinico va considerato come termine esclusivamente qualitativo dipendente da considerazioni proprie dello sperimentatore (o del *Writing Committee*) non basate su valutazioni obiettive, se non specificamente riportate.
P (valore)	Probabilità che si osservi una differenza uguale o maggiore per puro effetto del caso, nell'ipotesi che non via sia una reale differenza di effetto tra gli interventi confrontati. Convenzionalmente se tale probabilità è <5% (1/20; P<0,05) si considera il risultato come "statisticamente significativo".
Placebo	Sostanza somministrata al gruppo di controllo di uno studio clinico, teoricamente identica per caratteristiche organolettiche (aspetto, odore, sapore, peso) al trattamento sperimentale e ritenuta non esercitare nessun effetto sulla patologia in esame. Somministrare un placebo non è lo stesso che non somministrare nessun trattamento e a volte può indurre modificazioni fisiologiche misurabili. Se sia più appropriato non dare nulla o dare un placebo dipende dal disegno sperimentale dello studio, dalla specifica domanda clinica cui si intende rispondere e dalla misura dell'esito. La somministrazione di placebo a pazienti con una patologia definita pone anche problemi di natura etica da valutarsi in sede di stesura del protocollo ma anche di valutazione del peso metodologico dello studio.
Potenza	Si definisce "potenza" la probabilità di uno studio di poter osservare una differenza di esito fra due o più trattamenti/interventi della entità predefinita nel protocollo di studio con il livello di confidenza predefinito nello stesso protocollo, nel caso tale differenza esista. A parità di tutti gli altri elementi, la potenza aumenta con la dimensione del campione e con la precisione della misura dell'esito.

Prevalenza	Proporzione di soggetti che presenta uno specifico fenomeno o patologia in una popolazione definita a un momento temporale definito.
Protocollo	Documento che definisce in tutti i dettagli l'esecuzione di uno studio. I protocollo di ricerca vanno predisposti per tutti i tipi di studio clinico (intervenzionale, osservazionale, prospettico, retrospettivo, trasversale) ma anche per le revisioni sistematiche e le metanalisi. I punti chiave del protocollo includono la domanda clinica cui si intende rispondere, i criteri di inclusione e di esclusione, l'esito da considerare, la misura dell'esito, le tecniche statistiche da applicare. Per gli studi clinici vanno indicate tutte le procedure da applicare e la loro scadenza temporale, nonché le procedure di informazione dei soggetti e quelle per raccogliere e documentare il consenso alla partecipazione.
Randomizzazione a cascata ("cluster randomisation")	Il disegno sperimentale con randomizzazione a cascata prevede che gruppi di partecipanti siano randomizzati insieme allo stesso intervento. Esempi di randomizzazione a cascata includo l'allocazione di tutti i soggetti della stessa località, ospedale, scuola, medico generale, allo stesso intervento. In sede di analisi l'unità da analizzare deve essere la stessa dell'unità di randomizzazione. Se i dati sono analizzati utilizzando i singoli soggetti come unità si genera certamente un errore sistematico. Spesso il disegno per randomizzazione a cascata viene usato per rispondere a domande diverse rispetto alla randomizzazione individuale, anche se l'intervento è lo stesso. Per esempio, uno studio di prevenzione primaria con randomizzazione individuale può rispondere alla domanda sull'efficacia dell'intervento, mentre lo stesso studio con randomizzazione a grappolo per medico generale risponde alla domanda su quanto e con che forza viene recepita l'utilità dell'intervento dal medico.
Rapporto di tasso di rischio (HR; *hazard ratio*)	Grossolanamente equivalente al rischio relativo (RR) è una misura utile quando il tasso di rischio (*hazard rate*) non è costante nel tempo. Utilizzando informazioni raccolte su un arco di tempo sufficientemente lungo, è una misura sufficientemente accurata del rischio corso nel contesto di analisi di sopravvivenza. Diversamente dall'accezione comune lo HR può essere usato per eventi negativi e positivi. Il significato numerico non è diverso dalla *odds ratio* (in effetti i due termini possono essere derivati uno dall'altro).
Rapporto di verosimiglianza (*likelihood ratio*)	Il rapporto tra la probabilità che un soggetto con la condizione in esame presenti un risultato specifico a un test e la probabilità che un soggetto senza la condizione in esame presenti lo stesso risultato allo stesso test. Da questa definizione ristretta e usata in ambito diagnostico il rapporto di verosimiglianza è stato esteso a tutta una serie di altre tecniche di analisi statistica. Al di fuori degli studi su tecniche diagnostiche indica quanto bene si adattino i dati osservati al modello interpretativo utilizzato ("goodness of fit").
Registro	Raccolta sistematica di un insieme minimo concordato di informazioni relative a una patologia, una popolazione, una condizione clinica, un intervento medico specifico. Teoricamente dovrebbe coprire tutti i casi che si verificano nella popolazione compresa nell'ambito degli obiettivi del registro. Non esistono ancora criteri concordati per definire la validità di un registro. È anche difficile definire il peso relativo di un registro per determinare/influenzare le raccomandazioni di una linea guida, dato che un registro non è sottoposto a verifiche sistematiche di qualità. Il rischio di *bias* associato ad un registro per interventi (e spesso anche per patologie) è comunque molto elevato.
Revisione non sistematica	Una revisione o una metanalisi che non ha eseguito una ricerca esaustiva della letteratura pertinente e contiene solo una selezione degli studi eseguiti su una specifica domanda clinica, o che non ha definito i metodi per la ricerca e la valutazione degli studi in essa contenuti. Una revisione non sistematica e la relativa metanalisi possono facilmente fornire risultati fuorvianti.
Revisione sistematica	Una revisione della letteratura in cui sono stati utilizzati (e pre-specificati in un protocollo) metodi appropriati per identificare, valutare e riassumere studi che affrontano una domanda clinica predefinita. Può, ma non necessariamente deve, implicare la metanalisi dei risultati.
Riduzione assoluta del rischio (ARR; *absolute risk reduction*)	La differenza assoluta tra il rischio nel gruppo sperimentale e quella nel gruppo di controllo in uno studio. Si utilizza quando il rischio nel gruppo di controllo supera il rischio nel gruppo sperimentale e si calcola sottraendo il rischio assoluto nel gruppo sperimentale dal rischio assoluto nel gruppo di controllo. Questo indice non dà nessuna informazione sulla proporzione di riduzione del rischio fra i due gruppi. Per questa informazione si utilizza il rischio relativo (RR).

