Sopravvivere tra numeri e statistica

Renato Rossi

Ovvero tutto quello che avreste voluto sapere e non avete mai osato chiedere

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Indice

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Premesse

Perché questo libro? Mi sono deciso a scriverlo dopo aver constatato che la maggior parte dei medici non ama molto la statistica e la matematica, forse a causa di poco piacevoli ricordi scolastici. Non sono un esperto di statistica per cui ho usato un linguaggio semplice e facilmente comprensibile, per prima cosa a me stesso, cercando di ridurre al minimo l'uso dei numeri (verso i quali molti hanno una preclusione naturale) e gli esperti troveranno probabilmente queste pagine elementari. Anzi diciamo subito che un esperto di statistica sicuramente mi boccerebbe senza prendersi la briga di rimandarmi a settembre tanti sono gli errori e addirittura le imprecisioni di linguaggio da un punto di vista tecnico. Ma non preoccupatevi, noi siamo medici e non matematici e l'importante è la sostanza del discorso e non l'abito.

Perciò se il lettore avrà tratto anche solo l'impulso a meglio conoscere e comprendere i meccanismi che stanno alla base degli studi clinici, il mio scopo sarà raggiunto. Molte delle nozioni che seguiranno derivano dalla lettura, nel corso degli anni, di vari editoriali e commenti apparsi nelle maggiori riviste internazionali, dalla consultazione del sito EQM (Evidenza, Qualità e Metodo: http://www.evidenzaqualitametodo.it) che contiene una serie di articoli molto approfonditi sulle tematiche relative alla interpretazione degli studi clinici, infine da una serie di colloqui avuti con il dr. Alessandro Battaggia e con il dr. Fausto Bodini, che da anni si occupano di EBM e che pubblicamente ringrazio.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 1
L’inganno dell’esperienza

La conoscenza della letteratura è sempre più indispensabile al medico per esercitare la sua professione. Parafrasando Osler possiamo dire che un medico che pretendesse di curare i malati senza conoscere la letteratura scientifica è come il capitano di una nave che si barcamena in un oceano sconfinato senza conoscere le previsioni meteorologiche.

Tuttavia la mole di lavori pubblicati ogni anno rende impossibile la conoscenza di tutto quello che viene scoperto: nell’epoca della massima diffusione del sapere il medico rischia l’effetto inondazione. E’ indispensabile quindi avere dimestichezza con un metodo generale che serva a filtrare le informazioni ritenendo quelle importanti e cestinando le altre.

Ma quali sono i mezzi che noi abbiamo a disposizione per determinare se un farmaco (o più in generale un intervento terapeutico) è efficace o non lo è? In quale modo possiamo capirlo? Non potrebbe bastare l'esperienza clinica?

Qualcuno potrebbe dirmi: in fin dei conti io faccio questo lavoro da 20-30 anni, possibile che non serva a niente tutto il sapere e la fatica che ho accumulato nel corso degli anni?

Non voglio dire che l'esperienza clinica sia inutile, anzi essa è importantissima nel processo diagnostico ed in quello di gestione globale del paziente. Non occorre che mi dilunghi oltre su questo punto: non credo necessiti di dimostrazioni. Quello che voglio sostenere in queste note però è un'altra cosa, vale a dire che **l'esperienza ci può ingannare** quando dobbiamo giudicare della efficacia di un nostro intervento. Un vecchio medico di famiglia si divertiva a raccontare la seguente storiella, che però lui giurava essere vera. In una famiglia di contadini si ammala il nonno, ormai avanti negli anni, di una tosse persistente che non se ne voleva andare, per cui pensano bene di chiamare il medico. Costui arriva come vuole la tradizione a bordo di una malandata automobile dopo aver attraversato una strada polverosa immersa tra i campi di frumento, visita il nonno, sentenzia che si tratta di una semplice tracheite e prescrive uno sciroppo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

per la tosse, raccomandando alla figlia del vecchio contadino malato di somministrarne regolarmente un cucchiaio ogni otto ore. "Già che c'è, dottore, dia un'occhiata anche al mio bambino, che da qualche giorno mangia poco e si lamenta di mal di pancia". Il buon medico visita anche il piccino, tranquillizza la madre che si tratta di semplice indigestione e prescrive un farmaco in supposta, da somministrare per via rettale ogni dodici ore. Passano circa due settimane e un mercoledì, al mercato, il dottore incontra la figlia del vecchio contadino e si informa di come va il padre. "Benissimo" risponde la donna "quelle supposte che gli avete prescritto gli hanno fatto benissimo, è guarito in pochi giorni, anche se a dire la verità è stata una faticaccia a fargliele accettare. Anche mio figlio è guarito, lo sciroppo ha fatto miracoli". Questo aneddoto dimostra una cosa che ogni medico sul campo tocca spesso con mano: "**la guarigione non significa nulla**" e in medicina non è vero il detto "post hoc ergo propter hoc". L'esperienza di ciascuno di noi è sicuramente importante per il lavoro di tutti i giorni, ma non può essere la base per poter giudicare dell'efficacia di un farmaco o di un trattamento. Per poterlo fare abbiamo bisogno di un altro metodo, e questo metodo è rappresentato dagli studi clinici. E' evidente infatti che l'esperienza non può venirci in aiuto quando si tratta di **farmaci che non abbiamo mai usato** perché appena immessi in commercio. In questi casi su cosa dobbiamo basarci? Chiedere ai nostri amici informatori del farmaco non è molto realistico. Sarebbe come domandare all'oste se il vino che vende è buono. Quale oste direbbe che quello venduto nell'enoteca accanto è migliore?

Ma anche per i farmaci che usiamo da più tempo i nostri sensi finiscono con il giocarci brutti scherzi. Mi spiego con alcuni esempi che renderanno subito chiaro quello che voglio dire.

Casistica limitata

Supponiamo di avere a disposizione quattro farmaci per abbassare la pressione, il farmaco A, B, C, D. Supponiamo anche che dopo 5 anni di trattamento si verifichino i seguenti casi:

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

- con il farmaco A si hanno 40 infarti ogni 1000 pazienti trattati

- con il farmaco B si hanno 50 infarti ogni 1000 pazienti trattati

- con il farmaco C si hanno 40 infarti ogni 1000 pazienti trattati

- con il farmaco D si hanno 30 infarti ogni 1000 pazienti trattati

Mediamente un medico ha in carico 300 soggetti ipertesi, e sempre per pura probabilità, ipotizziamo che i pazienti siano così distribuiti:

- 75 assumono il farmaco A

- 75 assumono il farmaco B

- 75 assumono il farmaco C

- 75 assumono il farmaco D

In base a quanto detto il medico avrebbe la possibilità di registrare, dopo 5 anni:

- 3 infarti nel gruppo che assume il farmaco A

- 3,75 infarti nel gruppo che assume il farmaco B

- 3 infarti nel gruppo che assume il farmaco C

- 2,25 infarti nel gruppo che assume il farmaco D

Risulta chiaro senza bisogno di ulteriori spiegazioni che in base alla sua casistica il medico non potrebbe accorgersi della diversa efficacia dei quattro farmaci nel ridurre il rischio di infarto, perché imputerebbe la differenza, così piccola, al caso, mentre sappiamo in realtà che il farmaco B aumenta il rischio del 25% rispetto ad A e C e il farmaco D riduce lo stesso rischio del 25%, sempre rispetto ad A e C.

In questo primo esempio la piccolezza del campione trattato porta fuori strada il medico il quale sarà portato a ritenere che i quattro farmaci siano grosso modo equivalenti.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 1 – L’inganno dell’esperienza

Selezione dei pazienti

Supponiamo di avere a disposizione due farmaci per curare l'ipertensione, il primo farmaco (farmaco A) è noto per essere efficace ma provoca un effetto collaterale abbastanza fastidioso e cioè disfunzione erettile; il secondo farmaco (farmaco B) è pure esso efficace e provoca disfunzione erettile in una percentuale di casi inferiore, ma è poco maneggevole negli anziani perché può accentuare i problemi di memoria.

Può succedere quindi che se devo trattare un iperteso giovane (in cui la disfunzione erettile è più disturbante) sarò portato ad usare il farmaco B mentre se devo trattare un anziano userò più probabilmente il farmaco A. Se, dopo un certo numero di anni, andassi a controllare i miei ipertesi troverei probabilmente che chi prende il farmaco A va incontro ad una percentuale di infarto o ictus superiore a quella di chi prende il farmaco B. Erroneamente sarei portato a ritenere che il farmaco A è meno efficace del farmaco B a ridurre le complicanze dell'ipertensione (infarto e ictus). In realtà la mia analisi è **viziata** all'inizio per il fatto di aver somministrato prevalentemente il farmaco A a soggetti più anziani e quindi di per sé più propensi a sviluppare le complicanze della malattia ipertensiva rispetto ai più giovani. Al contrario ho somministrato il farmaco B a pazienti più giovani, di per sé meno soggetti ad avere le complicanze dell'ipertensione.

In gergo tecnico si dice che vi è un "**bias di selezione**". Vedremo meglio in seguito di che cosa si tratta. L'esempio che ho fatto è abbastanza grossolano, ne vedremo altri presi da studi clinici veri, ma per il momento è importante capire che la mia esperienza in questo caso può portarmi a conclusioni errate perché, in modo più o meno inconscio, io "seleziono" i pazienti da trattare.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 1 – L’inganno dell’esperienza

Evoluzione naturale della malattia

Le infezioni delle alte vie respiratorie sono molto frequenti e quasi sempre di natura virale, hanno un decorso autolimitato a 7-10 giorni (in genere) e, ovviamente, gli antibiotici sono inutili. Prendiamo ora un giorno qualsiasi di una giornata invernale e vediamo cosa succede in un affollatissimo ambulatorio: un medico sta visitando un paziente che lamenta da 3-4 giorni tosse, raucedine, mal di gola e scolo nasale. Dopo aver visitato scrupolosamente il malato il medico arriva alla conclusione di essere di fronte ad una banale virosi respiratoria e prescrive una terapia sintomatica. Dopo tre giorni il paziente richiede una nuova visita perché la tosse e il raffreddore persistono e la terapia si è dimostrata inefficace. Il paziente chiede al medico se non sia il caso di assumere un antibiotico. Il curante, un po’ perché teme di aver sottovalutato il quadro, un po’ perché non vuol entrare in contrasto con le richieste dell'assistito, finisce con accondiscendere e ordina un antibiotico. Dopo tre giorni il paziente guarisce. E' stato l'antibiotico? Probabilmente no, semplicemente l’infezione virale ha fatto il suo naturale decorso. Eppure nell'immaginario del paziente e nell'esperienza del medico si fa strada l'idea erronea che l'antibiotico serva nelle tracheo-bronchiti. Al contrario studi clinici effettuati con il sistema della randomizzazione e del doppio cieco (vedremo in seguito cosa significano queste buffe espressioni) hanno dimostrato l'opposto, e cioè che nelle flogosi respiratorie l'antibiotico, di solito, è inutile.

L’effetto placebo

Molti anni fa frequentavo il reparto di Medicina di un piccolo ospedale periferico. C'era un paziente affetto da asma bronchiale che si ricoverava con la precisione di un orologio svizzero ad ogni riacutizzazione per "farsi una flebo" di aminofillina e cortisonici. Una mattina eravamo tutti in

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

riunione nello studio del primario quando entra trafelata un' infermiera, tutta preoccupata, a riferirci che aveva messo la flebo di fisiologica all'asmatico (così lo conoscevano in reparto), che guarda caso si era ricoverato proprio quella mattina, flebo ordinata dal medico di guardia, ma poi si era dimenticata di iniettarci dentro i farmaci perché chiamata d'urgenza per un altro paziente. Come un sol uomo ci precipitiamo dall'asmatico e lo troviamo sorridente che tiene una quasi conferenza ai suoi vicini di letto, la flebo ormai del tutto consumata. Il primario gli chiede come sta e lui, serafico, risponde che sta benissimo e che anche questa volta "la flebo miracolosa" l'aveva vinta sulla sua malattia. Naturalmente non è neppure ipotizzabile che la semplice acqua possa essere efficace nella cura dell'asma ma la cosa si spiega se si tiene conto che si trattava di un paziente molto emotivo e suggestionabile e dell' **effetto placebo** dovuto alla flebo stessa e a tutta la "messa in scena" che la circonda (boccione, tubicini, ago infilato nel braccio, ecc.). Un osservatore esterno avrebbe potuto però trarre l' errata conseguenza che la semplice acqua di rubinetto è una cura per l'asma.

Cosa fare se l’esperienza ci inganna?

Ma allora, mi direte, come possiamo sapere se un farmaco funziona di più e meglio dell'acqua di rubinetto? Come possiamo eliminare l'effetto placebo? Come possiamo uscirne se l'esperienza, anche la più vasta, può giocarci di questi brutti scherzi? Cosa dobbiamo fare?

La scienza ha elaborato dei metodi per oggettivare (o meglio per cercare di rendere oggettivo) l'effetto di un farmaco o di un intervento. Questi metodi si chiamano **studi clinici**. In altre parole la medicina ha cercato di darsi una dignità scientifica fondata su di un metodo che sia oggettivo e ripetibile e che la possa, in qualche modo, togliere dalla volubile soggettività dell'operatore. Non è tutto oro quello che luce e ovviamente la medicina è una scienza sui generis diversa dalla

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

matematica: non sempre in medicina 2 + 2 = 4, qualche volta può essere 3 o 5. Tuttavia lo sforzo intrapreso in questi ultimi decenni ha permesso di costruire un corpus di conoscenze notevoli. Le quali però, e questo è un punto importantissimo che non va mai dimenticato, valgono a livello di popolazioni e di numeri statistici, ma vanno poi applicate, **tagliate su misura,** per ogni singolo paziente che ci sta davanti. In altre parole i risultati degli studi vanno **"trasferiti"** nella pratica del mondo reale e in quel particolare paziente. Ma avremo modo di tornare con più calma su questi aspetti.

Riprendiamo invece il discorso sugli studi clinici che, come abbiamo detto, sono il metodo che la medicina si è data per avere pure essa, al pari di altre branche del sapere, una veste scientifica. Quanti tipi di studi esistono?

Per fare le cose molto semplici diremo che ne esistono di due tipi. Gli studi del primo tipo, detti anche **studi di intervento**, sono noti come studi randomizzati e controllati (o anche con la sigla inglese RCT: randomized controlled trial); quelli del secondo tipo sono detti **studi osservazionali.** In che cosa si distinguono, come fare per capire se si tratta di studi del primo o del secondo tipo, quali sono i loro pregi e i loro difetti e altre cose sarà argomento dei prossimi capitoli.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 2
Come analizzare gli studi

Supponiamo di avere tra le mani una prestigiosa rivista internazionale di medicina e di metterci a sfogliarla. Ad un certo punto ci imbattiamo in un articolo che parla di un farmaco usato per trattare la tal o tal'altra condizione patologica. Si tratta di uno studio importante, vale la pena di ritenerlo oppure possiamo tranquillamente passare oltre? I suoi risultati sono affidabili e rilevanti per la mia pratica professionale?

Vediamo di procedere con ordine. Per prima cosa dobbiamo stabilire di che studio si tratta. Abbiamo visto che esistono sostanzialmente due tipi di studi: gli *studi sperimentali* e quelli *osservazionali*. Nei primi i ricercatori **somministrano** il trattamento oggetto dello studio, nei secondi il trattamento viene scelto dai medici curanti, dai pazienti o altro ma non dai ricercatori, che si limitano solamente a **osservare** quello che accade , che accadrà o che è già avvenuto. Quindi rispondendo ad una semplice domanda si può subito determinare se si tratta di uno studio di intervento o di uno studio osservazionale: ***gli sperimentatori hanno somministrato il trattamento oggetto dello studio?*** Se la risposta è positiva sarà uno studio di intervento, se negativa avremo di fronte uno studio osservazionale.

Vediamo più in dettaglio questo punto con due semplici esempi, ricavabili già leggendo l'abstract dello studio.

Primo abstract:
"Abbiamo reclutato 5000 pazienti a cui abbiamo somministrato aspirina o placebo. Dopo 5 anni siamo andati a vedere quanti infarti c'erano nel gruppo che assumeva aspirina e quanti in quello che assumeva placebo.