Riduzione relativa di rischio (RRR; *relative risk reduction*)	La riduzione proporzione di rischio tra i partecipanti assegnti al gruppo di sperimentale e quelli assegnti al gruppo di controllo in uno studio. È il complemento a 1 del rischio relativo (RRR=1-RR). È una misura di esito che però può essere fuorviante in assenza dell'indicazione del rischio assoluto nel gruppo di controllo.
Rischio assoluto (AR; *absolute risk*)	La probabilità che in un soggetto si verifichi l'esito specificato entro il periodo di tempo specificato. Può variare tra 0 e 1 (espresso come probabilità) o tra 0% e 100% (espresso come percentuale). Diversamente dall'uso comune, il termine "rischio" si può riferire a eventi sfavorevoli (p.es. il verificarsi di un ictus) o a eventi favorevoli (p.es. guarigione completa).
Rischio basale	Il rischio che un evento avvenga senza il trattamento attivo. È stimato dal rischio basale nel gruppo di controllo. È un elemento chiave per stimare la potenziale utilità di un intervento. A parità infatti di riduzione relativa di rischio, i soggetti con rischio basale più elevato possono beneficiare maggiormente dell'intervento (**NNTB** più piccolo).
Rischio relativo (RR)	Quante volte è più (RR>1) o meno (RR<1) probabile che si verifichi un evento in un gruppo a confronto con un altro gruppo. È il rapporto del rischio assoluto (**AR**) per ciascun gruppo. Sovrapponibile alla ***odds ratio*** (OR) quando vi sono pochi eventi. Nei documenti di origine USA a volte la *odds ratio* viene indicata col termine *relative risk*. Poiché le due misure non sono sovrapponibili se non in presenza di pochi eventi, e preferibile attenersi alla definizione rigorosa di RR come AR(test)/AR(controlli).
Sensibilità	La probabilità di avere un risultato positivo a un test nel caso sia presente la patologia in esame.
Serie di casi	Analisi di una serie di soggetti con una certa patologia. Non si possono eseguire confronti nelle serie di casi. L'evidenza associata alle serie di casi è debole. In ogni caso, non può essere definita "serie di casi" una serie di soggetti che non sia esplicitamente indicata come consecutiva (rimane a livello di rapporto multiplo di casi individuali).
Significativo (statisticamente significativo); significatività	Convenzionalmente nell'accezione "statisticamente significativo al livello del 5%"; si può esprimere come "l'intervallo di confidenza al 95% non include il valore corrispondente all'assenza di effetto". Significa che i risultati osservati nello studio (o risultati di ampiezza maggiore) possono essere giustificati dal caso solo 1 volta su 20. È preferibile evitare di utilizzare il termine "significativo" in altre accezione (p.es. in relazione all'entità di un esito clinico) a meno che sia chiaramente specificato che il termine non è utilizzato in senso statistico.
Specificità	La probabilità di avere un risultato negativo a un test nel caso non sia presente la patologia in esame.
Studio aperto	Uno studio nel quale sia i soggetti arruolati sia chi assegna il trattamento sia chi valuta l'esito sono a conoscenza dell'intervento cui il soggetto è stato allocato. È di limitato peso nella valutazione delle evidenze a causa dell'elevato rischio di *bias*.
Studio caso-controllo	Disegno sperimentale che esamina un gruppo di soggetti che hanno avuto un evento (di solito negativo) e un gruppo di soggetti che non hanno avuto lo stesso evento, in cerca della maniera nella quale l'esposizione ad agenti sospetti (di solito nocivi) differiva tra i due gruppi. Questo tipo di disegno sperimentale è principalmente utile per cercare di accertare la causa di eventi rari, come p.es., forme rare di tumore. Gli studi caso controllo possono generare solo ***odds ratio*** (OR), non **rischi relativi** (RR). Gli studi caso controllo generano evidenze più deboli degli studi coorte ma più affidabili delle serie di casi.
Studio controllato (CCT; *controlled clinical trial*)	Uno studio nel quale i partecipanti sono assegnati a due o più gruppi di trattamento. Normalmente limitato a quegli studi in cui il meccanismo di allocazione è diverso dalla randomizzazione. Nel caso di studi controllati nei quali l'allocazione è ottenuta per randomizzazione, si preferisce il termine "studio clinico randomizzato e controllato" o RCT ("randomised controlled trial"). Gli studi controllati e non randomizzati sono maggiormente soggetti a rischio di *bias* degli RCT.
Studio coorte	Un disegno non sperimentale di studio che segue un gruppo di soggetti (una coorte) ed esamina come gli eventi differiscano tra soggetti del gruppo. Uno studio che esamina una coorte che differisce in relazione all'esposizione a sospetti fattori di rischio (p.es., fumo di sigaretta) è utile per cercare di determinare se è probabile che l'esposizione causi specifici eventi (p.es. l'ictus). Gli studi coorte prospettici (che seguono i partecipanti a partire da un certo momento in avanti raccogliendo i dati a mano a mano che si presentano) sono più affidabili rispetto agli studi retrospettivi (che raccolgono dati già registrati in passato). L'evidenza associata agli studi coorte è relativamente bassa e si considera solo quando non è ragionevole attendersi livelli superiori di evidenza (vedi **Cap. 2**) per rispondere alla stessa domanda.

Studio crossover randomizzato	Uno studio in cui i partecipanti ricevono un certo trattamento del quale viene misurato l'effetto, quindi ricevono un altro trattamento e il relativo effetto viene nuovamente misurato. I trattamenti possono essere più di due. L'ordine dei trattamenti è assegnato per randomizzazione. A volte si utilizza un periodo senza trattamento prima dell'inizio dello studio e fra trattamenti successivi (periodi di "washout") per ridurre per quanto possibile l'interferenza fra trattamenti (effetto di trascinamento o "carry-over"). L'interpretazione dei risultati da studi randomizzati controllati in crossover è complesso e aumenta grandemente di complessità con il numero dei trattamenti consecutivi. Negli studi in crossover c'è il rischio – non facilmente misurabile – che un certo intervento possa esplicare i suoi effetti anche dopo la sua interruzione, o perché il periodo di *washout* non è sufficientemente lungo, o per altri fenomeni (apprendimento, trascinamento, induzione, interferenza metabolica). L'assenza di eterogeneità statisticamente significativa non è sufficiente ad escludere l'assenza di eterogeneità clinicamente rilevante. In linea di principio l'evidenza da studi crossover, anche se randomizzati, è più debole per gli esiti dei periodi successivi al primo, rispetto a quelli del primo periodo di trattamento.
Studio osservazionale	Studi nei quali si registrano i dati di interesse in una popolazione o in un campione seguito senza operare interventi sulla normale pratica clinica. Anche con questa definizione, tuttavia, uno studio osservazione prevede un adeguato protocollo di ricerca. Gli studi osservazionali non possono fornire evidenze di effetto. Sono invece appropriati per raccogliere evidenze in merito a prognosi, eziologia, incidenza e prevalenza. Fanno parte degli studi osservazionali le serie di casi, gli studi caso-controllo, gli studi coorte prospettici e retrospettivi.
Studio pragmatico	Studio randomizzato e controllato progettato per fornire risultati direttamente applicabili alla pratica clinica. Si contrappone a studio esplicativo che è progettato per ottenere una stima accurata di efficacia in condizioni ideali (come la maggioranza degli studi di fase III). Gli studi pragmatici reclutano una popolazione rappresentativa di coloro normalmente trattati, permette una normale aderenza basata sulle istruzioni normalmente fornite e analizza i risultati secondo la tecnica ***intention to treat***. Il vantaggio dello studio pragmatico è associato alla sua corrispondenza con la realtà. Il rischio è l'introduzione, nella stima degli effetti, anche di effetti dovuti a fattori di confusione normalmente evitati con i più stretti criteri di inclusione/esclusione e procedure di *follow-up* degli studi esplicativi. Per ridurre il potenziale impatto di tali fenomeni, gli studi pragmatici devono avere dimensioni adeguate e procedure di randomizzazione accurate.
Studio quasi randomizzato	Uno studio nel quale la tecnica di allocazione dei partecipanti alle diverse forme di trattamento non è davvero casuale (p.es., per data di nascita, per giorno della settimana).
Studio randomizzato	Studio nel quale la tecnica di allocazione dei partecipanti alle diverse forme di trattamento è determinata esclusivamente dal caso. Sono tecniche adeguate di randomizzazione il lancio della monetina (se la moneta è "onesta"), le tavole dei numeri casuali, le sequenze generate da computer con programmi di randomizzazione certificati.
Studio randomizzato e controllato (RCT; *randomised controlled trial*)	Studio nel quale i partecipanti sono assegnati per randomizzazione a uno o più gruppi, almeno uno dei quali (gruppo sperimentale) riceve un trattamento/intervento in valutazione e l'altro (gruppo di controllo o di riferimento) riceve il un trattamento alternativo o placebo o lo "standard of care". Questo disegno sperimentale permette una stima sufficientemente accurata dell'effetto relativo dei vari interventi. L'evidenza fornita dagli studi randomizzati e controllati è di livello elevato ma deve essere considerata probante solo quando più studi indipendenti generano risultati coerenti.
Studio sperimentale	Uno studio nel quale il ricercatore valuta gli effetti dell'alterazione voluta di uno o più fattori apportata in condizioni controllate. Uno di questi fattori può essere il trattamento/ intervento.
Studio trasversale ("cross-sectional")	Un disegno di studio che implica l'osservazione di una popolazione in merito a una esposizione o condizione o entrambe, per uno specifico arco temporale o uno specifico momento nel tempo. Può essere utilizzato per determinare la prevalenza di una condizione nella popolazione. Non può essere utilizzato per stimare la causalità di un intervento o trattamento.
Tasso di eventi ("event rates")	Nel determinare la potenza di uno studio basato su eventi, il tasso di eventi è più rilevante rispetto al numero di partecipanti. Quindi la stima di un effetto – e soprattutto la stima dell'affidabilità di un modello interpretativo del verificarsi degli eventi – dipende in primo luogo dal numero (o tasso) di eventi e successivamente dal numero di soggetti esaminati.