Secondo abstract:
"Abbiamo studiato 5000 soggetti iscritti al registro Amanti dello Scopone Scientifico della regione Tal dei Tali e abbiamo chiesto loro se prendevano aspirina o se non la prendevano. Dopo 5 anni siamo andati a vedere quanti infarti c'erano nel gruppo che prendeva aspirina e in quello che non la prendeva".

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 2 – Come analizzare gli studi

Di che tipo è il primo studio? La risposta è semplice : si tratta di uno studio di intervento perché la somministrazione di aspirina o di placebo è stata decisa dagli autori stessi dello studio. Al contrario il secondo studio è di tipo osservazionale perché gli autori dello studio non hanno deciso il trattamento ma si sono limitati a fotografare le cose come stavano e poi ad andare a vedere cosa succede dopo 5 anni. In altre parole l'assunzione di aspirina è stata decisa dai medici curanti o dai pazienti stessi in seguito alla lettura di una di quelle riviste per la salute o alla visione di una rubrica televisiva o dietro consiglio di amici o del barbiere (senza offesa per i barbieri ovviamente, nei cui saloni anzi spesso e volentieri vengono date delle ottime raccomandazioni sanitarie). In ogni caso **NON** sono stati gli autori dello studio a decidere il trattamento, essi si sono limitati a prendere atto, come dei semplici notai, di quello che stava avvenendo. Perché è importante distinguere tra studi di intervento e studi osservazionali? Perché gli studi osservazionali sono gravati da possibili *bias* che ne inficiano i risultati: essi possono essere utili in quanto permettono di generare delle ipotesi che dovranno però in seguito essere confermate da studi di intervento. Non è che gli RCT siano perfetti, come cavalieri senza macchia e senza paura, possono avere i loro bias e le loro debolezze, come vedremo più avanti. Se non capite cosa vuol dire bias non è il caso di farne un dramma. Fra poche pagine sarà tutto più comprensibile. Per ora è importante aver chiaro il concetto che gli studi osservazionali producono, in generale, dei risultati che sono meno affidabili di quelli di uno studio randomizzato e controllato, purché questo sia fatto con tutti i crismi che la procedura comanda. Bene, mi direte, abbiamo stabilito che siamo di fronte ad uno studio di intervento (qualche volta questi studi vengono anche detti sperimentali). E ora? Cari amici, siamo appena all'inizio del nostro esame e la strada è ancora lunga, ma cercherò, per lo meno, che non sia anche in salita altrimenti alla fine corro il rischio di ritrovarmi da solo. Stabilito che siamo di fronte ad uno studio sperimentale, vi sono vari punti da prendere in esame.

E' uno studio controllato? E' stata eseguita una randomizzazione? Lo studio è in cieco? Quali sono

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

gli end-point esaminati? Quanto è numeroso il campione arruolato e quanto è durato lo studio? Anche qui non spaventatevi se non riuscite a capire alcuni (o tutti) i termini usati. Alla fine vedrete che si tratta di una banalità al confronto dei problemi che dobbiamo affrontare ogni giorno.

Esiste un gruppo di controllo?

La prima domanda che ci dobbiamo porre a questo punto è se esiste un gruppo di controllo. Che cosa vuol dire questo?

Vuol dire che i pazienti arruolati (si dice anche reclutati ed è curioso l'uso di questi due termini che richiamano il linguaggio militare, quasi la partecipazione ad uno studio sia un obbligo di leva) sono stati divisi in due gruppi, a uno dei quali viene somministrato il farmaco mentre all'altro un placebo o un farmaco di confronto.

Se la risposta alla nostra domanda è positiva si tratta di uno studio sperimentale controllato, altrimenti abbiamo di fronte uno studio sperimentale non controllato. Per la verità la stragrande maggioranza degli studi pubblicati è di tipo controllato, mentre molto rari sono gli studi non controllati. Uno studio non controllato è uno studio in cui non esiste un gruppo di controllo. È lapalissiano. Un esempio potrebbe essere uno studio che volesse valutare l'efficacia dell'intervento chirurgico nella rottura degli aneurismi dell'aorta addominale. E' evidente che non si può usare un gruppo di controllo per motivi di tipo etico, in quanto un aneurisma aortico che si rompe, se lasciato a se stesso, porta a morte certa, per quanto mi risulta. In questo caso si dà per scontato che nel gruppo di controllo, non trattato, la mortalità sia del 100%: se l'intervento chirurgico porterà la mortalità al 75% possiamo dire che abbiamo ottenuto una riduzione del 25%. Eventualmente si potrebbe prevedere il confronto tra due diversi tipi di intervento chirurgico, per vedere quale è più efficace e se, in futuro, si rendesse disponibile un trattamento medico che riuscisse a riparare la parete aortica che si sta rompendo (come si fa con certe bombolette che iniettano dentro lo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

pneumatico bucato una sostanza espandente che lo gonfia e lo ripara) si potrebbe pensare a confrontare l'intervento chirurgico con una terapia alternativa non chirurgica.

In linea generale si può dire che al gruppo di controllo viene somministrato placebo, oppure un farmaco di riferimento quando ragioni etiche impediscono l'uso del placebo. Per esempio se si volesse determinare l'efficacia di una nuova statina negli infartuati, questa dovrebbe essere confrontato con una statina (farmaco che ha dimostrato di ridurre la mortalità in questa tipologia di pazienti): non sarebbe etico usare un placebo. Nei primi studi sulle statine negli infartuati queste vennero paragonate al placebo semplicemente perché non era nota la loro efficacia (si poteva anche speculare che fossero pericolose o inutili ed anzi è successo più di una volta che un farmaco si è dimostrato dannoso). Ma ora che la loro efficacia è stata ampiamente accertata è possibile organizzare solo studi che confrontino due statine diverse tra loro oppure una stessa statina a dosaggi diversi, per stabilire quale sia la dose più appropriata.

Abbiamo visto quindi che ci sono due gruppi di pazienti; il gruppo a cui viene somministrato il farmaco da studiare si chiama "**braccio intervento**", il gruppo a cui viene somministrato il placebo o il farmaco di confronto si chiama "**braccio controllo**". Per comodità continuerò a parlare di farmaci, ma è evidente che l'intervento può essere anche di tipo diverso, per esempio chirurgico, psicologico, o altro (screening, agopuntura, omeopatia). Un' annotazione circa gli studi che volessero determinare l'efficacia dell'agopuntura, che ci fa capire quanto debbano essere attenti gli autori nel loro disegno, pena essere sottoposti ad una valanga di critiche da parte degli esperti che lo valuteranno dopo la pubblicazione. Supponiamo per esempio di voler determinare se l'agopuntura funziona nella profilassi dell'emicrania meglio di un farmaco (per esempio un beta-bloccante). Allora arruolo il mio campione di pazienti e lo divido in due gruppi. Ovviamente al gruppo intervento faccio l'agopuntura e a quello di controllo somministro il beta-bloccante. Attenzione però, se facessi così non eliminerei l'effetto placebo (possibile) degli aghi. Se voglio

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

fare uno studio come le regole comandano al gruppo intervento devo fare l'agopuntura e somministrare contemporaneamente una pasticca di placebo che simuli il beta-bloccante, mentre al gruppo di controllo devo dare la pasticca di beta-bloccante e nello stesso tempo fare delle sedute di agopuntura simulata (in inglese sham acupuncture), vale a dire infilare degli aghi senza rispettare le zone dell' agopuntura tradizionale cinese. Solo in questa maniera i due gruppi saranno pienamente confrontabili.

Da quanto si è detto finora si capisce l'importanza del braccio di controllo: essa sta nel fatto che altrimenti è impossibile giudicare se i miglioramenti ottenuti dal trattamento sono dovuti alla evoluzione naturale della malattia, a un effetto placebo o a qualcosa d'altro che non conosciamo, come abbiamo visto negli esempi a proposito dei trabocchetti che ci può riservare l'esperienza. Un mio amico che di queste cose ne mastica più di me mi diceva: "Ricordati che chi corre da solo arriva sempre primo. Solo se gareggi con gli altri saprai il tuo valore reale, se sei un "figlio del vento" come Lewis (campione americano di corsa veloce) o un povero brocco". Un accenno soltanto alle tecniche con cui vengono reclutati i pazienti negli studi perché esse possono influire sui risultati finali o comunque sulla loro trasferibilità pratica. Per esempio è abbastanza frequente prevedere una prima fase di “epurazione” delle liste, se mi si passa questo termine. Così in uno studio che vuol valutare l’efficacia di una statina nel ridurre l’infarto si prevede una prima fase di otto settimane in cui tutti i soggetti arruolati assumono la statina e vengono poi scartati quelli in cui non si riesce ad ottenere una riduzione del colesterolol LDL di almeno il 20%. E’ ovvio che facendo in questo modo “seleziono” per lo studio i pazienti che rispondono meglio al trattamento e che potrebbero essere non sovrapponibili a quelli visti nella pratica. Altri esempi: in uno studio su un farmaco antipertensivo si prevede una prima fase in cui vengono selezionati e scartati i pazienti in cui il trattamento non produce una diminuzione della pressione di almeno il 10%; oppure in una prima fase si scartano tutti i pazienti che non assumono almeno l’80% delle dosi del farmaco

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

prescritto: è comprensibile come in questo modo vengano esclusi i pazienti poco complianti perché in essi il farmaco in esame produce troppi effetti collaterali oppure perché hanno delle forme morbose associate che impediscono o rendono difficile l'assunzione del farmaco.

Esistono vari modi di reclutare i pazienti. Uno è il cosiddetto "**opt-out**" in cui i soggetti vengono invitati, per esempio con una lettera o una telefonata, a partecipare al trial e se non rispondono i ricercatori possono contattarli nuovamente; un metodo alternativo invece è quello detto "**opt-in**" in cui, se il paziente non risponde al primo invito si presume che non voglia partecipare e non si può contattarlo di nuovo. E' stato dimostrato che i partecipanti arruolati secondo la tecnica "opt-in" sono in genere più sani di quelli reclutati secondo la strategia "opt-out" e la cosa appare anche logica. Si capisce però che i pazienti selezionati secondo la tecnica "opt-in" possono non essere rappresentativi dei pazienti reali.

Infine si possono prevedere dei criteri di inclusione e di esclusione così rigidi che ben difficilmente i pazienti corrispondono a quelli "veri" che vediamo tutti i giorni nei nostri ambulatori. Se per esempio in uno studio su un farmaco per il trattamento dello scompenso cardiaco si escludono tutti i soggetti con più di 65 anni, i diabetici, i nefropatici, quelli con BPCO, c'è da domandarsi a quale paziente reale possano poi applicarsi i risultati ottenuti. D'altra parte avendo una serie di RCT limitati a ben selezionate popolazioni si possono avere dati affidabili e specifici più mirati. Se per esempio in un trial vengono reclutati solo pazienti diabetici con scompenso cardiaco e insufficienza renale cronica abbiamo a disposizione uno studio per questa particolare categoria di pazienti. Il problema è che per ogni patologia bisognerebbe avere molti RCT ognuno con una tipologia particolare di pazienti, e la cosa è, ovviamente, irrealizzabile.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 2 – Come analizzare gli studi

E' uno studio randomizzato?

Abbiamo visto precedentemente che quando si esamina uno studio la prima cosa da vedere è se esiste un gruppo di controllo. Il passo successivo è chiedersi se la suddivisione nei due gruppi è avvenuta in maniera casuale (in gergo ***randomizzazione***). La randomizzazione ha lo scopo di evitare che nel gruppo trattamento finiscano per esempio soggetti con prognosi basale migliore (il che farebbe credere che i risultati sono dovuti al farmaco e non al fatto che i soggetti erano già in partenza meno propensi ad ammalarsi); inoltre evita che nel gruppo controllo finiscano soggetti più ammalati e perciò più a rischio (anche in questo caso si sarebbe portati a ritenere che eventuali benefici nel gruppo trattamento siano dovuti al farmaco).

Mi spiego con un esempio. Poniamo che io volessi reclutare i partecipanti ad uno studio su un nuovo farmaco anti-infarto fra i medici presenti ad un congresso sulla cardiopatia ischemica. Per decidere chi sottoporre a trattamento attivo e chi a placebo divido i soggetti in due gruppi: nel primo gruppo entrano tutti quelli che non portano giacca e cravatta e a loro darò il farmaco, nell'altro gruppo entrano tutti quelli che portano giacca e cravatta e a loro darò il placebo. Alla fine dello studio trovo che il farmaco ha prodotto meno infarti del placebo. Sono assolutamente sicuro che questo risultato sia dovuto al farmaco? In realtà no, perché potrebbe essere che chi si veste in maniera più tradizionale sia anche più anziano di chi si veste in modo informale. Non è detto che sia effettivamente così, ma potrebbe. Se così fosse, nel gruppo che ha assunto placebo sono andati molti più anziani e nel gruppo che ha assunto farmaco molti più giovani. I risultati trovati non sono quindi dovuti al trattamento ma semplicemente al fatto che i due gruppi non erano simili, non erano confrontabili, e quello assegnato al placebo aveva già di base un rischio più elevato di infarto.

Proprio per evitare distorsioni del genere esistono delle metodiche validate per procedere ad una

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

randomizzazione ottimale. Queste metodiche sono complesse, prevedono che ci siano dei numeri generati in modo random dal computer, che tali numeri siano poi associati ai pazienti in modo casuale, ecc. Però in sostanza il metodo è assimilabile a questo che sto per descrivere (che non piacerà a chi disegna studi, ma fa lo stesso; per noi basta e avanza). Si mettono in un sacchetto tanti bigliettini con ognuno un numero diverso (ogni numero corrisponde ad un paziente). In un altro sacchetto si mettono altrettanti biglietti in cui sta scritto F (per farmaco) e P (per placebo). Poi, al classico bambino bendato si fa pescare un bigliettino dal primo sacchetto (quello dei pazienti) e un bigliettino dall'altro sacchetto (quello del trattamento) e li si abbina. In questo modo si è sicuri che la suddivisione dei pazienti è del tutto dovuta al caso (a meno che il bambino non abbia la benda forata in modo da poter sbirciare, ma qui saremmo nella truffa). Come facciamo a sapere se la suddivisione dei paziente è stata effettuata con tecnica randomizzata? Di solito questa informazione può essere reperita già esaminando l'abstract dello studio e non richiede particolari competenze. Sapere invece se la randomizzazione è stata effettuata con tutte le regole previste è cosa più complessa, e su questa dovrebbero indagare gli esperti che si occupano per mestiere di queste cose quando fanno il pelo e il contropelo alla qualità metodologica dello studio.

Fino a questo punto siamo riusciti quindi a stabilire che ci troviamo di fronte uno studio sperimentale, controllato e randomizzato. In gergo questi studi vengono identificati, come abbiamo già anticipato, con la ormai nota sigla RCT (Randomized Controlled Trial), sigla che per noi ora non ha più segreti.

Lo studio è in cieco?

Questo è un altro aspetto degli studi di intervento estremamente importante e che è diventato anche molto popolare tra i medici. Spesso gli Informatori del Farmaco, per vantare l'efficacia superiore del loro prodotto, citano qualche studio e non mancano di rimarcare il fatto che si tratta di

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

uno studio in doppio cieco. Ma cosa s'intende per ***cecità*** di uno studio? Con questo termine si vuol dire che il medico sperimentatore non sa che tipo di trattamento sta somministrando (cioè non sa se a quel particolare paziente sta dando farmaco o placebo); nello stesso tempo anche il paziente è cieco e non sa se sta assumendo farmaco o placebo. Ecco il perché si dice ***doppio cieco:*** cieco il medico e cieco il paziente.

Perché dicevo che questo è un aspetto molto importante da valutare in uno studio? L'importanza della cecità deriva dal fatto che se il medico sa che tipo di trattamento sta somministrando potrebbe essere influenzato nella raccolta dei dati. Si pensi per esempio se si devono raccogliere dati circa la situazione psicologica del paziente dopo aver somministrato un antidepressivo o un placebo: se il medico sa che quel determinato paziente ha assunto placebo può esserne influenzato (anche in modo del tutto inconscio) e concludere che il paziente non ha avuto miglioramenti dalla terapia, mentre potrebbe ritenere e registrare dei benefici maggiori per i pazienti che sa assumere il farmaco. Questo vale soprattutto se ciò che lo studio si è proposto di registrare sono degli end-point soggettivi (fra poco capiremo cosa sono gli end-point), conta un po' meno se l'end-point registrato è oggettivo (per esempio il numero di ictus o di decessi che si verificano, perché questi non sono influenzabili dal medico che sta raccogliendo i dati). Lo stesso discorso vale ovviamente per i pazienti. Anche in questo caso se l'end-point è il numero di decessi non è che questo sia influenzabile da quello che crede o pensa il paziente, ma lo diventa se si tratta di registrare dei sintomi soggettivi come il dolore, le vertigini, la qualità di vita, eccetera. In realtà la mancanza di cecità potrebbe acquistare importanza anche se gli end-point sono di tipo oggettivo: infatti chi sa che sta assumendo placebo potrebbe avere una compliance al trattamento inferiore di chi sa di assumere farmaco attivo.

Tuttavia la doppia cecità è importante anche per altri aspetti. Supponiamo che un soggetto arruolato in uno studio presenti un effetto collaterale: se il medico e il paziente sanno che sta

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

assumendo farmaco attribuiranno ad esso l'effetto, se sanno che sta assumendo placebo potrebbero dargli meno importanza e non riferirlo (paziente) o registrarlo (medico).

E' stato dimostrato che la mancanza di cecità può portare a sovrastimare l'efficacia di un trattamento anche del 10-15%.

Ma come si fa a sapere se lo studio è in cieco? Di solito anche questa è una informazione reperibile facilmente dall'abstract. In alcuni casi però la doppia cecità si ottiene con degli escamotage. Per esempio uno studio si proponeva di determinare se l'artroscopia e il lavaggio articolare sono efficaci nella gonartrosi. Si tratta di un RCT in cui i pazienti vennero randomizzati all'intervento chirurgico oppure ad un intervento simulato. In pratica i pazienti arruolati nel gruppo di controllo venivano portati in sala operatoria e si praticavano loro due piccole incisioni a livello del ginocchio in modo da simulare l'artroscopia, senza però eseguire nessun intervento. Evidentemente in questo caso i medici non potevano essere in cieco. Lo studio dimostrò che a distanza di 2 anni l'efficacia dell'intervento sul dolore e sulla limitazione funzionale era paragonabile al placebo chirurgico. La mancanza di cecità dei chirurghi è stata superata con un trucco, facendo raccogliere i dati ad altri medici, diversi da quelli che avevano effettuato l'intervento e che non conoscevano che tipo di trattamento era stato praticato. Questo studio spiega bene sia l'importanza della cecità che del gruppo di controllo: se non ci fosse stato un controllo i miglioramenti evidenziati dopo l'intervento sarebbero stati ascritti a quest'ultimo mentre sono dovuti probabilmente alla evoluzione spontanea della malattia o all'effetto placebo dell'operazione stessa, che non si sarebbe potuto escludere se non ci fosse stata la doppia cecità (del paziente e del medico che raccoglieva i dati finali).

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 2 – Come analizzare gli studi

Quali sono gli end-point?

Siamo arrivati finalmente a parlare degli end-point (detti talora anche esiti o outcomes). Sicuramente li avrete sentiti nominare molte volte e magari vi sarete anche chiesti che cosa diavolo sono. Ebbene gli end-point non sono altro che "**quello**" che alla fine dello studio i ricercatori si sono proposti di **misurare,** definizione rozza ma che ha il pregio di essere facilmente comprensibile. Con alcuni esempi mi spiego subito. Supponete di essere dei ricercatori che vogliono valutare l'efficacia di un farmaco antipertensivo. Allora misurerete la pressione ai vostri pazienti prima della somministrazione del farmaco e dopo. La riduzione media della pressione che otterrete potrebbe essere l'end-point dello studio. Facciamo un altro esempio. State provando un farmaco per l'osteoporosi, confrontandolo con il placebo. Dopo cinque anni andate a contare quante fratture ci sono state. Quello è l'end-point dello studio.

Come si può capire da questi due esempi banali, sostanzialmente esistono due tipi di end-point: quelli **hard** (per esempio numero di decessi, infarti, ictus, ricoveri per scompenso cardiaco, fratture di femore, ecc.) e quelli cosiddetti **surrogati** (per esempio la pressione arteriosa, il colesterolo, la massa ossea, ecc). E' intuitivo che gli end-point hard sono quelli maggiormente utili nel determinare il beneficio clinico di un trattamento. Ma allora, mi direte, perché in molti studi si trovano end-point surrogati? La ragione sta nel fatto che questi ultimi sono più facili da ottenere e soprattutto richiedono un follow-up più breve: se sto provando un farmaco ipocolesterolemizzante, un conto è andare a vedere dopo 3 mesi quale è stata la riduzione del colesterolo, un conto è aspettare 5 anni per vedere se sono diminuiti gli infarti. Gli end-point surrogati sono quindi molto comodi per i ricercatori, li si usa perché si ritiene che in qualche modo siano correlati agli outcomes (abituatevi anche a questo termine) clinici: è ragionevole infatti pensare che se riduco il colesterolo riduco anche il rischio di infarto. Però purtroppo la medicina è una scienza un po' particolare dove

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

logica e ragionevolezza non sempre la fanno da padroni. In altre parole non è detto che ad un miglioramento di un end-point surrogato corrisponda un beneficio clinico. Vale quindi la regola che per giudicare realmente la bontà di un farmaco gli end-point surrogati non possono sostituire quelli clinici.

Per capirci faremo alcuni esempi.

Il primo riguarda uno studio che ha valutato l'efficacia della terapia ormonale sostitutiva nel migliorare l'assetto lipidico (lo studio è noto con la sigla PEPI). In questo lavoro circa 900 donne furono suddivise in 4 gruppi, in tre gruppi si usavano varie combinazioni di ormoni e nel quarto gruppo il placebo. Al termine dello studio venne evidenziato che la terapia ormonale sostitutiva riduceva i valori di colesterolo LDL e aumentava quelli dell'HDL. Bene, direte voi, ecco qui un trattamento che oltre a migliorare i sintomi della menopausa riduce il colesterolo cattivo e fa aumentare quello buono, abbiamo quasi trovato la "pallottola magica" per prevenire la cardiopatia ischemica nella donna proprio in un' età in cui essa diventa più vulnerabile a questa patologia. Apparentemente sembra tutto logico e ragionevole, ma purtroppo non è così. Quando poi vennero effettuati studi sulla terapia ormonale sostitutiva con end-point clinici hard (studio WHI, studio HERS, studio ESPRIT), sia in donne sane che cardiopatiche, si è visto che la terapia ormonale proposta per la menopausa non solo non protegge dalla cardiopatia ischemica, non solo non ha un ruolo protettivo sugli eventi cardiaci, ma addirittura ne aumenta il rischio.

Il secondo esempio è forse ancora più eclatante e suggestivo e viene citato in tutti i testi di metodologica degli studi clinici. L'esempio riguarda lo studio CAST in cui venne sperimentata la flecainide (un antiaritmico) nei soggetti infartuati con aritmie ventricolari minacciose. E' noto che nei pazienti post-infartuati vi è un aumento del rischio di morte improvvisa, specialmente nei primi giorni e nelle prime settimane dopo l'evento acuto. I soggetti più a rischio sono quelli che mostrano aritmie ventricolari frequenti e minacciose (fenomeno R/T, run di tachicardia ventricolare, BEV

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

polimorfi, ecc.). E' logico e ragionevole pensare che se riuscissi, con una terapia farmacologica, a sopprimere o ridurre di intensità queste aritmie ne avrei un vantaggio anche clinico, nel senso che avrei meno decessi per morte aritmica improvvisa. Lo studio CAST dimostrò infatti che il farmaco antiaritmico agiva positivamente sulle aritmie provocandone una netta riduzione. Tuttavia il trial venne sospeso anticipatamente. Come mai? L'analisi ad interim dei dati mostrava infatti un eccesso di morti nel gruppo trattato attivamente rispetto al gruppo di controllo che assumeva placebo. Questo studio è veramente paradigmatco in quanto dimostra elegantemente che un farmaco, che ha un effetto positivo su un end-point surrogato (aritmie), non necessariamente produce un beneficio clinico. Qualcuno parlò allora di “effetto cosmetico” del farmaco sull’elettrocardiogramma.

Un altro esempio ci viene dallo studio ILLUMINATE in cui venne somministrato torcetrapib oppure placebo a pazienti a rischio cardiovascolare. Il torcetrapib, inibendo una particolare proteina denominata CEPT, aumenta il colesterolo HDL: lo studio però venne interrotto anticipatamente perché, nonostante il farmaco producesse un aumento del 72% del colesterolo HDL ed una riduzione del 24,9% della frazione LDL, era gravato da un aumento degli eventi cardiovascolari e della mortalità.

Ho appena accennato alla cosiddetta **analisi ad interim**. Di cosa si tratta? E' una procedura di garanzia messa in atto per evitare brutte sorprese. Mentre il trial è in corso i dati preliminari vengono costantemente monitorati in modo da interrompere lo studio prima del termine previsto se il numero di eventi predefiniti (infarti, ictus, ricoveri per scompenso, fratture femorali, ecc.) dovesse superare una certa soglia rispetto all'altro braccio. Lo studio può essere sospeso anticipatamente sia perché il farmaco si dimostra "troppo" efficace rispetto al controllo sia nel caso opposto, quando il farmaco si dimostra meno efficace del controllo. Nel primo caso non sarebbe etico continuare lo studio privando i malati (sia quelli arruolati nel braccio controllo sia tutti i malati in

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

genere) di un trattamento che si è dimostrato chiaramente utile anche prima che lo studio finisca, nel secondo caso non è morale continuare a trattare soggetti con un farmaco che fa peggio del controllo. Tuttavia l'interruzione precoce di una studio comporta anche un rovescio della medaglia. Supponiamo per esempio che si sia reso disponibile un nuovo trattamento per i pazienti sieropositivi per HIV. Viene effettuato uno studio che paragona la nuova terapia con quella già disponibile per determinare se si riesce a ritardare la comparsa di AIDS conclamato. Lo studio dovrebbe durare 6 anni, ma dopo 3 viene interrotto perché l'analisi ad interim mostra un numero molto minore di insorgenza di AIDS con il farmaco nuovo. Tuttavia ci si accorge che questo farmaco provoca anche un aumento dell'ictus e dell'infarto fatali, ma la differenza non è significativa rispetto al farmaco di confronto. Non si può escludere però che tale differenza avrebbe potuto diventare rilevante se lo studio fosse durato i sei anni pianificati. In questo caso l'interruzione anticipata da un lato permette di estendere i benefici del nuovo farmaco nel ritardare la comparsa di AIDS a una vasta schiera di malati, dall'altro impedisce di valutare compiutamente il suo profilo di rischio.

Un ulteriore aspetto da considerare a proposito degli end-point è quello dei criteri diagnostici. Che cosa voglio dire con questo? Mi spiego con un esempio.

Se lo studio si propone di valutare di quanto migliorerà l'artrite reumatoide con un farmaco biologico, si dovranno stabilire dei criteri di attività della malattia (per esempio il numero di erosioni ossee oppure la valutazione del dolore tramite sistemi a punteggio) sia al baseline che al termine dello studio.

E' evidente che per alcuni end-point non è necessario definire nulla. Per esempio se si valutano i decessi, questi sono decessi punto e basta, non si corre certo il rischio che un medico usi dei criteri diversi di diagnosi. Un caso particolare però è rappresentato dall'end-point che valuta non i decessi in sé ma i decessi specifici (per esempio decessi per scompenso, per cancro mammario,

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

ecc.) perché in questa evenienza si potrebbero avere diversità interpretative. Per esempio in uno studio che volesse determinare la mortalità da cancro prostatico dopo intervento chirurgico o vigile attesa si potrebbe verificare quanto segue: la mortalità specifica risulta ridotta nel "gruppo intervento" rispetto al gruppo "vigile attesa", però la mortalità totale non è diversa tra i due bracci. Come mai? Una spiegazione può essere che lo studio non ha una potenza statistica tale da mettere in evidenza differenze sulla mortalità totale; un'altra però potrebbe essere che i decessi che si sono verificati nel gruppo chirurgico in seguito a complicanze post-operatorie (per esempio embolie polmonari o infezioni) sono stati classificati come decessi "non dovuti al cancro prostatico", il che porta all'apparente paradosso di una riduzione della mortalità specifica ma non di quella totale. Questo problema è stato sottolineato negli studi di screening oncologici, in cui molti esperti sostengono che l'unico end-point corretto da valutare è la mortalità totale e non quella specifica da cancro. Vediamo questo esempio, volutamente paradossale, ma che serve a spiegare meglio questo punto. In uno studio su uno screening oncologico succede che nel gruppo randomizzato allo screening la neoplasia viene scoperta molto più precocemente rispetto al gruppo non screenato, il tumore è in uno stadio operabile quindi i pazienti sono avviati all'intervento chirurgico; al contrario nell'altro braccio il tumore viene scoperto tardivamente, tanto da essere inoperabile, quindi quasi nessuno viene sottoposto all'intervento. Al termine dello studio si trova che i decessi "attribuiti" al tumore sono stati decisamente inferiori nel gruppo screenato, ma i decessi totali sono stati decisamente superiori per una elevata percentuale di complicanze post-operatorie. Se lo studio valutasse solo l'end-point "decessi da cancro" non darebbe una informazione corretta perché i pericoli dello screening sarebbe maggiori dei benefici, portando ad un aumento della mortalità totale.

Veniamo adesso ad un aspetto particolare, quello degli **end-point secondari**, croce e delizia degli esperti di "critical appraisal", cioè di quei tizi un po' rompiscatole e bastian contrari che si divertono

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

a fare le pulci ai trials. Quando si studia il disegno di un trial si definiscono uno o più **end-point primari** che si andrà poi a misurare. Conoscere qual è l'end-point primario di un trial è di capitale importanza perché è quello sul quale viene tarata la potenza statistica. In altri termini è l'end-point primario che definisce lo scopo dello studio stesso. E' diventata prassi comune definire anche uno o più end-point secondari. Non è detto che, dal punto di vista clinico, l'end-point primario sia più importante di quello secondario, ma è quello sul quale si dovrebbe interpretare statisticamente il trial. Infatti è su quest'ultimo che si sono basati tutti i calcoli statistici preliminari. Gli end-point secondari sono utili se vanno nella stessa direzione di quelli primari, ma se non è così può essere fuorviante trarre delle conclusioni affidabili basandosi solo su di loro. Ma gli autori, nelle loro conclusioni, tengono sempre conto di questo "caveat"? Purtroppo non è così. Per esempio in uno studio vengono paragonati due farmaci antipertensivi e l'end-point primario sia costituito dall'infarto non fatale. Alla fine dello studio questo end-point non differisce statisticamente tra i due gruppi, ma si registra una diminuzione di end-point secondari (stroke, eventi cardiovascolari totali, interventi di rivascolarizzazione coronarica) in uno dei due gruppi. Gli autori concludono che un regime antipertensivo è più efficace di quello di paragone, ma non danno importanza al fatto che i due farmaci sono equivalenti per l'end-point primario, mentre la riduzione di alcuni end-point secondari richiederebbe conferme da ulteriori studi. Non si vuol qui sostenere che gli end-point secondari non siano importanti, possono di per sé essere pienamente validi, ma andrebbero interpretati con più cautela perché una valutazione rigorosa dello studio, dal punto di vista statistico, dovrebbe fare sempre riferimento all'end-point primario. Una riduzione "statisticamente significativa" di un outcome secondario potrebbe esserlo solo in via nominale, cioè solo apparente. Infatti il potere "matematico" dello studio si riflette unicamente sull'outcome primario, ed è su questo, come dicono gli esperti, che è stato "speso" tutto il suo potere statistico.

Accettare per certa una significatività statistica di un end-point secondario significa accettare un

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

margine di errore che potrebbe essere troppo elevato. I risultati derivanti da end-point secondari possono fornire informazioni supplementari rispetto a quelle trovate con l'end-point primario se sono concordanti. In caso contrario dovrebbero essere considerati soprattutto delle ipotesi da valutare in uno studio successivo. Insomma un'interpretazione dello studio basata solo su end-point secondari andrebbe sempre guardata con prudenza, per quanto prestigiosa sia la rivista che pubblica il lavoro.