Tendenza (*trend*)	Termine rigorosamente limitato all'analisi di regressioni e di frequenze marginali in tabelle di contingenza complesse. Assolutamente da evitare – e privo di valore semantico – se riferito a differenze di esito interessanti ma non statisticamente significative.
Validità	Termine ambiguo. Si usa per indicare la rigorosità o la consistenza di uno studio. Si considera "internamente valido" uno studio se il modo in cui è progettato ed eseguito porta a risultati non *biased* e fornisce una stima accurata dell'effetto che si intende misurare. Si considera "esternamente valido" uno studio i cui risultati sono applicabili ai normali pazienti trattati nella pratica clinica quotidiana. L'uso del termine "validità" implica un giudizio soggettivo da parte degli esaminatori degli studi e non è quantificabile.
Validità esterna (generalizzabilità)	La validità dei risultati di uno studio al di là dello studio stesso. Uno studio randomizzato e controllato fornisce evidenza diretta di causalità entro lo studio stesso. È necessario un ulteriore passo logico sostenuto da informazioni specifiche per applicare i risultati dello studio a un ambito più generale. Tuttavia, uno studio viene condotto proprio per stimare se un certo intervento può essere utile alla popolazione che non ha partecipato allo studio. Si assume quindi che i risultati di uno studio siano generalizzabili all'intera popolazione a meno che vi siano evidenze del contrario. Evidenze di effetti consistenti (**vedi**) in differenti *setting* e differenti popolazioni forniscono evidenza in favore di validità esterna. Se l'evidenza dell'effetto è raccolta solo da *setting* atipici (p.es. solo in centri di eccellenza mentre la maggior parte dei casi è vista in centri di primo livello o in medicina territoriale) allora bisogna mantenere un sano scetticismo rispetto alla generalizzabilità dei risultati. La generalizzabilità dipende non solo, anche se principalmente, dai criteri di arruolamento (inclusione ed esclusione), ma anche dalla popolazione dalla quale il campione in studio è stato estratto (vedi anche **applicabilità**).
Valore predittivo negativo	La probabilità di non avere la patologia considerata in presenza di un risultato negativo del test in esame (da non confondere con specificità).
Valore predittivo positivo	La probabilità di avere la patologia considerata in presenza di un risultato positivo del test in esame (da non confondere con sensibilità).
Vero negativo	Soggetto con la condizione in esame secondo il *gold standard*, che abbia un risultato negativo al test sotto studio.
Vero positivo	Soggetto con la condizione in esame secondo il *gold standard*, che abbia anche un risultato positivo al test sotto studio.

Risultati ottenuti da un singolo studio clinico

In linea di principio si considera che i risultati di un singolo studio non siano probanti (cioè non possano costituire elemento per modifiche alla pratica clinica). Per essere considerati probanti, gli stessi risultati devono essere ottenuti in due studi indipendenti condotti in campioni indipendenti estratti casualmente dalla stessa popolazione.

Questo principio è motivato dal fatto che un singolo studio è normalmente progettato con una numerosità campionaria adeguata per escludere il rischio di falso positivo (errore di tipo I, errore alfa, rischio di vedere un effetto inesistente) con probabilità del 5%. Indipendentemente perciò dal valore di "P" effettivamente osservato, il risultato ottenuto ha il rischio di essere falso una volta su 20, limite non accettabile per essere utilizzato in medicina. Invece, la probabilità che due studi indipendenti, ciascuno progettato per un alfa del 5%, indichino entrambi un risultato non vero è pari a $0{,}05 \times 0{,}05 = 0{,}0025$. Quindi il rischio per i pazienti che deriva dall'utilizzare in pratica clinica i risultati di due studi indipendenti coerenti è pari allo 0,25%, cioè accettabile.

Possono però essere progettati studi capaci di escludere un falso positivo con probabilità pari a 1% o meno. In questo caso il risultato osservato può essere trasferito con pochi rischi direttamente alla pratica clinica. Tuttavia in tali casi bisogna considerare la effettiva dimensione del campione. Spesso infatti tali studi sono di grandi dimensioni e coinvolgono soggetti fra loro eterogenei (p.es. di nazioni o continenti diversi con le ovvie conseguenze sul profilo genetico e sulle abitudini di vita). In tal caso, o vi sono evidenze probanti che l'evoluzione della patologia e la risposta al trattamento non è influenzata da tali fattori, oppure bisogna comunque valutare se l'applicabilità dei risultati può essere ragionevolmente estesa alla popolazione in cui si sta operando (da cui il principio di "applicabilità diretta" utilizzato nella valutazione del grado delle evidenze).

Potrebbero però verificarsi situazioni in cui uno studio progettato per escludere falsi positivi con probabilità <0,05, osservi il risultato voluto con P<0,001 (limite massimo di accuratezza per stimare la P da parte di un sistema che non utilizzi un supercomputer; valori più piccoli sono privi di significato). Diventa accettabile questo risultato anche senza conferma? In realtà no. Il valore di P osservato è uno degli infiniti valori distribuiti secondo la curva normale intorno al valore vero (anche la probabilità è una variabile). Quindi, quando si definisce un livello di accettabilità, è implicito che si accettano il valore della soglia (0,05) e tutti i valori più piccoli della soglia. Esiste quindi una certa probabilità che la stima puntiforme del valore di P sia più piccola del livello di soglia, ma ciò era implicito nella definizione stessa di livello soglia. Il problema diventa più chiaro se espresso in termini di confidenza. Se lo studio era progettato per potere avere una confidenza del 95% di non accettare dei falsi positivi, il risultato effettivamente osservato non cambia il livello di confidenza per quello studio: potrà solo suggerire a cambiare il livello di confidenza da considerare per progettare uno studio successivo. Questa situazione potrebbe essere paragonata alla situazione di un aereo di linea in volo. L'aereo è progettato per volare a una velocità di 800 km/h in aria calma (errore alfa – o P critica per affermare la significatività – di progetto dello studio), che in tal caso corrisponde a 800 km/h rispetto al suolo. È possibile che, con un forte vento in poppa, si raggiungano i 1·000 km/h rispetto al suolo, ma la velocità rispetto all'aria rimane 800 km/h. Quindi, vedere sfrecciare un aereo a 1·000 km/h non ci permette di affermare che la velocità di quell'aereo è superiore a quella di progetto (o che un P osservato è più "significativo" o "più piccolo" di quello di progetto), anche se non ci dispiace arrivare prima. Tutta la discussione sulla potenza (vedi **2.8**) corrisponde invece a tenere conto del fatto (e cercare di neutralizzarlo aumentando la numerosità del campione) che è anche possibile incontrare un forte vento contrario che riduce la velocità rispetto al suolo (riduce il valore di P osservato a parità di tutte le altre condizioni), anche se si continua a volare a 800 km/h rispetto all'aria in movimento.

Endpoint primario ed endpoint secondari

Qualsiasi studio clinico è programmato con un obiettivo, un esito e un "endpoint". L'obiettivo è la domanda di natura clinica cui si vuole dare una risposta. L'esito è lo specifico aspetto clinico capace di fornire la risposta. Lo "endpoint" è la misura dell'esito clinico dalla quale – attraverso algoritmi statistici – si ricava la risposta.

Uno studio clinico viene però progettato in funzione di una specifica capacità di esclusione di falsi positivi (l'aspetto dei "falsi negativi" è correlato alla potenza). Questo limite è un limite complessivo dello studio. Quindi, se si intende rispondere a più domande, bisogna verificare prima di tutto se gli esiti che rispondono alle diverse domande sono correlati tra loro. Se lo sono, allora bisognerà distribuire l'errore alfa totale tra i diversi "*endpoint*" correlati (ovvero, si stima la numerosità campionaria non più per alfa, ma per alfa diviso *k*, dove *k* è il numero di *endpoint* correlati: regola di Bonferroni). Se gli *endpoint* non derivano da esiti correlati – situazione molto rara in uno studio clinico proprio per la necessaria specificità degli interventi terapeutici che si utilizzano – allora si deve stimare la numerosità campionaria per ciascuno degli *endpoint* indipendenti e utilizzare la numerosità massima trovata. In ogni caso, la numerosità campionaria di studi con *endpoint* multipli è molto elevata. Per questo motivo, gli *endpoint* cosiddetti "primari" (che significa "dimostrabili") sono normalmente limitati a uno o due.

D'altra parte sarebbe assurdo "sprecare" tutte le informazioni raccolte al di là di quelle indispensabili a stimare l'*endpoint* primario. Si possono quindi porre altre domande oltre a quella principale, e valutare la risposta ottenuta nello studio. Ciò è del tutto legittimo, sia che queste risposte "secondarie" siano previste in sede di progettazione, sia che vengano osservate dopo la conclusione dello studio ("ex post-facto"). Tuttavia le implicazioni sull'applicabilità (forza) dei risultati sono sostanzialmente diverse.