In **conclusione**, un consiglio: quando qualcuno vi presenta uno studio magnificando le virtù del tal farmaco chiedete sempre quali erano gli end-point considerati dai ricercatori e abituatevi a dubitare se si tratta di end-point surrogati. Dubitare però non vuol dire non "efficace" perché nulla vieta che un farmaco che ha a sua dimostrazione solo studi su end-point surrogati non possa, in futuro, disporre di studi che ne dimostrino l'utilità anche su outcomes clinici importanti. In genere il problema riguarda farmaci immessi in commercio recentemente, essi sono ancora troppo nuovi per avere già a loro merito studi con esiti clinici che richiedono molti pazienti arruolati e vari anni d'uso. In questi casi è utile **sospendere il giudizio**, come facevano certi filosofi del buon tempo antico, e aspettare. Nel frattempo conviene usare farmaci alternativi più vecchi (di solito il mercato è sovrabbondante) e sperimentati e di cui si conosce meglio, proprio perché da più tempo in uso, il profilo di sicurezza a lungo termine. Questa strategia permette di evitare, con una certa ragionevolezza, di incorrere in effetti collaterali non noti, che sono più spesso prerogativa dei farmaci più recenti e usati da minor tempo. Un altro consiglio è quello di valutare sempre con occhio critico la superiorità di un trattamento rispetto ad un altro se questo giudizio si basa solo su end-point secondari.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Quanti sono i pazienti arruolati?

Il numero dei pazienti arruolati nello studio è un altro parametro importante da valutare e facile da reperire già dall'abstract. E' intuitivo che tanto più numeroso è il campione arruolato e tanto più lungo il follow-up tanto più i risultati dovrebbero essere validi e affidabili. La numerosità del campione è importante perché solo con certi numeri si può avere la potenza statistica per rilevare determinati eventi. E' evidente che se un trattamento ha lo scopo di ridurre un evento che già di per sé non è frequente bisogna arruolare molti pazienti (migliaia o decine di migliaia) per poterlo rilevare. La numerosità del campione viene quindi ritenuta, di solito, sinonimo di studio clinico importante. Questo però può portare anche a delle incongruenze. Supponiamo per esempio che un farmaco riduca l'ictus rispetto al farmaco concorrente e che però la differenza possa diventare statisticamente significativa (vedremo in seguito cosa significa questa espressione) solo se si studia un numero molto elevato di pazienti (per esempio 30-40.000). Succede allora che differenze marginali vengono amplificate perché si sono reclutati moltissimi soggetti: si ottiene una significatività statistica ma l'utilità clinica di questa informazione è discutibile. In effetti si assiste sempre più spesso alla organizzazione di mega-trial che hanno lo scopo di mettere in evidenza differenze di efficacia tra due farmaci molto piccole, che non risulterebbero se la casistica fosse meno numerosa. Aumentare la casistica può quindi essere un escamotage per evidenziare benefici piccoli. Mi direte: ma scusa se un farmaco è efficace lo è tanto sui piccoli quanto sui grandi numeri. La faccenda purtroppo non funziona così. Per farmi capire farò un esempio perché penso che gli esempi servano molto di più di tante parole. Dovrò purtroppo usare dei numeri (anche se cercherò di usarli molto semplici) e alcuni concetti che per ora a molti possono non essere chiari. Lo diventeranno in seguito, per il momento l'importante è seguire il filo del ragionamento. Supponiamo di avere un nuovo farmaco che riduce il rischio di infarto e di volerlo confrontare con un farmaco già ampiamente usato. Arruolo quindi 5.000 soggetti e li divido in

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

maniera randomizzata in due gruppi di 2.500 ciascuno. Al primo gruppo somministro il nuovo farmaco, al secondo gruppo quello più vecchio. Dopo cinque anni vado a contare quanti infarti ci sono stati nei due gruppi.

Ecco i risultati:

- Gruppo farmaco nuovo: 2.500 trattati e 140 infarti
- Gruppo farmaco vecchio: 2.500 trattati e 160 infarti

Sembra quindi che il nuovo farmaco sia più efficace del vecchio, ma la statistica è una cosa strana, non basta dimostrare che si hanno meno infarti, bisogna anche dimostrare che tale riduzione non è dovuta semplicemente al caso ma, come si dice, è **significativa dal punto di vista statistico**. Per fare questo si calcola la riduzione relativa del rischio di infarto avuta con il nuovo farmaco e si trova che questa è del 12,5% (in effetti se si fa 160 meno 12,5% si avrà 140). Più avanti impareremo anche a calcolare il rischio relativo (non è così difficile come potrebbe sembrare) e anche che esso si esprime in questo modo: RR = 0,875. Per il momento non è importante che capiate questo strambo modo di esprimersi, ma intanto cominciate con il prenderne confidenza.

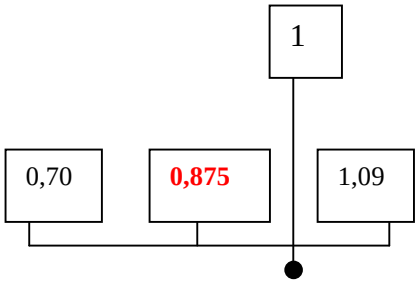
Però il calcolo del rischio non basta ancora, bisogna anche trovare il cosiddetto intervallo di confidenza al 95% (che si scrive così: IC95%). Un risultato per essere significativo dal punto di vista statistico deve avere un IC95% che NON comprende l'unità (cioè il numero 1). Se per sfortuna lo comprende vuol dire che la differenza trovata conta poco o nulla perché **non è significativa statisticamente**. Nel caso dell'esempio in esame l'IC95% va da 0,7 a 1,09 e comprende perciò il numero 1. Questo vuol dire che il 12,5% in meno di infarti trovati con il nuovo farmaco non conta quasi nulla e perciò che i due trattamenti devono essere considerati di efficacia paragonabile. Non state a chiedervi per ora come mai l'IC95% non deve comprendere il numero 1 e credetemi sulla parola. Quando spiegherò queste cose potete tornare a rileggervi questo punto e tutto vi sembrerà liscio come l'olio. Anzi per complicare un po' le cose riporto una tabella che

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

probabilmente per molti sarà poco comprensibile ma che mostra come in generale vengono riportati i risultati di uno studio (in modo semplificato, ovviamente). Intanto cominciate ad abituarvi a questo nuovo linguaggio.

	Farmaco nuovo	Farmaco vecchio
Trattati	2500	2500
Infarti	140	160

RR 0,875
IC95% 0,70 – 1,09



Ovviamente un risultato del genere può non far comodo a chi produce il nuovo farmaco, allora che fa? Arruola ben 30.000 pazienti, dividendoli poi in due bracci di 15.000 ciascuno. Somministra come di prassi nuovo e vecchio farmaco e dopo 5 anni tira le somme.

Ecco i risultati:

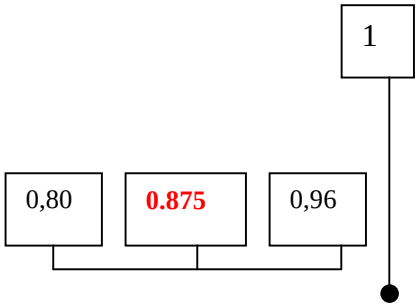
- Gruppo nuovo farmaco: 15.000 trattati e 840 infarti
- Gruppo vecchio farmaco: 15.000 trattati e 960 infarti

Se calcoliamo di quanto è stata la riduzione del rischio di infarto con il nuovo farmaco troveremo che essa è sempre del 12,5% (infatti 960 meno 12,5% fa appunto 840). L'RR sarà sempre 0,875, ma l'intervallo di confidenza al 95% va da 0,8 a 0,96 e non interseca il numero 1. Pertanto la differenza trovata è diventata miracolosamente significativa. Anche qui riporto la tabella così vi mantenete in esercizio.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

	Farmaco nuovo	Farmaco vecchio
Trattati	15.000	15.000
Infarti	840	960

RR 0,875
IC95% 0,80 – 0,96



Un consiglio, ritornate su queste pagine quando in seguito il modo di rappresentare i dati della tabella vi sarà più familiare.

In conclusione si vede come una differenza di efficacia tra due farmaci piccola diventa rilevabile dalla statistica se si aumenta il numero dei soggetti studiati. Tutto questo ha degli evidenti vantaggi nel senso che permette di scoprire differenze che altrimenti, con meno casi arruolati, non si potrebbero vedere, ma porta d'altra parte ad enfatizzare benefici che clinicamente potrebbero essere scarsi. L'organizzazione di mega-trial con decine di migliaia di pazienti arruolati ha lo scopo di evidenziare differenze nei trattamenti che non risulterebbero con meno soggetti. Queste differenze acquistano importanza in termini statistici di popolazione quando vengono trattati milioni di persone, ma possono essere minime per il singolo paziente. Nel caso esemplificato, se volessimo tradurre il tutto in termini facilmente comprensibili, potremmo anche dire che bisogna trattare 125 soggetti per 5 anni con il nuovo farmaco per avere un infarto in meno. Ciò vuol dire anche che per 5 anni ne tratto inutilmente 124, i quali assumeranno il farmaco per non avere nessun beneficio in più e correranno il rischio di eventuali effetti collaterali. Questo modo di vedere le cose fa riferimento al cosiddetto **NNT** (numero di soggetti che è necessario trattare per evitare un evento in un determinato lasso di tempo). Impareremo in seguito anche a calcolarlo, questo famoso NNT, per ora mi serve solo per dire che il nuovo farmaco potrebbe essere utile solo se non

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

porta ad un maggior numero di effetti collaterali gravi in più rispetto al vecchio farmaco. Se per ipotesi avessi ogni 125 pazienti trattati per 5 anni un infarto in meno rispetto al vecchio trattamento ma 3 uremie terminali in più, il beneficio andrebbe a farsi benedire. In altre parole il rapporto rischi/benefici non sarebbe favorevole. Questo è un punto molto importante: se il disegno del trial non prevedesse di registrare anche le uremie terminali potrei non aver ben chiaro il reale profilo di sicurezza del farmaco e ritenere che sia efficace perché riduce gli infarti, ma non pericoloso per il rene. Purtroppo non sempre i trials riportano tutti gli eventi avversi gravi (SAEs = Serious Adverse Events) che si verificano, anche quelli che apparentemente non sembrano legati al trattamento in esame, e questo porta a non avere un quadro chiaro del reale impatto sulla salute del farmaco in prova. Un altro aspetto da considerare poi è il costo della nuova terapia: se questo fosse molto elevato vale la pena investire molti soldi per avere un infarto in meno ogni 125 trattati per 5 anni o è preferibile investire i fondi in altri progetti sanitari con un costo ed un'efficacia più favorevoli? Come si può intuire le risposte non sono affatto semplici ma quando si giudica dell'utilità di un farmaco vi sono molti aspetti da considerare, non ultimi quelli economici, soprattutto in tempi di risorse sanitarie limitate.

Qual è la durata dello studio?

La durata dello studio (detta anche follow-up) dipende naturalmente dall'end-point che si vuol misurare. Per esempio se si vuol vedere se un farmaco antipertensivo è in grado di ridurre la pressione più del placebo può bastare un follow-up di qualche mese. Se al contrario si vuol valutare se lo stesso farmaco è in grado di ridurre le complicanze della malattia ipertensiva (come per esempio l'ictus o l'infarto o lo scompenso cardiaco) è necessario disegnare uno studio con un follow-up adeguato della durata di almeno qualche anno. Quando si valuta uno studio bisogna

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

quindi sempre chiedersi se la durata dello studio è adatta a valutare gli outcomes previsti. Per esempio in uno studio si vuol valutare l'efficacia di un nuovo farmaco proposto per il morbo di Alzheimer. Come end-point si sceglie di misurare lo stato funzionale e psichico tramite un questionario somministrato prima dell'inizio dello studio e dopo sei mesi. E' evidente che per una malattia a decorso cronico e progressivo come l'Alzheimer la valutazione dell'efficacia della terapia a sei mesi è probabilmente insufficiente a determinare se il farmaco è utile o meno a ridurre la disabilità e la progressione del morbo.

Esempio conclusivo

Come esempio finale di alcuni dei punti che abbiamo considerato finora porterò uno studio che aveva randomizzato meno di 200 donne a terapia ormonale sostitutiva o placebo. Dopo 12 mesi nel gruppo in terapia attiva si osservò una riduzione della proteina C reattiva (PCR) rispetto ai valori basali e al gruppo di controllo. La PCR elevata è un noto fattore associato al rischio cardiovascolare. A questo punto allora una domanda facile per i lettori: secondo voi è corretto, sulla base di questo studio, concludere che la terapia ormonale sostitutiva (TOS) protegge il cuore? Credo che chiunque mi abbia seguito, anche distrattamente, fino a questo punto non troverà difficoltà a rispondere che non è assennato trarre conclusioni di questo tipo. In effetti il supposto ruolo protettivo della TOS sulle malattie cardiovascolari è stato smentito clamorosamente dallo studio WHI che aveva arruolato più di 16.000 donne con un follow-up di circa 5 anni . Il confronto tra lo studio precedente e il WHI non è neppure pensabile, non solo perché il primo ha valutato un endpoint surrogato (PCR) e il secondo degli endpoint clinici ben più importanti (mortalità, infarto, ictus, tromboembolismo venoso, ecc.), ma anche per la numerosità del campione e la diversa durata.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 3
Analisi per sottogruppi

Quando si disegna uno studio di solito si predefiniscono degli end-point che poi si andranno a misurare: quanti infarti fatali e non fatali si avranno alla fine dello studio, quanti ictus, eccetera. Abbiamo visto nel capitolo sugli end-point che la potenza statistica dello studio viene tarata sull'end-point primario, ma possono essere previsti anche degli end-point secondari. Spesso i ricercatori però non si limitano ad analizzare i dati sull'intero campione arruolato ma scompongono quest'ultimo in vari sottogruppi. Per esempio si decide di andare a vedere se tra tutti i pazienti arruolati il trattamento si è dimostrato più utile in particolari tipi di pazienti (nelle donne piuttosto che negli uomini, nei diabetici piuttosto che negli obesi, ecc.) oppure se il trattamento porta a risultati diversi nei soggetti che fumano rispetto a chi non fuma o ancora in coloro che assumono regolarmente il farmaco rispetto a chi ha una cattiva compliance farmacologica e così via.

Questo modo di procedere è utile perché permette di ricavare molte informazioni, ma esse vanno sempre prese con cautela in quanto i risultati potrebbero essere dovuti al caso. E' buona norma quindi considerare i risultati derivanti da una ***analisi per sottogruppi*** come un'ipotesi che dovrebbe essere convalidata da studi successivi. Comunque le analisi per sottogruppi vengono considerate più affidabili se erano originariamente previste nel protocollo del trial, lo sono meno se vengono effettuate a posteriori e senza essere state originariamente predefinite. L'ideale sarebbe non solo che l'analisi fosse già stata definita prima, ma che venisse usata la cosiddetta "randomizzazione stratificata" che permette una eguale distribuzione nei due bracci dei vari sottogruppi. Nello studio Val-heFT pazienti con scompenso cardiaco vennero trattati con valsartan o placebo. Siccome è noto che i betabloccanti incidono sulla prognosi dello scompenso si decise addirittura di randomizzare i pazienti che assumevano e non assumevano betabloccanti ai due gruppi così che la loro presenza nel gruppo trattamento e nel gruppo controllo fosse omogenea. Ma perché diciamo che le analisi per sottogruppi possono dare risultati inaffidabili?

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 3 – Analisi per sottogruppi

I moderni metodi di analisi statistica effettuati con potenti elaboratori elettronici permettono di valutare decine e decine di dati e qualche volta salta fuori qualche risultato che apparentemente ha una significatività statistica, ma in realtà è dovuto al semplice gioco del caso. Per esempio se viene fatto uno studio su 20.000 pazienti ipertesi e si decide di fare un'analisi per sottogruppi dividendo i pazienti per età (maggiori o minori di 65 anni), per la presenza o meno di diabete, per lo stato di fumatore (fumo si/no), per il sesso (maschi/femmine), per l'uso o meno di aspirina (asa si/no), per la presenza o meno di colesterolo alto (colesterolo > 200 mm/ml si/no) eccetera, si ottiene un numero incredibile di combinazioni. L'analisi computerizzata potrebbe allora mostrare che il farmaco, che nello studio nel suo insieme non ha evidenziato risultati favorevoli, sia invece efficace in una certa sottopopolazione (per esempio nelle donne con più di 65 anni, con colesterolo > 200 ma senza diabete e che non fumano). E' probabile che questo risultato sia semplicemente dovuto ad una combinazione casuale. E' stato dimostrato che tante più sono le analisi per sottogruppi effettuate in un determinato campione e quindi tante più sono le combinazioni, tanto più aumenta la probabilità di trovare un risultato apparentemente significativo dal punto di vista statistico ma in realtà dovuto solamente al gioco del caso. Chi si occupa di valutare criticamente gli studi clinici considera una serie di aspetti: come si è già accennato, per esempio si controlla se l'analisi per sottogruppi era stata pianificata o meno nel protocollo dello studio, oppure viene valutata l'entità dell'effetto trovato nei vari sottogruppi rispetto al campione totale, o ancora la plausibilità o meno dei risultati, ecc. Si tratta di aspetti notevolmente complessi, per quanto ci riguarda basti ricordare il seguente principio: considerare in genere preliminari e quindi meritevoli di ulteriori studi i risultati di una analisi per sottogruppi, ancorché ben progettata e condotta.

Restano da esaminare le **analisi a posteriori (o post-hoc analysis**). Di che cosa si tratta? Con questo termine si intendono quelle analisi non contemplate nel protocollo di ricerca che vengono

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

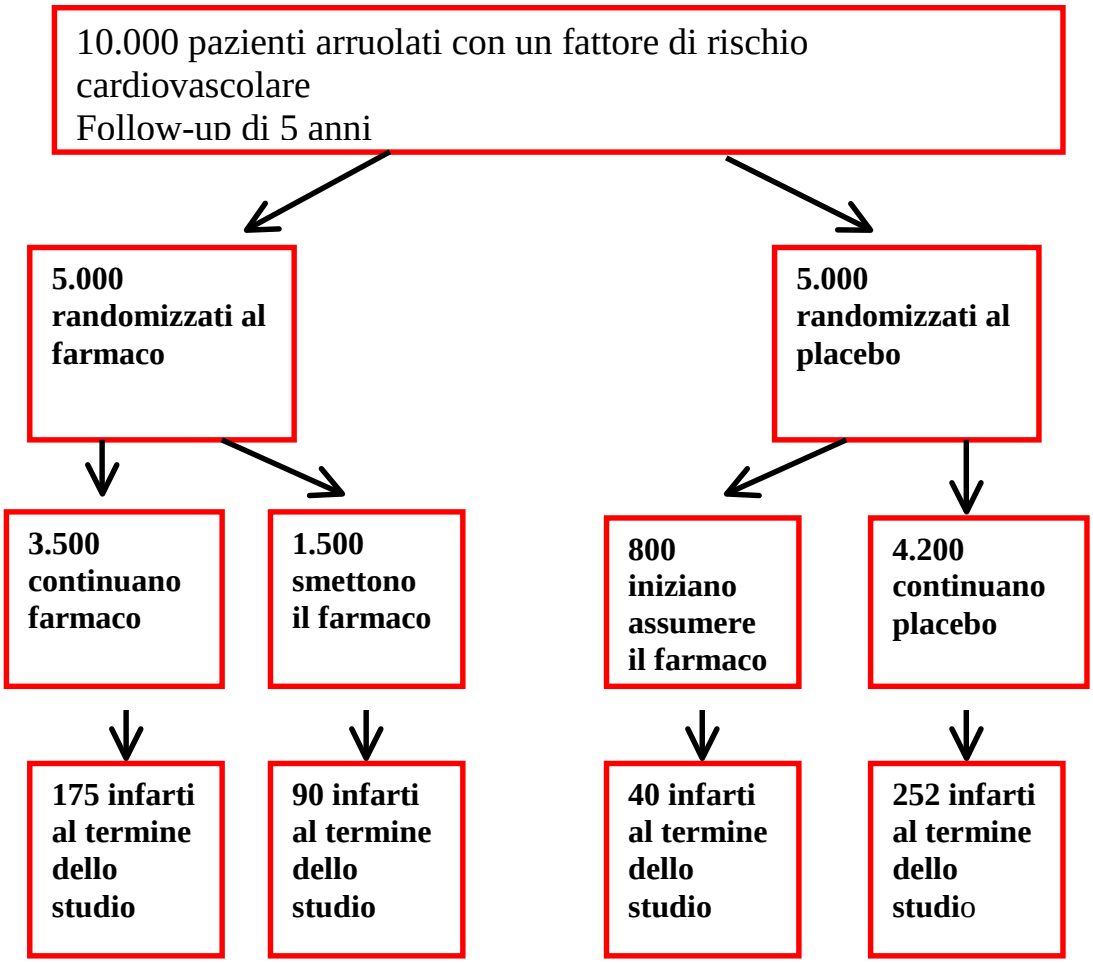
effettuate dopo la conclusione della raccolta dati. E' un fenomeno molto frequente in letteratura: per esempio in uno studio è stato calcolato che circa metà dei lavori cardiologici considerati erano analisi a posteriori effettuate su vari sottogruppi. Durante uno studio, infatti, vengono raccolti molti dati sui pazienti (per esempio il tipo di dieta, la loro attività fisica, il tipo di farmaci assunti, le loro abitudini sessuali, voluttuarie, lo stato socio-economico, i loro viaggi, i vari passatempi, ecc). Succede così che dopo la fine dello studio venga la curiosità di andare a vedere per esempio quello che succede in certi tipi di pazienti rispetto ad altri. Alcuni ricercatori, per dimostrare quanto possa essere ingannevole questo tipo di analisi, ne fece una sui dati di uno studio che aveva valutato i beta-bloccanti nel post-infarto: risultò che il farmaco era molto **più efficace nei pazienti nati sotto un determinato segno zodiacale** rispetto ad altri! Ora, per quanta fede si possa avere nei segni zodiacali e nell'oroscopo, è un po' difficile ammettere che questo risultato abbia validità scientifica. Vale quindi l'avvertenza di sempre: le analisi a posteriori permettono di esplorare molteplici aspetti ma è opportuno prendere con le molle questi dati e considerarli più che altro alla stregua di ipotesi. Qualcuno è arrivato a definire le analisi a posteriori una specie di "tortura dei dati".

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 4.
L'intenzione a trattare

Dico subito che qui stiamo affrontando un aspetto che è abbastanza complicato da spiegare e da capire. Con il termine "***intenzione a trattare o intention to treat "*** si intende che l'analisi dei risultati viene effettuata considerando i pazienti secondo il gruppo cui erano stati originariamente assegnati e non se hanno assunto o meno il farmaco. Cercherò di spiegarmi meglio con un esempio, altrimenti detta così la cosa può apparire anche più astrusa di quanto sia in realtà. Dei ricercatori si propongono di valutare un farmaco che, da studi precedenti, si sa ridurre l'infarto nei pazienti con tre o più fattori di rischio. Il loro scopo è vedere se funziona anche nei pazienti con un solo fattore di rischio. Arruolano così 10.000 pazienti con un fattore di rischio per infarto (diabete, ipertensione, colesterolo > 250 mg/dL) e li randomizzano in due gruppi di 5.000 ciascuno; a un gruppo somministrano il farmaco in esame e all'altro somministrano un placebo. Si valuterà il numero di infarti dopo 5 anni. Tuttavia succede, per vari motivi (scarsa compliance, effetti collaterali pesanti, ecc.) che 1.500 pazienti del gruppo trattamento smettono il farmaco precocemente. Al contrario a 800 pazienti del gruppo controllo viene somministrato ad un certo punto del follow-up il farmaco perché diventano ad alto rischio cardiovascolare per la comparsa di altri fattori di rischio e non sarebbe etico lasciarli senza trattamento. Nello schema sottostante viene rappresentato graficamente quanto si verifica durante lo studio e al termine del follow-up. In definitiva si verifica che hanno assunto il farmaco 4.300 pazienti (3.500 del gruppo trattamento e 800 del gruppo controllo) mentre non lo hanno assunto 5.700 pazienti (4.200 del gruppo controllo e 1.500 del gruppo trattamento). Come si vede ci sono stati 90 infarti nel gruppo randomizzato a farmaco che ha smesso il trattamento e 175 nel gruppo che ha continuato; invece nel gruppo controllo si verificano 40 infarti nel gruppo che ad un certo punto ha iniziato il farmaco e 252 nel gruppo che ha continuato il placebo.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati



Sommando gli infarti non in base al gruppo cui erano originariamente allocati i pazienti ma in base all'assunzione o meno del farmaco (modo di analizzare il trial che viene detto "**as treated**") si ottiene :

- 215 infarti nel gruppo che ha assunto il farmaco (175 + 40)
- 342 infarti nel gruppo che non ha assunto il farmaco (252 + 90)

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Con questo metodo il trattamento ottiene una riduzione significativa del rischio di infarto del 18%. Questo si esprime in gergo tecnico con la seguente formula: RR 0,82 ; IC95% 0,69-0,98.Tuttavia se i dati vengono analizzati secondo **l'intenzione a trattare** (cioè valutando quanti infarti ci sono stati nei 5.000 pazienti del gruppo trattamento e nei 5.000 del gruppo controllo indipendentemente dall'aver o meno assunto il farmaco) si ottiene:

- infarti nel gruppo trattamento 265 (175 + 90)
- infarti nel gruppo controllo 292 (252 + 40)

La riduzione del rischio ottenuta con il trattamento è del 10% . Questo si esprime in gergo statistico con la formula: RR 0,90 , IC95% 0,76-1,07. Come si capirà meglio in seguito questo risultato ***non è significativo.*** Ebbene lo ammetto, i numeri di questo esempio sono stati appositamente elaborati per mostrare che se non si fa un'analisi intention to treat ma "as treated" si possono ottenere misure di efficacia differenti. Un altro modo di analisi è noto come "**per protocol**" in cui si considerano solo i pazienti che hanno rispettato il protocollo del gruppo a cui erano assegnati. Anche questa modalità può portare a distorsioni. Non è detto che sia sempre così ma è una eventualità possibile. Ma perché, direte voi, non è giusto fare un'analisi as treated o per protocol? Un motivo, per esempio, potrebbe essere che i 1.500 pazienti che hanno sospeso il farmaco siano quelli che sono più a rischio, oppure più fragili dal punto di vista clinico, o coloro che non tollerano il trattamento, ecc. Contarli come appartenessero al gruppo placebo o comunque sottrarli al gruppo trattamento va a rompere la randomizzazione, cioè quel particolare artificio tecnico che gli studiosi mettono in atto per far sì che i due gruppi (trattamento e controllo) siano sovrapponibili, con fattori di rischio noti e non noti equamente distribuiti. Se fosse vero che i 1.500 che interrompono il trattamento sono i soggetti più anziani o con comorbilità e quindi di per sé già a rischio maggiore di infarto, incorporarli nel gruppo placebo o sottrarli al gruppo trattamento diventa, se vogliamo usare un termine sportivo, sleale.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 4 – L’intenzione a trattare

Un’analisi non intention to treat può portare a sovrastimare l’efficacia di un trattamento, anche se non sempre è così. In ogni caso, come ho appena detto, ignorare "l'intention to treat" significa rompere quel delicato meccanismo della randomizzazione messo in atto appositamente perché i due gruppi confrontati siano paragonabili e quindi ci scompiglia le carte in tavola. Siccome risulta difficile per un lettore medio capire con quale modalità i ricercatori hanno effettuato l’analisi, è opportuno focalizzare sempre l’attenzione su questo punto: se i ricercatori non dichiarano esplicitamente che l’analisi e stata eseguita secondo tale modalità, può essere che qualcosa non quadri.

Dato che si tratta di un aspetto ostico, per illustrare l'intention to treat farò due esempi tratti da due studi reali che dimostrano come un'analisi non intention to treat possa portare a interpretazioni fuorvianti.

Il primo è lo studio HOT (Hypertension Optimal Treatment) in cui erano stati arruolati quasi 19.000 pazienti ipertesi con pressione arteriosa diastolica (PAD) compresa tra 100 e 115 mmHg. I pazienti vennero randomizzati in tre gruppi: un gruppo doveva raggiungere una PAD inferiore a 90 mmHg (gruppo A), un gruppo aveva l’obiettivo di arrivare a valori inferiori a 85 mmHg (gruppo B), in un gruppo la PAD doveva scendere sotto gli 80 mmHg (gruppo C). Per arrivare a questi obiettivi inizialmente veniva somministrato un calcio-antagonista e se non si raggiungeva il valore prefissato di PAD si potevano aggiungere altri ipotensivi. Lo scopo dello studio era di dimostrare che a più bassi valori di PAD raggiunti con la terapia corrispondeva una riduzione degli eventi cardiovascolari. Al termine dello studio il numero di eventi era simile nei tre gruppi.

Lo studio quindi risultò negativo e non riuscì a dimostrare (esclusa la sottopopolazione di pazienti diabetici) che raggiungere una PAD inferiore a 80 mmHg è meglio che arrivare a una PAD inferiore a 90 mmHg. Tuttavia nell'abstract dello studio si legge che la più bassa incidenza di eventi

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

cardiovascolari si è avuta per una PAD di 82,6 mmHg. Come si è arrivati a queste conclusioni? Semplicemente osservando a quali valori di PAD si avevano meno eventi ma indipendentemente dal gruppo in cui si trovavano i pazienti. Ciò significa che alcuni dei pazienti che avevano avuto meno eventi si trovavano nel gruppo A, altri nel gruppo B, altri ancora nel gruppo C. In altre parole si è fatta una analisi dei risultati prendendo i pazienti un po’ di qua e un po’ di là e non secondo il gruppo a cui erano stati randomizzati, quindi non in base alla intenzione a trattare. Così succede che lo studio HOT viene comunemente citato a dimostrazione di maggior efficacia della terapia aggressiva dell’ipertensione, ma si ignora che in realtà il trial ha avuto esito negativo.