I risultati di un *endpoint* primario sono probanti, vale a dire che possono essere direttamente utilizzati per modifiche della pratica clinica.

I risultati di *endpoint* secondari, invece, non sono probanti. Ciò non significa che non siano veri, ma solo che non è detto che lo siano e ciò indipendentemente dal valore di "P" osservato: questo è il significato della frase che compare negli studi più seri e recita «il valore di P riferito a questi risultati è citato con mero significato descrittivo». Tuttavia, se l'*endpoint* secondario era stato pianificato *a priori*, i risultati ottenuti possono essere utilizzati per stimare con buona accuratezza se la domanda clinica relativa merita di essere sottoposta a verifica in un nuovo studio, e quale debba essere la dimensione di tale studio.

Quando invece gli *endpoint* secondari sono stati identificati "ex post-facto", occorre notevole cautela anche solo per la progettazione di un nuovo studio. *endpoint* non pianificati possono risultare "statisticamente significativi" o per pura casualità ("the play of chance" negli articoli) o perché si sono modificati continuamente i termini dell'*endpoint* fino a trovare qualcosa con P<0,05 (cosiddetto "data dredging") e riportarlo come "*endpoint* secondario definito *a posteriori*".

In ogni caso i risultati di *endpoint* secondari non possono influenzare la pratica clinica finché non sono sottoposti a verifica in studi specificamente ed adeguatamente progettati (con le eccezioni indicate in

seguito valide però solo per casi davvero clamorosi). Si possono ricordare, a questo proposito, i risultati degli studi ELITE (*Lancet* 1997; **349**: 747-752) ed ELITE II (*Lancet* 2000; **355**: 1582-1587) per confrontare la tollerabilità renale fra captopril e losartan in pazienti anziani con scompenso cardiaco). Alla fine dello studio ELITE, l'endpoit secondario mortalità era significativamente inferiore con losartan. Alla fine dello studio ELITE II, l'*endpoint* primario mortalità era sovrapponibile tra i due.

Analisi intermedie ed analisi di sottogruppi

Questi due aspetti sembrano molto distanti tra loro, ma condividono lo stesso problema statistico. Tutto comincia dall'errore alfa accettato in sede di progettazione dello studio, tipicamente il 5%. Questo significa che lo studio è dimensionato per osservare falsi positivi (effetti inesistenti) una sola volta su venti. "Nel suo complesso" è una dizione letterale: il rischio totale disponibile per quello studio è il 5%. Se questo viene "speso" tutto per la sola analisi dell'*endpoint* primario, la confidenza che il risultato rifletta i fatti è del 95%. Se viene "speso" in altre analisi, alla fine la confidenza che il risultato sia "vero" è minore. "Quanto" minore dipende da quante analisi sono fatte. L'approccio più semplice è di considerare la confidenza da assegnare al risultato finale come pari alla confidenza non spesa: se in uno studio programmato per un errore alfa del 5% si sono eseguite due analisi intermedie (o si sono esaminati due sottogruppi) oltre all'analisi finale e per ciascuna analisi si definisce "si considerano significativi i risultati con P<0,05", allora la confidenza attribuibile a questo risultato è pari all'85% (1-0,05×3: significa che un P nominale pari a 0,05 per questa analisi corrisponde a un P reale pari a 0,15; oppure che per ottenere una confidenza del 95% bisogna ottenere un P nominale pari a 0,0167). Una confidenza così bassa è chiaramente inaccettabile per trarre conclusioni cliniche da applicare a pazienti veri. Esistono tecniche complesse che permettono di "spendere" poco errore alfa nelle analisi intermedie e mantenerlo quanto più possibile disponibile per l'analisi finale, ma questo va pianificato prima dello studio e, comunque, il risultato finale è un aumento della dimensione del campione non molto diversa da quella ottenuta considerando come P critico il valore di P/*k* (regola di Bonferroni).

Anche per eventuali sottogruppi ogni analisi va pianificata come *endpoint* primario correlato all'analisi globale. Anche in questo caso si può "spendere" diversamente l'errore alfa sulle diverse analisi, ma anche questo va programmato prima dello studio e, soprattutto, nei risultati deve essere ben evidenziato il grado di confidenza attribuibile a ciascuna analisi (cosa che, purtroppo, non viene quasi mai fatta). Per questo motivo le analisi di sottogruppi pianificati vanno esaminate con cura per quanto attiene alla confidenza con la quale i risultati osservati possono suggerire modifiche di pratica clinica. Le analisi di sottogruppi non pianificati hanno senso solo per indirizzare la futura ricerca, ma sono prive di valore probante per modificare la pratica clinica (vedi anche "*endpoint* secondari non pianificati" in **2.2**). Si può ricordare qui il caso classico dello studio ISIS-1 in cui venne osservato, nell'analisi di sottogruppi non pianificati, che, tra l'altro, l'atenololo era maggiormente efficace nei pazienti nati sotto il segno della bilancia. Si può anche ricordare lo studio PRAISE (efficacia dell'amlodipina vs placebo in pazienti con scompenso cardiocircolatorio; *N Engl J Med* 1996; **335**: 1107-1114). L'analisi globale non evidenziava differenze di mortalità; all'analisi per sottogruppi non pianificati si identificò un sottogruppo (eziologia non ischemica) nel quale amlodidipina riduceva significativamente la mortalità. Lo studio PRAISE-2 (parzialmente discusso in *Am Heart J* 2004; **147**: 151-157), progettato per confermare questo risultato, riscontrò che la mortalità nei due gruppi di trattamento era invece sovrapponibile.

L'approccio all'analisi di sottogruppi può assumere un aspetto apparentemente più "tecnico" rispetto al semplice confronto di un sottogruppo verso un altro (o il sottogruppo di riferimento), che consiste nell'analisi di interazione. In questo approccio, si valuta l'effetto non tanto dell'appartenenza al sottogruppo A, B, C, ... sulla probabilità di ricadere nel gruppo con esito positivo o negativo, come nella normale analisi logistica multivariata, ma si stima l'interazione fra il fattore di classificazione principale (generalmente il trattamento cui un soggetto è stato randomizzato) e l'appartenenza al sottogruppo. In tal modo si ritiene di poter evidenziare, se l'interazione è statisticamente significativa, che l'effetto del trattamento è differente in funzione della caratteristica che definisce il sottogruppo (tipicamente: sesso, gravità della patologia, ...). Purtroppo anche questo approccio può essere fuorviante, anche se in grado minore rispetto alla tradizionale analisi di sottogruppi. Infatti, anche l'analisi di interazione introduce un *bias*, specie se i sottogruppi non erano definiti *a priori* e quindi sono sbilanciati, in relazione a tutte le altre analisi statistiche. Anche in questo caso vale la regola generale che, per poter definire significativa un'interazione, la P osservata deve essere più piccola della P critica definita, come indicato sopra, come P/*k*, dove *k* è il numero di analisi pianificate (incluse le analisi di interazione). Utilizzando l'analisi di interazione è possibile ottenere fattori di correzione meno importanti, ma non è mai possibile evitarli. Per una discussione più approfondita di questo aspetto, si veda: Lagakos SW. The Challenge of Subgroup Analyses – Reporting without Distorting. *N Engl J Med* 2006; **354**: 1667-1669.

Metanalisi ed eterogeneità

Si discute spesso dell'importanza dell'eterogeneità nelle metanalisi, perché una elevata eterogeneità fra studi diminuisce il valore delle metanalisi come guida alla pratica clinica.

In primo luogo bisogna distinguere l'eterogeneità clinica dall'eterogeneità statistica.

L'eterogeneità clinica si riferisce al fatto che gli studi clinici sono stati condotti in maniera diversa e non assimilabile tra loro, quindi sui diversi studi potrebbero avere agito influenze casuali e/o sistematiche diverse. Dal punto di vista puramente statistico ciò è un vantaggio perché, aumentando la variabilità, rende più probanti (statisticamente) eventuali risultati "statisticamente significativi". Tuttavia potrebbe anche costituire uno svantaggio in termini clinici, perché parte dei risultati ottenuti potrebbero non essere applicabili in un determinato contesto (nel caso di queste linee guida, alla popolazione italiana residente e trattata nell'ambito del sistema sanitario e delle strutture cliniche italiane). Se esiste una eterogeneità clinica, l'unico approccio di rilievo è la valutazione di scenari diversi in funzione di diversi parametri di riferimento (analisi di sensibilità). Se l'analisi di sensibilità porta comunque agli stessi risultati, allora l'eterogeneità clinica non costituisce problema, altrimenti è preferibile affinare la metanalisi dopo avere definito con la massima precisione possibile quali sono i parametri di riferimento coerenti con la situazione specifica che si vuole esaminare.

L'eterogeneità statistica stima invece quanto è probabile che i risultati siano distribuiti in maniera omogenea o meno nei diversi studi, e dipende sia dai risultati dei singoli studi, sia dalla loro dimensione (o da altri fattori che ne determinano il "peso"), sia dall'effetto medio osservato. In questo modo è possibile stimare quanto è probabile che la distribuzione osservata dei risultati possa essere osservata per puro effetto del caso (comunemente stimata con il test Q di Cochran). Una eterogeneità statisticamente significativa può – ma non necessariamente – indicare un errore sistematico o in alcuni studi o nella selezione degli studi. Tuttavia il test Q indica solo se si deve respingere o no l'ipotesi di omogeneità fra studi, ma non ne stima l'entità. Recentemente si è iniziato a utilizzare a questo scopo l'indice I^2, che misura la dimensione della eterogeneità vera e può essere interpretato come la percentuale della variabilità totale dell'effetto calcolato nella metanalisi, da attribuirsi alla eterogeneità fra studi. Chiaramente, un effetto gravato da elevata eterogeneità (nell'ordine del 70%~75%) va considerato con maggiore cautela rispetto ad un effetto gravato da minore eterogeneità (nell'ordine del 20%~30%), indipendentemente dalla capacità del test Q di osservare la "significatività statistica" della stessa eterogeneità. Naturalmente, come tutti i risultati statistici, l'indice I^2 è davvero informativo se accompagnato dal suo intervallo di confidenza al 95%, che permette di valutare anche l'accuratezza della stima di eterogeneità.

Quindi i risultati delle metanalisi hanno maggiore o minore valore probante per il trasferimento nella pratica non solo in funzione dell'entità dell'effetto osservato (sempre considerando il relativo intervallo di confidenza) ma anche in funzione della eventuale eterogeneità clinica (valutandone l'impatto rispetto all'utilizzo dei risultati nella popolazione e nel sistema sanitario bersaglio) e della eterogeneità statistica considerata in primo luogo come dimensione dell'eterogeneità vera ($I^2 \pm IC_{95}$) e poi anche come significatività.

Esaminando, p.e., la metanalisi di confronto tra endoarteriectomia e *stent* nell'ictus (*Stroke* 2005; **36**: 905-911; l'articolo riporta i dati del chi quadrato; lo I^2 è stato calcolato dai dati presentati nelle tabelle) si osserva: per l'esito "morte o ictus entro 30 giorni" una eterogeneità statistica significativa (P=0,035) ma un eterogeneità vera non rilevante (I^2 61,4%) che, tuttavia, presenta un intervallo di confidenza di tale ampiezza (IC_{95} 0%-85,5%) da suggerire estrema cautela nell'interpretazione dell'analisi; per l'esito "morte o ictus a 1 anno", una eterogeneità statistica significative (P=0,016) e una eterogeneità vera ugualmente elevata (I^2 75,9%; IC_{95} 21%-93%). Sulla base di questi risultati si evidenzia la necessità di studi clinici comparativi meglio progettati o più standardizzati per ridurre l'eterogeneità che riduce il valore della metanalisi. Inoltre, dato che gli studi considerati sono eterogenei, continueranno a contribuire alla dispersione della stima delle nuove metanalisi, i nuovi studi dovranno essere o di dimensioni decisamente maggiori (peso maggiore nella metanalisi) o in numero consistente.

Modo di esprimere i risultati nelle metanalisi (e negli studi)

La modalità con cui sono espressi i risultati negli studi e nelle metanalisi può influenzare in modo importante il peso apparente di un risultato rispetto al suo peso vero. Nel valutare la rilevanza dei risultati espressi per una raccomandazione di una linea guida – quindi nel valutare quanto sia rilevante per suggerire modifiche di pratica clinica – gli elementi chiave da prendere in considerazione sono il peso assoluto e l'accuratezza con cui tale peso può essere stimato, sempre tenendo conto delle considerazioni già indicate sopra. L'elemento fondamentale è la stima dell'effetto assoluto. Nessuna stima di merito può basarsi su altre modalità di espressione dei risultati.

Tralasciamo per sostanziale irrilevanza nella definizione di linee guida gli esiti espressi in forma continua (differenze di medie) limitandoci ai risultati espressi in forma di categoria, generalmente dicotomica (p.es. vivo/morto a una certa distanza da una procedura).

In queste situazioni sono possibili molti modi di esprimere lo stesso risultato: riduzione relativa di rischio (o *relative risk reduction*, RRR), riduzione assoluta di rischio (*absolute risk reduction*, ARR), *odds ratio*, *hazard ratio* (rapporto di tasso di rischio), NNT, ...

Le definizioni dei termini sono indicate nel glossario. Queste considerazioni generali tengono conto solo del peso delle diverse misure. Il primo aspetto da considerare è quanto descrittive sono le diverse statistiche della reale entità dell'effetto. Quella che può generare più confusione è la riduzione relativa di rischio perché è espressa come variazione percentuale di una entità non presentata (cioè il rischio assoluto di riferimento). Chiaramente, se il rischio di riferimento è molto piccolo, anche una grande riduzione relativa è di importanza solo marginale, mentre se il rischio di riferimento è grande anche una piccola riduzione relativa può essere applicata utilmente (per i pazienti) nella pratica. Quindi questa misura non viene considerata utile per valutare un effetto.

Un livello inferiore di confusione è associato al rapporto del tasso di rischio o *hazard ratio*. La *hazard ratio* è il tasso di eventi per persona-tempo. La *hazard ratio*, allora, indica semplicemente, per qualsiasi coppia di soggetti presa a caso, una dal gruppo con *hazard rate* più alto (p.es. controlli) e una dal gruppo con *hazard rate* più basso (p.es. trattati) il rapporto di *odds* per il verificarsi dell'evento. Una *hazard ratio* di 2 corrisponde quindi al 67% di probabilità di avere prima l'evento per il soggetto di controllo rispetto al soggetto trattato (*odds*). Una semplice trasformazione dell'*hazard ratio* (HR) in *odds* (non *odds ratio*, che si ottiene in altro modo) si ottiene dalla: *odds* = HR/(1-HR). L'informazione ottenuta dalla *hazard ratio* è quindi riferita sostanzialmente alla probabilità che un evento avvenga prima o dopo, ma indica in maniera solo indiretta la probabilità che l'evento avvenga in assoluto in un periodo di tempo predeterminato, che è la domanda di normale interesse nel definire il peso di un certo risultato per le raccomandazioni pratiche. Anche se la *hazard ratio* possiede un interesse rilevante nel seguire nel tempo eventi con rischio non costante (p.es. il rischio di recidiva di ictus), deve quindi essere usata con cura. Uno dei problemi principali consiste nel modo in cui è costruito il modello interpretativo. Spesso la HR viene calcolata da modelli contenenti diverse variabili esplicative (variabili utilizzate per "suddividere" il rapporto di rischio tra diversi fattori di interesse). Il primo aspetto critico è che ci devono essere abbastanza eventi per ciascuna variabile. Quindi bisogna valutare non il numero di casi, ma il numero di eventi: occorrono 5-10 eventi per variabile per costruire un modello valido. Inoltre il rischio deve essere omogeneo nel tempo per ciascuna delle variabili considerate. Se ciò non è vero, il modello può dare risultati fuorvianti. Infine occorre che le variabili esplicative abbiano un senso di per sé. Infatti, se si lascia al computer di scegliere un modello solo su basi matematiche, date sufficienti variabili esplicative, una qualche relazione risulterà significativa per caso, ma è facilmente priva di significato clinico.

Uno dei metodi usati per esprimere i risultati è quello che usa la *odds ratio*. Per eventi rari corrisponde al rischio relativo. Per eventi frequenti diverge dal rischio relativo in maniera anche importante. La scelta se usare il rischio relativo o la *odds ratio* dipende sia dal tipo di studio (negli studi prospettici sono utilizzabili entrambi; negli studi retrospettivi si può utilizzare solo la *odds ratio*) sia dal numero di eventi (se sono rari i due metodi sono equivalenti, se non sono rari la *odds ratio* può essere fuorviante).