Il secondo esempio è uno studio, pubblicato da ricercatori canadesi, sulla efficacia dello screening del cancro della prostata. In questo caso quindi non si tratta di terapia farmacologica ma di un intervento diverso (screening). Lo studio suggeriva che lo screening è efficace nel ridurre la mortalità. Tuttavia solo il 23,1% dei soggetti invitati allo screening aveva risposto, mentre quelli che non avevano risposto erano stati inseriti nel gruppo non screenato. In tal modo si è creato un evidente bias di selezione per cui i due gruppi (screenati e no) non erano paragonabili. Ormai dovrebbe essere chiaro che cosa vuol dire questa espressione: è evidente che chi risponde ad un invito allo screening è di solito più giovane e più in salute di chi non risponde, pertanto “non è leale" paragonare i due gruppi. Si doveva invece prendere chi aveva risposto e randomizzare costoro in due gruppi, uno sottoposto allo screening e uno che funzionava da controllo. Inoltre quasi 1000 pazienti che originariamente facevano parte del gruppo controllo furono in seguito sottoposti allo screening e gli autori allora li inserirono nel gruppo screenato, analizzando i dati non più secondo l'intention to treat (in questo caso sarebbe più giusto dire "intention to screen"). Lo studio venne infatti ampiamente criticato, proprio per questi gravi errori metodologici. Ci fu anche

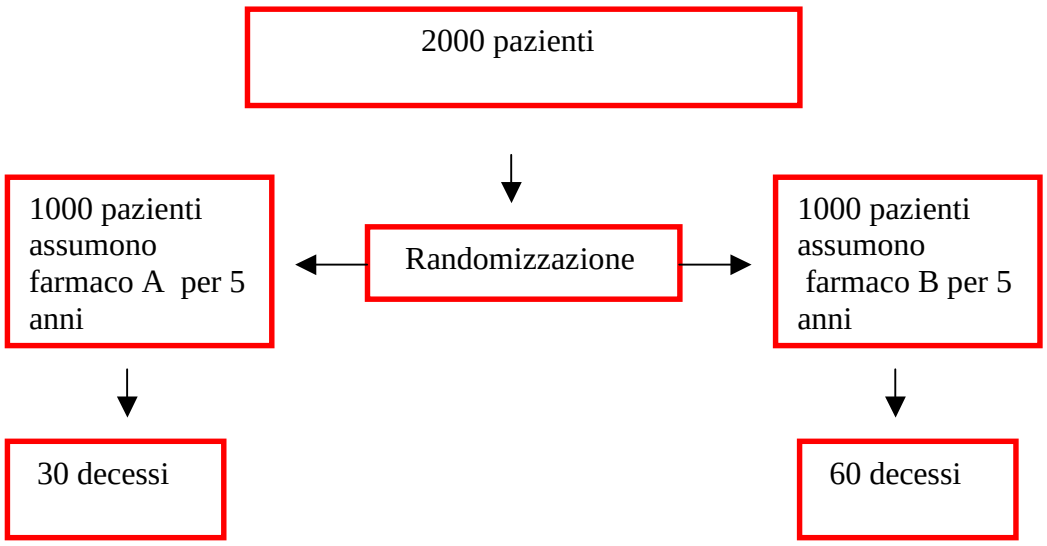
Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

chi si prese la briga di rianalizzare i dati secondo "l'intention to screen": la riduzione della mortalità nel gruppo screenato scompariva come per incanto.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 5
Le misure di efficacia

E' venuto il momento, purtroppo, di introdurre i numeri. Lo so, molti non si trovano a loro agio, complici forse vecchie reminiscenze di liceo. Per questo motivo cercherò di rendere l'esposizione il più semplice possibile, avvalendomi di tabelle ed esempi. Il discorso parte da una domanda: una volta disegnato e portato a termine un RCT abbiamo il problema di come esprimere i risultati ottenuti, in altre parole di quali misure e quali numeri usare per esprimere l'efficacia di un determinato trattamento. Supponiamo di arruolare 2.000 pazienti per vedere se il farmaco "A" riduce i decessi rispetto al farmaco "B". Suddividiamo i pazienti in modo randomizzato in due gruppi di 1000 ciascuno. Ad un gruppo somministriamo per 5 anni il farmaco "A", all'altro il farmaco "B". Dopo 5 anni andiamo a contare quanti decessi si sono verificati nei due gruppi. I risultati siano che nel gruppo A ci sono stati 30 decessi e nel gruppo B ci siano stati 60 decessi. Lo schema sottostante riassume i risultati dello studio.



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Come si vede la percentuale di decessi è di 30 su 1.000 (cioè 3 su 100) nel gruppo trattato con "A" e di 60 su 1.000 (cioè 6 su 100) nel gruppo trattato con "B".

Un primo modo di esprimere i risultati dello studio è quello di far riferimento al **numero assoluto** di eventi che si sono verificati ogni cento pazienti trattati nei due bracci dello studio. In questo caso avremo che nel braccio A i decessi sono stati il 3% e nel braccio B il 6%. Questa percentuale viene detto ***rischio assoluto*** e si esprime con la sigla AR (dall'inglese Absolute Risk). Mi pare che fin qui le cose siano semplici e lapalissiane: il rischio assoluto di morte col farmaco "A" sarà del 3%, con il farmaco "B" sarà del 6%. Questo vuol dire che la riduzione del rischio assoluto (ARR = Absolute Risk Reduction) ottenuta con il farmaco "A" rispetto al farmaco "B" sarà data dalla differenza tra i due rischi assoluti (6% - 3% = 3%).

Naturalmente siccome vogliamo fare le cose per bene useremo il gergo statistico e scriveremo che l'ARR ottenuta con "A" sarà uguale al 3%

La tabella sottostante riassume questo primo modo di esprimere i risultati di uno studio.

Risultati dello studio espressi in termini di Rischio Assoluto

AR di A = 3%
AR di B = 6%
ARR = 3%

Il rischio assoluto si calcola dividendo il numero degli eventi per il numero dei pazienti arruolati
Braccio A = 3 diviso 100 = 0,03 (3%)
Braccio B = 6 diviso 100 = 0,06 (6%)

La riduzione del rischio assoluto si calcola sottraendo il rischio assoluto di A al rischio assoluto di B = 6% - 3% = 3%

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Si noti che la percentuale si può esprimere (anzi spesso la troverete espressa in questo modo) con il corrispondente numero decimale. Così 3% diventa 0,03 e 6% diventa 0,06.

Ma vi è un altro modo di esprimere i risultati dello studio, molto importante in quanto permette di **paragonare** l'efficacia di vari farmaci tra loro e soprattutto di valutare il reale impatto clinico del trattamento. Si tratta del cosiddetto ***NNT (Number Needed to treat).*** Come dice il termine stesso, l'NNT esprime il numero di pazienti che bisogna trattare per evitare un evento (in questo caso un decesso) in un determinato periodo di tempo. Il ragionamento è semplice. Riferendoci al nostro studio è evidente che se il farmaco "A" permette di risparmiare 3 decessi ogni 100 soggetti trattati rispetto al farmaco "B", per evitare un decesso bisognerà trattare circa 33 pazienti (cioè 100 diviso 3). L'NNT viene qualche volta riportato negli studi clinici, purtroppo non sempre. Se si rendesse necessario determinarlo lo si può fare con una semplice formula, come esemplificato nella tabella sottostante.

NNT = 100 diviso riduzione del rischio assoluto o ARR Nell'esempio: 100 diviso 3 = 33,3

Perché diciamo che l'NNT è un parametro molto importante? Perché un semplice numero permette di paragonare l'efficacia di vari trattamenti e di avere anche un'idea precisa del beneficio clinico ottenuto e della sua reale importanza.

Facciamo un esempio concreto: supponiamo di avere a disposizione un farmaco che riduce la percentuale di ictus quando somministrato ai pazienti diabetici. Tuttavia esaminando i dati degli studi e calcolando l'NNT ci accorgiamo che se lo diamo ai diabetici ipertesi bisogna trattarne 40 per 5 anni per avere un ictus in meno, se lo diamo ai diabetici non ipertesi per avere un ictus in

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

meno bisogna trattarne 150 per 5 anni. Nel primo caso avremo un NNT a 5 anni di 40, nel secondo caso l'NNT a 5 anni sarà di 150. Risulta facilmente comprensibile che nei diabetici ipertesi il farmaco è molto più efficace che nei diabetici non ipertesi perché per avere lo stesso risultato, nello stesso arco di tempo, devo trattare meno pazienti.

En passant, si noti che l'NNT in assoluto più favorevole è 1. Un NNT = 1 significa che si evita un evento per ogni paziente trattato. Per avere un NNT di 1 bisogna che nel gruppo trattamento ogni 100 pazienti ci siano zero eventi mentre nel gruppo controllo ogni 100 trattati ci devono essere 100 eventi.

Il calcolo dell'NNT permette di valutare anche un altro aspetto: **quanto pazienti saranno trattati inutilmente**. Infatti l'effetto di un farmaco è del tipo "tutto o nulla": l'ictus o il decesso o l'infarto ci sono o non ci sono. Ciò significa che con un NNT di 40 tratterò inutilmente (e li esporrò agli effetti collaterali della terapia) 39 pazienti, con un NNT di 150 il numero di pazienti trattati per niente sale a 149! In questo caso si parla di NNT – 1.

Purtroppo, come dicevo, spesso l'NNT non viene riportato negli studi (o perlomeno nell'abstract), ma ogni volta che ci vengono magnificate le virtù di un trattamento dovremmo sempre alzare la mano e domandare l'NNT. Così non di rado ci renderemmo conto di quanto siano poco efficaci terapie che magari pensiamo essere salvavita e di quanto modesti siano i risultati ottenuti nonostante tutti i nostri sforzi. Se posso esprimere un concetto semi-filosofico, direi che l'NNT permette di confrontarci con la nostra **piccolezza.**

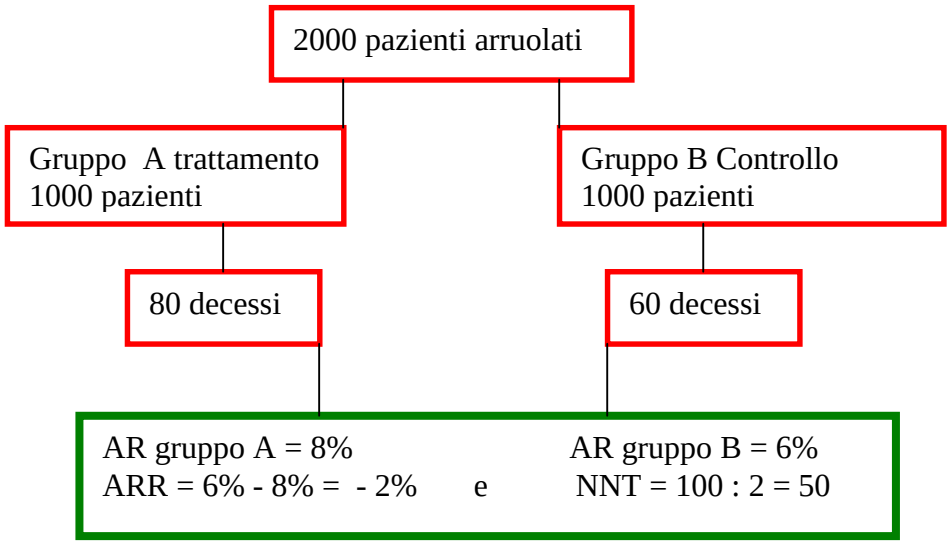
A questo punto conviene fare una considerazione perché **i numeri talvolta sono dei birbanti matricolati e ci fanno confondere.** Nell'esempio appena fatto abbiamo considerato uno studio in cui l'end-point era un evento negativo, cioè il decesso. Vediamo invece cosa succede ***quando l'end-point è un evento positivo.*** Mi direte: ma cosa stai blaterando, come può un end-point

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

essere positivo? E' possibile e l'esempio che segue lo dimostra. Supponiamo di voler testare contro placebo un farmaco che cicatrizza l'ulcera duodenale e di stabilire come end-point primario dello studio la percentuale di ***guarigioni endoscopiche*** dopo 4 settimane di terapia. In questo caso si va a valutare un evento positivo (la guarigione) e non uno negativo come i decessi. Cosa succede allora per il rischio assoluto e per l'NNT? Non succede nulla di particolare, solo che la riduzione del rischio assoluto sarà un numero negativo (perlomeno nell'ipotesi che il farmaco faccia meglio del placebo). Vediamo cosa succede al nostro studio sul farmaco che cicatrizza l'ulcera: si arruolano 100 soggetti per parte e dopo 4 settimane si vede che si è ottenuta la cicatrizzazione dell'ulcera in 30 pazienti del gruppo placebo e in 80 del gruppo trattamento: il rischio assoluto del placebo è 30% (= 0,30) e quello del farmaco è 80% (= 0,80). La riduzione del rischio assoluto (ARR) sarà data da 30% - 80% = - 50% (= -0,50). Come si vede l'ARR è un numero negativo ma è del tutto logico se si pensa che l'end-point valutato era un evento positivo (la percentuale di guarigione delle ulcere) e non un evento negativo. Insomma ci viene buona la vecchia regoletta del liceo per cui ***"meno per meno fa più"*** e quella contraria del ***"più per meno fa meno***". Per l'NNT però le cose non cambiano: se ogni 100 pazienti se ne hanno 50 in più rispetto al placebo che guariscono dall'ulcera vuol dire che basta trattare due pazienti perché uno guarisca. L'NNT sarà quindi uguale a 2. Per il calcolo si farà sempre 100 diviso l'ARR che però dovrà essere scritto senza il segno meno davanti (100 diviso 50 = 2). E' più difficile da spiegare che da capire, basta solo farci un poco la mano senza lasciare che la fobia per i numeri prenda il sopravvento.

E se per caso un trattamento dovesse ottenere risultati peggiori del controllo? Lo vediamo subito.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati



Come si capisce facilmente in questo caso si ha un evento in più ogni 50 soggetti trattati e non un evento in meno. Più correttamente, in questi casi, si parla non di NNT ma di NNH (Number Needed to Harm).Il concetto di NNH torna buono quando si vogliono valutare gli effetti collaterali di un farmaco. Ritorniamo al nostro esempio del farmaco antiulcera che cicatrizzava l'80% delle ulcere contro il 30% del placebo. I ricercatori però oltre a questo end-point principale ne vogliono valutare un altro, secondario, che è l'insorgenza di epatite acuta non virale, che potrebbe essere un effetto avverso del trattamento. Alla fine delle 4 settimane dello studio si nota che nel gruppo trattato con placebo si hanno zero casi di epatite acuta mentre nel gruppo in trattamento se ne hanno 10. Questo vuol dire che ogni 100 pazienti trattati si hanno 10 casi in più di epatite acuta rispetto al placebo, vale a dire che l'NNT (o meglio l'NNH) è di 100 diviso 10 = 10. In altre parole ogni 10 pazienti trattati si svilupperà una epatite acuta. A scanso di equivoci preciso che questo e, in generale, anche gli altri sono esempi inventati e non si riferiscono a nessun farmaco in particolare, servono solo a scopo didattico. Se si volesse allora riassumere il risultato complessivo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

dello studio potremmo dire che ogni 10 pazienti trattati cicatrizzo 5 ulcere in più rispetto al placebo (ricordo che l'NNT era = 2) ma nello stesso tempo avrò anche un caso di epatite acuta in più. Come si può facilmente comprendere questi calcoli sono di estrema importanza quando si deve valutare il ***profilo beneficio/rischio di un farmaco***.

Supponiamo per esempio di avere uno studio su un farmaco antitumorale per il cancro mammario che viene confrontato col placebo. L'end-point primario dello studio sia di valutare di quanto il farmaco riduce il rischio di recidiva neoplastica dopo 5 anni, ma gli studiosi vanno a misurare anche un end-point secondario, il numero di ictus fatali. Alla fine del follow-up si trova che il farmaco riduce in effetti le recidive di tumore mammario con un NNT di 10 ma nello stesso tempo si trova un NNH per ictus fatale di 12. Significa che ogni 100 pazienti trattate ce ne saranno 10 che avranno un beneficio in quanto non avranno una recidiva, ma nello stesso tempo ce ne saranno 8,33 che saranno morte per ictus. Si capisce bene come questi due semplici dati permettano di valutare più compiutamente l'impatto globale del trattamento sulla salute.

Però a questo punto mi aspetto un'obiezione. Mi direte: ma se nel capitolo sugli end-point ci hai rotto le scatole dicendo che bisogna sempre stare attenti a valutare gli outcomes secondari perché tutto il potere statistico dello studio è impegnato per quello primario e anche una eventuale significatività statistica di un end-point secondario è solo apparente! In effetti questa è la regola, tuttavia quando un end-point secondario è un evento collaterale, criterio di prudenza vuole che il dato venga preso per buono o comunque considerato con molta attenzione, persino nel caso la significatività statistica non ci sia.

Per finire possiamo introdurre anche un altro parametro dal nome difficile (ma non spaventatevi). Si tratta del cosiddetto **LLH** (Likelihood of being helped or harmed) che potremmo tradurre come la probabilità di trovare beneficio o di essere danneggiati da un farmaco. Il parametro si trova con questa formuletta: NNH/NNT (NNH diviso NNT).

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Facciamo un esempio per capirci. In uno studio un farmaco usato per la dispepsia non ulcerosa risulta utile in 40 pazienti su 100 per cui il suo NNT sarà di 2,5. Però nello stesso tempo può provocare cefalea in 25 pazienti su 100 per cui l'NNH sarà di 4.Per trovare l' LLH basta dividere 4 per 2,5 = 1,6. ***Possiamo dire che quando LLH > 1 il beneficio è superiore ad un determinato danno (o effetto collaterale), se LLH < 1 vuol dire che la probabilità di avere un effetto collaterale è superiore al beneficio, se LLH = 1 rischio e beneficio si equivalgono.***
Come si vede è un modo raffinato per mettere in relazione NNT ed NNH. Nel caso in esame LLH = 1,6 significa che il beneficio è superiore al danno ed è ovvio essendo l'NNT inferiore all'NNH (si ottiene un beneficio ogni 2,5 pazienti trattati e un danno ogni 4 pazienti trattati). Così nello stesso studio se avessimo avuto un altro effetto collaterale del farmaco (per esempio vertigini) con un NNH di 2, avremmo avuto un LLH = 2 diviso 2,5 = 0,8, quindi sfavorevole.E' possibile in questo modo calcolare LLH per tutti gli effetti favorevoli rispetto a quelli sfavorevoli a patto di poter conoscere i vari NNT ed NNH.