Tutti questi modelli, tuttavia, se hanno il pregio di stimare il rapporto di efficacia tra due procedure, non sono clinicamente "informativi". In altri termini, anche se "misurano" il rapporto di efficacia tra due procedure, non ci dicono nulla sulla reale utilità. Per stimare l'utilità si ha bisogno di sapere quanti pazienti trarranno beneficio dall'una o dall'altra. Qui interviene il "numero di pazienti da trattare" o NNT. Più specificamente si usano i termini NNTB (*number needed to benefit*) per eventi da considerare in senso positivo, e NNTH (*number needed to harm*) per eventi da considerare in senso negativo. Il rapporto NNTH/NNTB (con il suo intervallo di confidenza; definito anche LHH o "Likelihood of Being Helped Versus Harmed" [Burneo JG, Wiebe S. Chapter 3: Outcome and adverse effect measures in neurology. In: Candelise L, Hughes R, Liberati A, Uitdehaag BMJ, Warlow C, eds. Evidence-Based Neurology, Management of Neurological Disorders. Oxford, UK: Blackwell Publishing, BMJI Books; 2007: 20.]) fornisce una stima del margine di sicurezza dell'intervento. Lo NNT è il numero di pazienti da trattare per prevenire un evento negativo o per ottenere un evento positivo in più rispetto all'intervento di confronto. Deve specificare il trattamento, la sua durata e l'evento prevenuto (o ottenuto). Si calcola come 1/ARR e va sempre riportato con il suo intervallo di confidenza. In questo modo si confrontano gli interventi sulla base del numero di soggetti che possono beneficiarne, e dalla stima dell'intervallo di confidenza e degli eventuali carichi aggiuntivi (di rischio, disagi e costi) si può valutare quanto cogenti siano i risultati di uno studio clinico (o di una metanalisi) per eventuali modifiche di pratica clinica. Questo approccio è quello prevalentemente utilizzato in queste linee guida.

È bene insistere sul fatto che la misura NNTB/NNTH non è una misura assoluta (invariante rispetto alle condizioni al contorno) ma è legata a una procedura specifica e a un tempo specifico di esposizione a rischio. La stessa procedura può produrre NNTB/NNTH differenti per tempi di esposizione differenti quando il rischio è variabile nel tempo. Questo aspetto è particolarmente rilevante nel caso del rischio di ictus. Un caso evidente è riportato nel Capitolo relativo all’endoarteriectomia carotidea nella stenosi sintomatica.

Modelli prognostici e carte del rischio

Queste linee guida fanno riferimento sia a modelli prognostici sia a carte del rischio. Le cosiddette "carte del rischio" sono una semplificazione dei modelli prognostici. Il modello prognostico permette, sulla base di un algoritmo che considera informazioni facilmente disponibili "al letto del paziente", di stimare la probabilità di un certo esito. Le carte del rischio utilizzano alcune di queste informazioni – generalmente le più frequenti o quelle che maggiormente influenzano la stima di probabilità – e le suddivide per grandi categorie (non necessariamente omogenee tra loro) in modo da rendere più semplice la stima del rischio, senza dover utilizzare dei calcoli.

Le due tecniche soffrono di due principali problemi. La stima del rischio di un evento, proprio per il modo in cui è stata calcolata, è una stima di popolazione e significa che un soggetto con determinate caratteristiche appartiene a una popolazione che manifesterà un certo numero di eventi di un certo tipo in un certo tempo. Non è possibile determinare se il singolo soggetto appartiene a quelli che manifesteranno l'evento o a quelli che non lo manifesteranno. Ciò richiede prudenza nell'affrontare la discussione del rischio presumibile con il singolo paziente per evitare di indurre indebite paure (nonché eccessivo carico di terapie non correttamente seguite) o, peggio, indebite sicurezze.

Il secondo problema è più radicale ed è dovuto alla natura stessa della procedura di costruzione del modello prognostico. Ciascun modello nasce dall'esame di una specifica coorte di soggetti. Non è mai possibile affermare che tale coorte rappresenti la popolazione attuale alla quale lo si intende applicare. Pertanto nessun modello prognostico e nessuna carta del rischio è da considerarsi suggerimento autorevole per modifiche di pratica clinica senza una validazione indipendente su una coorte indipendente estratta dalla stessa popolazione. La cosiddetta "validazione interna", ovvero ottenuta applicando il modello a sottogruppi casuali della stessa popolazione dalla quale esso deriva, è condizione necessaria per la sua validità, ma non è condizione sufficiente proprio perché i vari sottogruppi sono per definizione omogenei alla coorte da cui derivano, ma non sono necessariamente omogenei alla popolazione bersaglio.

I modelli e le eventuali carte del rischio derivati da questi modelli e validati su coorti indipendenti sono da considerarsi elemento importante e probante agli effetti di modifiche della pratica clinica, soprattutto se possono esprimere il risultato in termini di NNTB e NNTH (con il loro intervallo di confidenza), cioè in una forma che permette una stima diretta dell'eventuale beneficio o rischio per i pazienti.

Modelli non validati e relative carte di rischio non validate indipendentemente vanno considerati come ipotesi di lavoro, magari anche interessanti, ma da considerarsi con cautela prima di affidare a tali stime eventuali cambiamenti di pratica clinica che, ad una verifica indipendente, potrebbero rivelarsi, se non controproducenti (nel caso stimassero un rischio inferiore a quello reale), inutili (però con aggravio di carichi sul paziente senza benefici prevedibili).

3.8. Predittività di correlazioni, curve di regressione, analisi di sopravvivenza

Correlazioni, regressioni e analisi di sopravvivenza sono tecniche largamente utilizzate negli studi clinici esaminati per definire queste linee guida. Tutte queste tecniche sono estremamente utili per determinare scenari realistici di utilizzo di tecniche, procedure e interventi sia di prevenzione sia di trattamento. Tuttavia l'impiego dei risultati ottenuti da queste tecniche deve considerare tutti gli aspetti indicati nelle osservazioni precedenti.

Il primo aspetto da considerare riguardo la predittività dei risultati, e quindi la forza con la quale possono suggerire modifiche di pratica clinica, riguarda la corrispondenza del campione esaminato con la popolazione bersaglio. A questo proposito si applicano le stesse considerazioni già riportate in 3.5.

Il secondo aspetto riguarda il numero e la natura dei predittori utilizzati nella costruzione dei modelli di correlazione, regressione e sopravvivenza (in questo caso soprattutto utilizzando la tecnica del rischio proporzionale di Cox). Si devono applicare per questi aspetti le stesse considerazioni riportate al punto3.6. In particolare, nel caso della tecnica del rischio proporzionale secondo Cox, occorre sempre tenere presente che essa è tanto valida quanto i predittori utilizzati, e che anche in questo caso necessita un certo numero di eventi (5-10 almeno) per ogni predittore utilizzato. Inoltre, la validità del modello derivato dipende da quanto è presumibile che sia costante il rischio associato a ciascun predittore su tutto l'arco dei valori del predittore stesso effettivamente presenti nel campione.

L'aspetto più critico e spesso trascurato è che queste tecniche possono essere predittive esclusivamente nell'ambito dei valori di predittori utilizzati per la loro costruzione. In altre parole, una correlazione tra un fattore A e un risultato B può essere valida per tutto il dominio dei valori di A esaminati nella

correlazione, mentre è priva di significato per valori al di fuori di tale dominio. La stessa cosa si applica alle stime di regressione: valori immediatamente al di sotto del valore minimo di A e immediatamente al di sopra del valore massimo di A utilizzati per costruire la regressione, possono essere associati a valori di B completamente al fuori della curva di regressione stimata. Inoltre, se una curva di regressione permette di stimare B a partire da valori di A, non necessariamente – anzi, molto raramente – la stessa curva permette di stimare A a partire da valori di B. Infine, per le curve di sopravvivenza, tutti questi aspetti si integrano a suggerire cautela per evitare conclusioni – e quindi utilizzi – scorretti. Una curva di sopravvivenza vale solo ed esclusivamente per il periodo di tempo coperto dall'osservazione. Le conclusioni associate a tale curva non sono estendibili oltre il periodo osservato se non come generatore di ipotesi da confermare prospetticamente in maniera indipendente. È infatti possibile – e per estensioni sufficientemente lunghe è certo – che si osservino avvicinamenti tra le curve o inversioni di tendenza o perdita di significato a causa di altri fattori non considerati nella curva ma prognosticamente più rilevanti con il passare del tempo. Non è inoltre trasferibile il risultato di una curva di sopravvivenza per un certo esito ad altri esiti per quanto clinicamente correlati (p.es. non necessariamente – anzi molto raramente – una curva di sopravvivenza riferita a un esito composito è anche applicabile ai singoli componenti di tale esito).

Quindi l'approccio generale in queste linee guida è stato quello di considerare particolarmente rilevante per suggerire modifiche di pratica clinica quelle stime di correlazione, regressione e sopravvivenza, condotte su popolazioni per quanto possibile simili alla popolazione bersaglio italiana, utilizzando predittori ragionevolmente correlati all'esito clinico considerato, che esaminassero esiti clinici per quanto possibili robusti e semplici. Per le regressioni e le curve di sopravvivenza si sono prese in primo luogo in considerazione quelle che coprissero un dominio dei valori dei predittori quanto più prossimo al campo realmente osservabile nella pratica clinica. Per le curve di sopravvivenza si sono considerate di particolare peso quelle riferite ad esiti clinici semplici e robusti, su campioni seguiti per un arco di tempo sufficientemente ampio da essere omogeneo o quanto meno confrontabile con il periodo per il quale si prevede si seguire i pazienti esaminati, stimate a partire da un numero sufficientemente grande di eventi bersaglio, e ove possibile confermate in almeno due studi indipendenti (vedi 3.2).

3.9. Significatività e potenza – rappresentatività (campione)

Nell'esame degli studi utilizzati per la costruzione delle linee guida si è dato peso, certamente, all'esistenza di risultati "statisticamente significativi", ma anche alla potenza degli studi e alla rappresentatività del campione. Questi aspetti sono strettamente correlati alla dimensione del campione utilizzato negli studi (la cosiddetta "numerosità campionaria").

Espresso in termini pratici, e al di là delle formule matematiche utilizzate, la dimensione di un campione deve essere tale da permettere di identificare come "statisticamente significativa" una differenza di effetto che sia spiegabile solo raramente da effetti casuali, ma anche tale da poter essere effettivamente rilevata se esiste, e da poter essere applicabile a tutta la popolazione da cui è stato estratto il campione. Questi termini corrispondono a: significatività, potenza, e rappresentatività (che corrisponde, implicitamente, a predittività nei limiti discussi di seguito).

Un risultato statisticamente significativo in uno studio di buona potenza può essere privo di valore predittivo per l'applicazione alla popolazione, se il campione non rappresenta – per aspetti demografici, genetici, fisiopatologici – la popolazione cui si intende applicare il risultato o se l'*endpoint* primario non risponde esattamente alla domanda clinica di interesse per la popolazione considerata in queste linee guida. Questo solo aspetto, indipendentemente dalla significatività del risultato, causa la diversa classificazione della forza di una evidenza esterna considerata in queste linee guida (significato di "direttamente applicabile").

Tuttavia, la significatività statistica per sé non è un elemento sufficiente a dare "forza" probante a uno studio. Anche in questo caso il valore di "P" è elemento necessario ma non sufficiente. Elemento altrettanto rilevante è l'intervallo di confidenza del risultato, la sua natura, il suo peso clinico, il sospetto (o l'evidenza) di *bias*. Conviene ricordare che la significatività statistica è solo una stima di quanto sia probabile che la stessa differenza possa essere osservata per puro caso nel caso si ripetessero le stesse analisi (da cui necessità della definizione di uno e uno solo *endpoint* ben definito o, se più di uno, adeguato incremento della numerosità campionaria). Conviene anche ricordare che il valore effettivo di "P" osservato e riportato non aumenta minimamente la forza dell'evidenza rispetto a quanto ipotizzato nel protocollo di studio.

La "potenza" di uno studio clinico è uno degli aspetti forse meno comunemente considerati nelle discussioni ma riveste una rilevanza notevole per la costituzione di linee guida. Tecnicamente, la "potenza" di uno studio è la sua capacità di osservare una differenza di esito fra i due o più gruppi al livello di significatività scelto. Quindi, una volta che lo studio è stato concluso e la differenza è stata osservata al livello di significatività prefissato, la questione della potenza è di minore rilevanza. Tuttavia, la potenza può anche essere definita come la percentuale di studi condotti su campioni indipendenti estratti dalla stessa popolazione che darà una differenza statisticamente significativa (nei limiti prefissati

di P) se l'effetto del trattamento è vero. Estendendo questa definizione fino a comprendere, negli *n* campioni estratti indipendentemente dalla popolazione, tutta la popolazione, allora la potenza dello studio è una stima, per quanto grossolana, della proporzione di popolazione che può "rispondere" al trattamento come si è visto nello studio purché il campione sia davvero rappresentativo della popolazione stessa. Per questo motivo la forza delle evidenze esterne utilizzate in queste linee guida tiene conto anche della potenza degli studi.

Bias, fattori di confusione, stratificazione

Quando si discute di valore probante di uno studio clinico agli effetti di una linea guida non si può trascurare di valutare quanto i risultati dello studio possono essere spostati rispetto alla realtà dei fatti.

Uno scostamento dei risultati osservati dai risultati "veri" può essere dovuto a diversi fattori, primo fra tutti il fatto che il campione studiato in realtà rappresenta solo una parte della popolazione reale (p.e. a causa di criteri di inclusione e di esclusione peraltro spesso indispensabili a garantire la necessaria sicurezza ai partecipanti allo studio).

Una deviazione sistematica dei risultati osservati dal risultato "vero" si definisce "*bias*". Di solito il *bias* tende a deviare, in maniera quasi sempre non voluta, i risultati osservati verso esiti maggiormente favorevoli all'intervento in esame rispetto all'esito vero. Il *bias* si può manifestare in molti modi: all'arruolamento (p.e. un clinico conservatore assegnerà i casi tendenzialmente più gravi al trattamento di riferimento mentre uno "progressista" li assegnerà al trattamento nuovo); durante la conduzione dello studio (p.e. gestendo in maniera differente i soggetti assegnati all'uno o all'altro trattamento); e alla valutazione dei risultati (p.e. valutando sistematicamente in maniera più favorevole i risultati ottenuti con l'uno o con l'altro trattamento). Questi *bias* possono essere contenuti utilizzando le normali tecniche di randomizzazione (assegnazione del trattamento su base casuale) e di cecità (evitando che chi assegna il trattamento e chi ne valuta i risultati sia a conoscenza dell'effettivo trattamento assegnato al singolo paziente). Anche in questo caso le procedure di randomizzazione e di cecità devono essere tali da impedire di fatto, e non solo formalmente, la conoscenza del trattamento assegnato. Pertanto la valutazione di queste tecniche non si limita a verificare che siano scritte, ma entra nel dettaglio della valutazione di come sono state applicate e quindi se siano da considerarsi valide o meno. Uno studio per il quale non si potesse stimare se le tecniche di randomizzazione e cecità erano valide, è stato considerato come *bias*ed (cioè affetto da errore sistematico) in queste linee guida.

Tuttavia bisogna considerare un'altra fonte di potenziale errore sistematico, costituita dal campione esaminato rispetto al campione arruolato. È infatti raro che tutti i soggetti arruolati nello studio giungano alla fine dello stesso. I casi di interruzione dell'osservazione possono quindi essere una fonte importante di *bias*, specie se costituiscono una proporzione importante (>5%) del campione arruolato. Infatti, in tal caso non è detto che i soggetti che completano lo studio rappresentino la stessa popolazione rispetto ai soggetti che erano stati arruolati, per cui le conclusioni non sono più applicabili a tale popolazione nonostante i soggetti arruolati fossero rappresentativi. Inoltre, bisogna considerare che se una certa proporzione di soggetti abbandona lo studio, è presumibile che una simile (ma di solito maggiore) proporzione di soggetti abbandonerà il trattamento nella realtà clinica. Quindi la stima dell'effetto di una procedura deve tenere conto di questo fenomeno. Ciò si ottiene con l'analisi dei dati ITT (*intention-to-treat*).

Indipendentemente dalla definizione data a questa procedura, il principio è semplice: i risultati devono dare conto di tutti i soggetti arruolati che non abbiano fisicamente rifiutato il trattamento prima della prima dose, ovvero, N soggetti hanno almeno teoricamente iniziato il trattamento, N risultati devono essere valutati. Naturalmente per i soggetti che non hanno concluso l'osservazione andranno fissate (pre-fissate, ovvero definite prima dell'inizio dello studio) regole precise per determinare l'esito. Le regole più sicure sono quelle conservative (o *worst-case*) secondo le quali i soggetti che non terminano lo studio sono per definizione classificati come fallimento oppure sono assegnati al risultato peggiore osservato tra tutti quelli che hanno concluso lo studio. Con questo principio si evita di favorire indebitamente il trattamento in esame rispetto al trattamento di riferimento. Naturalmente si possono condurre analisi di sensibilità per diverse regole di assegnazione, valutando eventuali discrepanze sul piano del valore clinico, ma le analisi di sensibilità hanno rilevanza solo dopo che l'analisi *worst-case* ha confermato che un effetto esiste.