Vi è infine un terzo modo di esprimere i risultati degli studi, ed è purtroppo spesso il solo modo in cui questi ci vengono presentati. Si tratta del ***rischio relativo*** che esprime il rischio di evento del farmaco oggetto del trattamento **rispetto** al rischio che si è avuto con il controllo. Esiste una formula per calcolarlo, ma dato che probabilmente ce la dimenticheremo, conviene fissarsi in mente un metodo, che una volta imparato, non dovremmo più scordare. Dico subito che è un metodo non ortodosso che farà inorridire gli esperti, ma dato che funziona con me penso che andrà bene anche a voi. Riprendiamo il nostro esempio in cui con il farmaco "A" il rischio di decesso era del 3% e con il farmaco "B" era del 6%. Per comodità supponiamo che con il farmaco di confronto (nel nostro esempio il farmaco B) il rischio di decesso sia "1". In realtà sappiamo che non è così perché abbiamo visto che nel gruppo "B" il rischio di decesso è del 6%, però ***fingiamo***

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

(ripeto, per nostra comodità) che questo 6% sia uguale ad 1. Quindi ci domandiamo: fatto "1" il rischio con il farmaco B, qual è il rischio con il farmaco A? In altre parole se il 6% trovato con il farmaco B noi lo abbiamo fatto uguale a 1, il rischio con il farmaco A, che era il 3%, quanto sarà? La risposta mi sembra facile, ***rispetto*** a 1 del farmaco "B" il rischio con "A" è esattamente la metà (infatti 3% è la metà di 6%), cioè sarà 0,5. Esprimeremo tutto questo dicendo che il ***Rischio Relativo (RR = Relative Risk) di "A" rispetto a "B" è uguale a 0,5***. Faccio notare che dire 0,5 è come dire 50%. Quando troverò scritto che il tal farmaco è risultato più efficace del placebo con RR = 0,5, vuol dire che se con il placebo facciamo uguale a 1 il rischio di un determinato evento, con il farmaco tale rischio scende a 0,5, cioè al 50%. Nello stesso modo se troveremo scritto che RR = 0,75 vuol dire che il rischio col trattamento attivo è il 75% di quello avuto con il controllo, e così via. Qualche volta al posto di RR si trova scritto ***HR (Hazard Ratio)*** che per i nostri scopi possiamo tranquillamente considerare assimilabile al rischio relativo. In realtà mentre RR è un rapporto tra due rischi assoluti, l'HR è un rapporto tra due Hazard Rate. Qui bisognerebbe introdurre le curve di Kaplan-Meier ed il modello a rischi proporzionali di Cox, ma già i nomi sono tutto un programma ed io in statistica sono solo un praticone né voglio essere preso a bastonate dai miei lettori. L'argomento verrà, comunque, trattato succintamente nel capitolo 11.

La formula che permette di calcolare l'RR è mostrata nella tabella che segue.

Formula generale per il calcolo del Rischio Relativo o RR RR = Rischio Assoluto del farmaco diviso Rischio Assoluto del controllo Nell'esempio fatto: AR farmaco 3% AR controllo 6% RR = 3% diviso 6% = 0,5 (50%)

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Comunque se non ricordiamo la formula ci possiamo arrivare anche imbastendo una proporzione che si basa sul ragionamento di prima: AR controllo sta ad "1" come AR farmaco sta ad "X" (dove X è il rischio relativo che dobbiamo trovare). Alla fine si arriva sempre allo stesso risultato:

6 : 1 = 3 : X

X = 3 : 6 = 0,5 (= 50%)

L'RR permette poi di calcolare la ***Riduzione del Rischio Relativo*** (RRR = Relative Risk Reduction). Nel caso del nostro esempio essa è 0,5: infatti se abbiamo fatto uguale a 1 il rischio del controllo, il rischio relativo del trattamento è 0,5 e la riduzione del rischio deriverà dalla sottrazione: 1 meno 0,5 = 0,5. Tuttavia siccome gli esperti di statistica non amano le cose semplici l' RRR non viene mai espresso in questa maniera ma **nel suo corrispondente percentuale** per cui 0,5 = 50%. Diremo quindi che il trattamento ha ridotto il rischio relativo del 50%.

Così se in uno studio risulta un RR di 0,75 vuol dire che RRR è uguale al 25%, se RR è di 0,40 l'RRR è del 60%, se RR 0,91 l'RRR sarà del 9% e così via. Basta farci un po' la mano e poi la cosa viene automatica.

E se il farmaco fa peggio del controllo cosa succede per il rischio relativo? Riprendiamo l'esempio in cui abbiamo parlato di NNH. In quel caso i decessi nel gruppo trattato erano l'8% e nel gruppo controllo il 6%:

- **AR trattati 8%; AR controlli 6%**
- **RR = 8% diviso 6% = 1,333**

Come si vede ***l'RR è superiore a 1*** e non inferiore; un RR superiore a 1 significa che il farmaco testato si è ***comportato peggio del controllo;*** non si parla naturalmente di riduzione del rischio

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

relativo ma di ***aumento*** perché il trattamento ha provocato un incremento degli eventi del 33,3%.

Per riassumere, ecco alcune domande a cui sarà facile rispondere:

- RR = 0,82, cosa vuol dire? il farmaco ha fatto meglio o peggio del controllo, e di quanto?
- RR = 1,25, cosa vuol dire? il farmaco ha fatto meglio o peggio del controllo, e di quanto?

E infine una domandina da cento milioni (non di euro però): se l'RR è uguale a 1 cosa significa? Ha fatto meglio il farmaco testato o il controllo? Credo che tutti ormai sappiano la risposta: ***un RR = 1 significa che il match è finito alla pari.***

E' importante a questo punto sottolineare ancora una volta che purtroppo i risultati di uno studio vengono ***presentati troppo spesso solo in termini di riduzione del rischio relativo***. La cosa è comprensibile se si tiene conto che in questo modo l'entità del beneficio ottenuto viene molto enfatizzata. Nell'esempio che abbiamo fatto, un conto è dire che si è ottenuta una riduzione del rischio assoluto del 3% e un' altra è dire che si è ottenuta una riduzione del rischio relativo del 50%.

Ma la cosa è ancora più sottile: ***far riferimento al rischio relativo non permette di paragonare l'entità di due interventi.*** Come sempre un esempio ci viene in aiuto. Premetto che anche in questo come nelle altre esemplificazioni i numeri presentati sono del tutto di fantasia e non corrispondono ad alcuno studio reale, servono solo a scopo didattico. Pensiamo a un farmaco che riduca il rischio di frattura del femore, rispetto al placebo, in termini relativi del 50%. Il farmaco "A" viene confrontato con il placebo "P" in un primo studio in cui sono state arruolate 2000 donne in post-menopausa con pregressa frattura di femore e alla fine dello studio si contano 20 fratture nel gruppo in trattamento attivo e 40 nel gruppo placebo. Ormai siamo in grado di capire i risultati espressi nella tabella sottostante.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

2000 donne in post-menopausa con pregressa frattura trattate con “A” o “P”	Gruppo trattato con “A” 1000 donne	Gruppo trattato con “P” 1000 donne
Numero di fratture dopo 5 anni	20	40
Rischio assoluto (**AR**)	2%	4%
Riduzione del rischio assoluto (**ARR**)	2%	
Rischio relativo (**RR**)	0,5 (= 50%)	
Riduzione del rischio relativo (**RRR**)	50%	
NNT	50	

Vediamo invece cosa succede per il farmaco “A”, confrontato sempre con il placebo “P”, in uno studio in cui sono state reclutate 10.000 donne in post-menopausa ma **senza** pregressa frattura di femore, trattate sempre per 5 anni e in cui ci siano state 30 fratture nel gruppo in trattamento attivo e 60 nel gruppo placebo. Costruiamo ancora la nostra tabella.

10.000 donne in post-menopausa con pregressa frattura trattate con “A” o “P”	Gruppo trattato con “A” 5000 donne	Gruppo trattato con “P” 5000 donne
Numero di fratture dopo 5 anni	30	60
Rischio assoluto (**AR**)	0,6%	1,2%
Riduzione del rischio assoluto (**ARR**)	0,6%	
Rischio relativo (**RR**)	0,5 (= 50%)	
Riduzione del rischio relativo (**RRR**)	50%	
NNT	166	

Come si può facilmente vedere in entrambi gli studi il farmaco “A” ottiene una riduzione del rischio relativo di frattura rispetto al placebo del 50%, ma nel primo studio, quello in cui erano arruolate donne con pregressa frattura, basta trattare 50 donne per evitare un evento (NNT = 50), nel secondo studio, in cui sono state arruolate donne senza pregressa frattura, occorre trattare per 5 anni ben 166 donne per evitare un evento (NNT = 166). Se ci limitassimo solo all' RRR potremmo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

concludere che il farmaco è ugualmente efficace sia in prevenzione secondaria che primaria perché la riduzione del rischio relativo è sempre del 50%, in realtà è circa tre volte più efficace in prevenzione secondaria perché per evitare un evento basta trattare un terzo delle donne che invece occorre trattare in prevenzione primaria. E' ovvio che chi ha interesse a magnificare l'efficacia di un trattamento farmacologico tenderà a presentare i risultati in termini di riduzione del rischio relativo e non di riduzione del rischio assoluto e di NNT.

Quando ci vengono presentati i risultati di uno studio solo con la riduzione del rischio relativo

dovremmo chiedere poche cose: quanti erano i trattati e quanti gli eventi nei due gruppi. Con questi pochi dati ora siamo in grado di stabilire da soli i reali benefici che ci possiamo attendere dal trattamento. Ma vi avverto subito, non sarà affatto facile ottenere questi numeri, anzi quasi di sicuro non li otterrete per niente, bisognerà prendere il lavoro originale e andarseli a cercare. Purtroppo neppure nell'abstract questi numeri vengono spesso riportati ma ci si limita al rischio relativo. Insomma per sapere tutto bisogna fare un po' di fatica!

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 6
La differenza è significativa?

Gli esami non finiscono mai, diceva qualcuno. Nel nostro caso potremmo dire che quelle che non finiscono mai sono le difficoltà. Una volta che abbiamo espresso i risultati dello studio che stiamo esaminando in termini matematici, dobbiamo affrontare una questione cruciale: **la differenza che abbiamo trovato tra il trattamento e il placebo (o il farmaco di confronto) è significativa dal punto di vista statistico oppure è dovuta semplicemente al caso?** Si tratta di una domanda fondamentale, come si può capire, perché solo se la differenza è significativa possiamo attribuire un reale valore clinico al risultato che abbiamo trovato (non è sempre detto che ad un risultato statisticamente significativo corrisponde un reale beneficio clinico, come vedremo in seguito, ma per ora accantoniamo questo aspetto).

Come si fa a dire che il risultato è statisticamente significativo? Come possiamo essere sicuri che il dato non è dovuto al semplice gioco capriccioso della casualità ? Affermare che un risultato è statisticamente significativo mentre è puramente dovuto al caso viene definito dagli esperti in statistica ***"errore alfa***". Gli studiosi si sono messi d'accordo e dicono: accettiamo una probabilità di errore alfa inferiore al 5%; se questa probabilità è inferiore al 5% possiamo dire che il risultato è statisticamente significativo. A noi non interessa sapere come viene calcolata questa probabilità di errore, visto che il nostro mestiere è quello di fare i medici e non i matematici. Importa invece conoscere come si esprime. Ebbene si esprime con la **famosa "P**".

> La "P" esprime la probabilità che un risultato sia dovuto al caso. Se questa probabilità è inferiore al 5% il risultato si definisce statisticamente significativo

Perciò quando troviamo che un determinato farmaco ha ridotto il rischio di infarto di una certa

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

percentuale con “P” < 0,05 significa che il dato è significativo perché la “P”, cioè la probabilità di errore alfa, è inferiore al 5% (ricordo che dire 0,05 e dire 5% è la stessa cosa espressa in modo diverso).

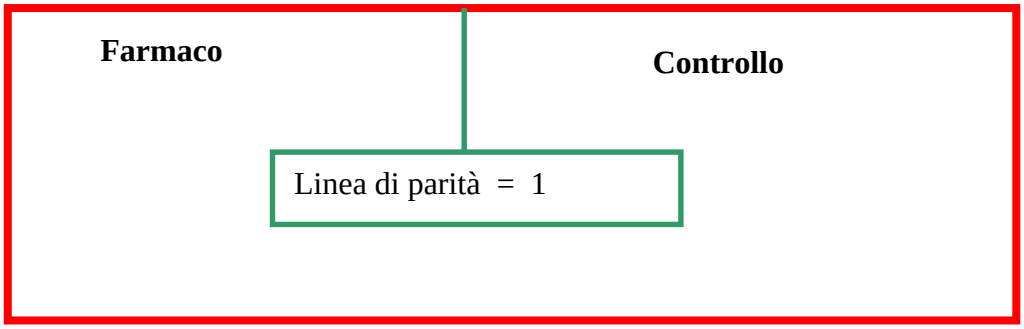
Vi sembra poco una probabilità di errore alfa del 5%? Se uno studio dimostra che un farmaco riduce l'ictus di un tot rispetto al placebo e che il risultato è statisticamente significativo perché la P è uguale a 0,049 vi sentireste tranquilli? Una P = 0,049 significa che c'è il 4,9 % di probabilità che quel risultato sia semplicemente casuale. Formalmente il risultato è statisticamente significativo ma proviamo a metterla in altro modo: voi state comodamente viaggiando con la vostra auto e ad un certo punto dovete attraversare un ponte e vi accorgete che all’imbocco qualche bontempone ha messo un bel cartello con su scritto: "Attenzione questo ponte crolla all’incirca ogni 20,4 passaggi" (in altre parole c'è una probabilità del 4,9% che crolli quando state passando voi). Vi sentireste tranquilli? Vi sentireste tranquilli perché la probabilità è di un crollo ogni 20,4 passaggi invece che ogni 20? Questo per dire che tanto più bassa è la P tanto più possiamo stare tranquilli che il caso non ci ha messo lo zampino. Però la significatività statistica non è altro che una invenzione dell’uomo, una convenzione: che differenza vi può essere dal punto di vista clinico tra una P di 0,049 e una di 0,051? Eppure la prima permette di affermare la significatività statistica, la seconda no!!

En passant dirò che vi è anche l'***errore beta***, che è l'opposto dell'errore alfa: si considera statisticamente non significativo un risultato che invece lo è. Gli studiosi si sono messi d'accordo per accettare una probabilità di errore beta inferiore a 0,1 (cioè 10%).

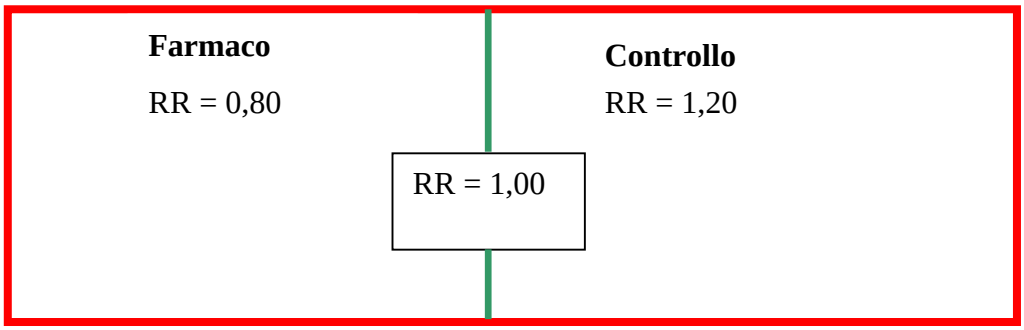
E' arrivato adesso il momento di complicare ancor più la faccenda e di spiegare il famigerato

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

intervallo di confidenza. Per farlo ritorniamo un attimo indietro all'RR, cioè al rischio relativo. Ricorderete che esso può essere inferiore a 1 (l'intervento fa meglio del controllo), superiore a 1 (l'intervento fa peggio del controllo) oppure può essere uguale a 1 (intervento e controllo sono uguali). Adesso immaginiamo di costruire una tabella a due colonne in cui nella colonna di sinistra si mette il farmaco oggetto dello studio e nella colonna di destra il farmaco di confronto o il placebo; la linea centrale verticale che separa le due colonne corrisponda al valore 1.