Il *bias* però non è la sola fonte di scostamento tra risultato vero e risultato osservato. Possono esistere altri fattori, definiti come "fattori di confusione", che hanno lo stesso effetto. I fattori di confusione sono variabili non considerate nello studio, che influenzano variabili che sono invece considerate nello studio. Non essendo considerati, i fattori di confusione possono esercitare un'influenza non quantificabile che sfugge all'osservazione. Non esistono tecniche semplici per evitare l'effetto di fattori di confusione. Certamente, la migliore risposta a questo rischio è data da una stima adeguata della numerosità campionaria e dalla randomizzazione e cecità, tale per cui eventuali deviazioni attribuibili a potenziali fattori di confusione siano uniformemente distribuite sui due o più gruppi esaminati. D'altra parte, esistono anche fattori di confusione prevedibili (p.es. età, sesso, eziopatogenesi della patologia

considerata, scolarità in certi casi ...) che quindi possono essere registrati e per i quali è possibile apportare una correzione in sede di analisi. Pertanto, in sede di pianificazione dello studio deve essere condotta un'accurata valutazione dei potenziali fattori di confusione noti o presumibili, così da registrarli e successivamente tenerne conto in analisi multivariate o analisi di sensibilità dei risultati. Resta il fatto che se il campione è di dimensioni adeguate ed adeguatamente randomizzato l'influenza dei fattori di confusione è marginale rispetto ai *bias* discussi in precedenza.

Esistono poi situazioni in cui fattori di confusione distribuiti in maniera asimmetrica nella popolazione possono esercitare un peso rilevante sui risultati (p.es. una condizione presente in una piccola frazione della popolazione che però modifica completamente la risposta a un certo trattamento). In tal caso è preferibile ricorrere a tecniche di stratificazione del campione, così che i vari strati siano tutti adeguatamente rappresentati e si possa stimare la risposta al trattamento della popolazione dopo correzione per l'effetto del fattore. Anche questo approccio può, tuttavia, condurre a un errore sistematico. Infatti la stratificazione permette una stima più accurata dell'effetto del trattamento tramite analisi multivariata con il fattore di confusione, ma di per sé non permette di valutare l'effetto del fattore di confusione. Per potere stimare allo stesso tempo anche la correlazione tra livello del fattore di confusione ed esito clinico considerato, bisogna considerare questo come *endpoint* separato dello studio, e quindi aumentare conformemente la numerosità campionaria, allo stesso modo di un'analisi per sottogruppi.

In queste linee guida si è attribuito maggiore livello di evidenza agli studi per i quali erano chiaramente indicate e valutabili come valide le misure prese per la riduzione del *bias* e per considerare di limitata rilevanza l'impatto di potenziali fattori di confusione.

Effetto di farmaco e effetto di classe

Normalmente gli studi clinici sono effettuati con singoli farmaci, mentre le metanalisi tendono ad aggregare farmaci appartenenti alla stessa classe terapeutica. Gli effetti stimati da uno studio clinico possono quindi evidenziare ciascuno l'effetto di un singolo farmaco mentre egli effetti stimati da una metanalisi tendono ad evidenziare l'effetto complessivo di una classe terapeutica di farmaci presumibilmente omogenei.

I suggerimenti relativi ad una certa modifica di pratica clinica che derivano da un insieme coerente di studi clinici sono quindi normalmente applicabili ad un singolo farmaco e non sono estendibili ad altri farmaci anche se appartenenti alla stessa classe terapeutica e presentano (apparentemente) lo stesso meccanismo d'azione. Ciò potrebbe essere superato, con estensione delle conclusioni a un'intera classe terapeutica (o a uno specifico sottoinsieme di questa) qualora diversi insiemi coerenti di studi clinici condotti ciascuno con un diverso farmaco appartenente alla stessa classe, forniscano risultati sostanzialmente coerenti tra loro e privi di *bias* importanti. Tuttavia, gli effetti di classe sono generalmente meglio valutati per mezzo di adeguata revisione sistematica e metanalisi con stima quantitativa dell'eterogeneità.

I suggerimenti derivanti invece da una metanalisi sono di solito da considerarsi validi per un'intera classe senza distinzione – quindi senza preferenza – per un qualsiasi farmaco appartenente a quella classe. All'interno di una classe possono comunque essere eseguite stime per sottogruppi di studi. A questo proposito, tuttavia, vanno considerate tutte le cautele relative alle analisi di sottogruppi. Inoltre, non è sufficiente determinare lo scostamento dell'effetto di un sottogruppo dall'effetto globale né il confronto dell'effetto di un sottogruppo di studi rispetto agli altri studi raggruppati, perché in questo caso si tende a concentrare tutto l'effetto osservato nel solo fattore "gruppo di studi", senza tenere conto delle influenze di tutti gli altri fattori di variabilità. Si possono eseguire, per ridurre l'impatto fuorviante di tale concentrazione di variabilità, o analisi di sensibilità, oppure analisi di interazione tra effetto primario considerato e classificazione degli studi in diversi sottogruppi. In entrambi i casi, tuttavia – e in particolar modo quando esistano forti sbilanciamenti di numerosità tra sottogruppi –, è bene esaminare insieme alla eventuale significatività dell'effetto osservato, anche l'intervallo di confidenza dell'effetto e la misura quantitativa di eterogeneità.

In queste linee guida ci si è prevalentemente attenuti ai risultati di metanalisi per la stima di eventuali effetti di classe, e ai risultati di insiemi omogenei di studi per la stima di eventuali effetti di farmaco.

Significatività clinica e significatività statistica

Se la definizione di significatività statistica è chiara, quella di significatività clinica lo è molto meno.

La significatività statistica (sintetizzata nel valore numerico di "P") indica esclusivamente la probabilità che l'effetto osservato possa essere osservato casualmente nel caso sia vera l'ipotesi zero in base alla quale è stato progettato lo studio. Se il valore calcolato è più piccolo di quello predefinito in sede di progettazione dello studio è lecito respingere l'ipotesi zero. Si noti che in tale definizione non è minimamente vincolante l'ipotesi fatta né il livello critico di probabilità. Quindi, affermare che esiste una significatività statistica non permette di trarre nessuna conclusione se non è chiaramente definito il livello

critico prefissato di probabilità e non è chiaramente esplicitata l'ipotesi zero da respingere. Nella maggior parte dei casi l'ipotesi di partenza è che non vi siano differenze di effetto fra due interventi e il livello critico prefissato è il 5%. In tali casi, quindi, ottenere un risultato statisticamente significativo indica che si può respingere con il 95% di confidenza l'ipotesi che i due interventi abbiano lo stesso effetto nella popolazione rappresentata dal campione esaminato.

Il problema critico agli effetti di raccomandazioni utili per la pratica clinica è quale possa essere l'impatto sui malati di un risultato così espresso. Per stimare tale impatto è più ragionevole considerare invece l'intervallo di confidenza della differenza di esito osservata, espressa in termini di numero di pazienti da trattare. Inoltre un risultato isolato non permette di trarre conclusioni definitive, e un insieme di risultati è meglio stimato con una revisione sistematica e una metanalisi che forniscano una stima dell'effetto con il suo intervallo di confidenza e una stima quantitativa di eterogeneità.

Ciò non significa che sia indispensabile attendere il risultato di numerosi studi e della relativa metanalisi per apportare modifiche importanti alla pratica clinica. Possono esistere situazioni –peraltro piuttosto rare – in cui l'effetto osservato è di tale rilevanza agli effetti della tutela della salute dei malati, che anche sulla base di un unico studio è opportuno modificare la pratica clinica (p.es. per riduzione sostanziale di mortalità osservata in uno studio con *bias* trascurabile, piccolo NNTB e grande NNTH). Tuttavia ciò non esime dal cercare conferma dei risultati ottenuti – non necessariamente con altri studi randomizzati e controllati – né esime dalla valutazione accurata caso per caso del rapporto tra benefici presumibili e rischi prevedibili (cioè dalla considerata valutazione clinica del singolo caso alla luce dei dati disponibili). In ogni caso il grado della raccomandazione relativa rimane inferiore rispetto ai risultati confermati da insiemi di studi omogenei o metanalisi con scarsa eterogeneità, senza tuttavia ridurre l'importanza pratica dell'effetto.