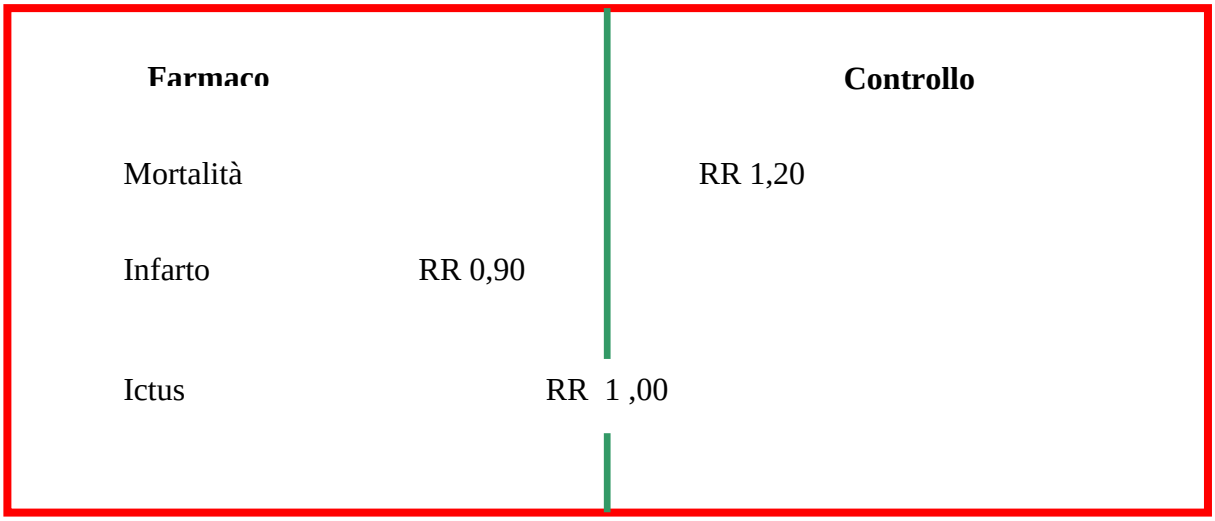
Nella rappresentazione grafica dell'RR, se questo è inferiore a 1 il valore andrà nella colonna di sinistra, dove c'è il farmaco, se RR è maggiore di 1 andrà nella colonna di destra dove c'è il controllo. Se infine l'RR fosse uguale a 1 esso corrisponderebbe alla linea verticale che divide le due colonne.



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 6 – La differenza è significativa?

Proviamo adesso a rappresentare graficamente i risultati di uno studio in cui un farmaco antipertensivo è stato confrontato con un farmaco di riferimento e siano stati valutati i seguenti end-point: mortalità totale, infarto non fatale, ictus.



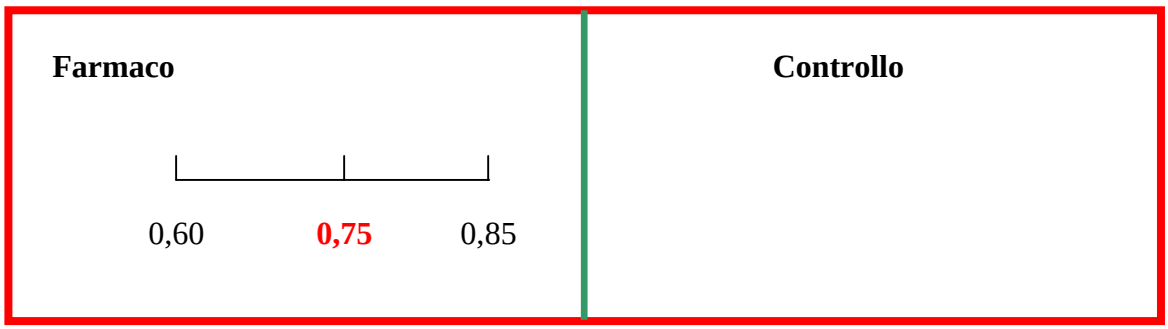
La lettura della tabella è la seguente:

- il farmaco testato ha fatto peggio del controllo sulla mortalità totale: un RR = 1,20 significa che con il farmaco la mortalità è aumentata del 20%
- il farmaco testato ha fatto meglio del controllo per quanto riguarda l'infarto non fatale, riducendo il rischio del 10%
- non c'è differenza tra farmaco e controllo per quanto riguarda l'ictus.

Però, per la serie le cose semplici non vanno bene, dobbiamo considerare che l'RR che abbiamo trovato è una stima unitaria del rischio, ma in realtà non è così, non può essere un singolo numero. Per farla breve dobbiamo immaginare di ripetere per 100 volte la stima dell'RR (non spaventatevi,

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

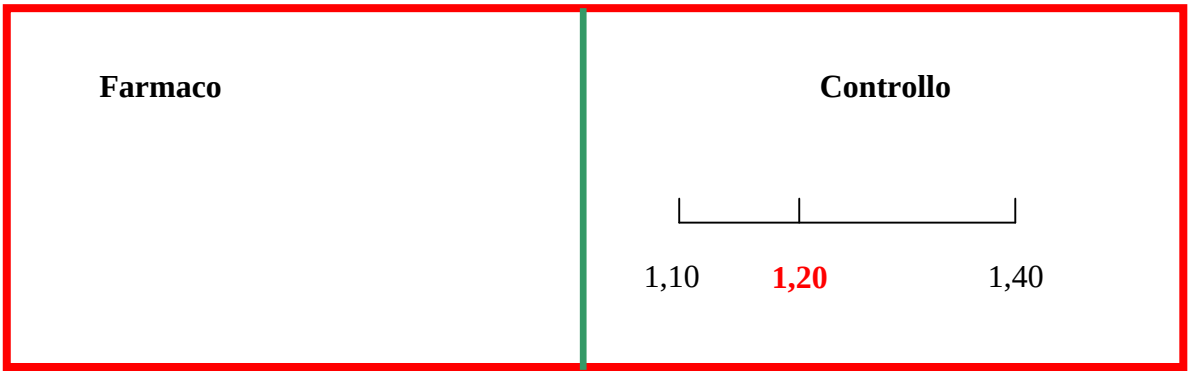
non spetta a noi farlo). Così facendo troveremo 100 RR diversi, il 95% di questi RR viene a trovarsi entro un certo intervallo che viene detto intervallo di confidenza al 95% (IC95%). Lo so, così non è troppo chiaro, e allora facciamo un esempio. Supponiamo di trovare un RR = 0.60. Ripetiamo per 100 volte la stima dell'RR e troveremo 100 valori e poniamo che 95 di questi valori siano compresi tra 0.50 e 0.70, mentre 5 valori cadranno fuori da questo intervallo, cioè potranno essere più piccoli di 0.50 o più grandi di 0.70. Allora diremo che l'RR trovato è di 0.60 ma il suo IC95% va da 0.50 a 0.70. Detto in altro modo **l'intervallo di confidenza al 95% esprime il range entro cui può trovarsi l'RR con una probabilità del 95%.** Per capire ancora meglio l'IC95% possiamo pensare all'RR come ad un treno che viaggi da Napoli a Milano. A una certa ora noi sappiamo che il treno dovrebbe trovarsi in una stazione intermedia tra Roma e Firenze, ma in realtà non sappiamo esattamente quale sia per cui diciamo che con una probabilità del 95% si trova comunque in un tratto compreso tra Firenze e Roma. Facciamo un altro esempio. In un trial si trova che un farmaco riduce la probabilità di sviluppo di cancro mammario del 25%, quindi l'RR sarà 0.75, ma il suo IC95% va da 0.60 a 0.85. Ciò vuol dire che in realtà la riduzione del rischio (con una precisione del 95%) potrebbe, nella migliore delle ipotesi, essere del 40% (RR = 0.60) e nella peggiore del 15% (RR = 0.85). La tabella che segue esprime questo esempio.



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Come si vede chiaramente dalla rappresentazione grafica sia l'RR che il suo IC95% sono sempre inferiori a 1, quindi sono sempre nella colonna di sinistra. L'RR, pur spostandosi all'interno del suo intervallo, non potrà mai essere uguale o superiore a 1. Si può affermare pertanto che il risultato così trovato è statisticamente significativo perché non potrà mai essere che il farmaco faccia venire più tumori mammari del controllo. Tutto questo si trova scritto negli studi come segue: **RR 0.75; IC95% 0.60-0.85.**

Prendiamo un esempio opposto: in un trial un farmaco non solo non riesce a ridurre il rischio di infarto rispetto al controllo, ma addirittura lo fa aumentare del 20%. Pertanto l'RR sarà 1.20. Il suo IC95% trovato va da 1.10 a 1.40. Ciò vuol dire che nella migliore delle ipotesi l'aumento del rischio è del 10%, nella peggiore del 40%. La rappresentazione grafica sarà quella schematizzata nella tabella che segue.

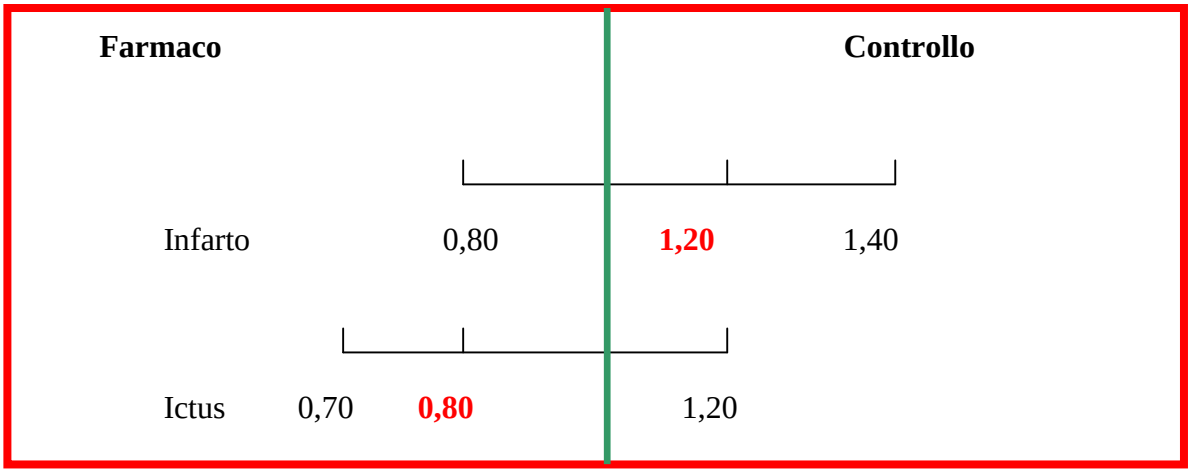


In questo caso l'RR sarà sempre superiore a 1 perché il suo IC95% non interseca mai la linea verticale dove sta' la parità. Anche qui diremo che il risultato così trovato è statisticamente significativo perché non potrà mai essere che il farmaco faccia venire meno infarti del controllo. Tutto questo si trova scritto negli studi come segue: **RR 1.20; IC95% 1.10-1.40.**

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 6 – La differenza è significativa?

Prendiamo infine il terzo e ultimo caso, quello in cui l'IC95% intersechi la linea verticale dove sta la parità. Poniamo che uno studio abbia valutato l'efficacia di un farmaco nel ridurre il rischio di infarto e di ictus. I risultati trovati sono rappresentati nella tabella che segue.



Come si può ormai facilmente arguire, per quanto riguarda l'infarto il farmaco ha provocato un aumento del rischio del 20% (RR = 1.20), ma il suo intervallo di confidenza interseca la linea di parità, quindi potrebbe anche essere che il rischio venga ridotto del 20%. Per quanto riguarda l'ictus il farmaco riduce il rischio del 20% (RR = 0.80) ma il suo intervallo di confidenza interseca la linea di parità e il rischio, per quanto ne sappiamo, potrebbe anche essere aumentato del 20%.

Diremo allora che la differenza trovata per i due end-points **non è statisticamente significativa**. In effetti non sappiamo realmente se il rischio aumenti o si riduca dato che l'intervallo di confidenza interseca la linea di parità e quindi potrebbe essere maggiore o minore dell'unità.

Gli esempi e le tabelle dovrebbero essere chiari, ma se non avete compreso vi conviene rileggerli perché si tratta di un punto fondamentale per capire i risultati degli studi così come poi li troverete riportati.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 6 – La differenza è significativa?

In conclusione possiamo dire che l'IC95% ci permette di vedere se l'RR è statisticamente significativo: se comprende il numero 1 non lo è, se al contrario non lo comprende lo è.

Riporto ora i risultati del Women's Health Study, uno studio randomizzato e controllato in cui erano state arruolate quasi 40.000 donne di almeno 45 anni e senza storia di malattie cardiovascolari o neoplastiche. Le pazienti vennero trattate con aspirina oppure placebo e seguite per oltre 10 anni. Lo scopo dello studio era di stabilire se l'aspirina fosse in grado di ridurre l'incidenza di neoplasie. Il risultato fu il seguente: sviluppo totale di neoplasie: RR 1.01; CI95% 0.94-1.08 (vi fu un aumento del numero totale di neoplasie dell'1% nel gruppo aspirina, ma il dato non è significativo, quindi i due trattamenti hanno pareggiato). Lascio ora a voi l'interpretazione degli altri dati:

Cancro mammario: RR 0.98; IC95% 0.87-1.09
Cancro del colon: RR 0.97; IC95% 0.77-1.24
Cancro polmonare: RR 0.78; IC95% 0.59-1.03
Mortalità totale: RR 0.95; IC95% 0.81-1.11
Mortalità da cancro polmonare: RR 0.70; IC95% 0.50-0.99

Siccome le cose appena dette sono troppo semplici qualcuno si diverte a complicarle e invece di riportare i risultati di uno studio secondo la formula classica li riporta in questo modo:

Riduzione del rischio relativo del 25% (dal 10% al 45%)

Conviene però non lasciarsi spaventare: la riduzione del rischio relativo del 25% corrisponde ad un RR di 0,75 mentre i due numeri tra parentesi corrispondono ai due estremi dell'intervallo di

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

confidenza, dove 10% corrisponde a 0,90 e 45% a 0,55%. Così se vogliamo usare la modalità solita di esprimere i risultati scriveremo: RR 0,75 (IC95% 0,55 – 0,90).

Nello stesso modo se troviamo scritto: Riduzione del rischio relativo del 15% (da – 10 a 25) sappiamo che RR = 0,85 con IC95% da 0,75 a 1,10.

Ovviamente così come è possibile calcolare l'intervallo di confidenza per il rischio relativo, lo si può fare anche per la riduzione del rischio assoluto e per l'NNT. Prendiamo questo esempio: in uno studio l'intervento ha ridotto l'end-point considerato, in valori assoluti, del 2,5% (ARR = 0,025) e quindi si ha un NNT di 40. Se l'intervallo di confidenza dell'ARR varia da 1% a 5% (o se si preferisce da 0,01 a 0,05) il corrispondente NNT varierà da 20 (ARR 5%) a 100 (ARR 1%). Questo vuol dire che nella migliore delle ipotesi basta trattare 20 persone per evitare un evento, nella peggiore bisogna trattarne 100.

Come si può immaginare l'IC95% può essere largo, stretto, vicino o lontano dalla linea di parità. Questo può dipendere da vari fattori come per esempio la numerosità della popolazione arruolata nello studio, la sua potenza statistica, ecc. Però quello che interessa a noi è un'altra cosa, cioè la possibilità di trarre delle conclusioni di tipo clinico dalla morfologia e dalla posizione dell'IC95% rispetto alla linea di parità.

Se l'intervallo di confidenza è largo e una delle estremità si avvicina alla linea di parità (per esempio 0.40 - 0.98 oppure 1.03-2.04) significa che il risultato è poco riproducibile e il suo significato clinico probabilmente poco importante. Invece se l'IC95% è stretto e lontano dalla linea di parità (per esempio 0.35-0.55 oppure 1.60-1.80) significa che il risultato è riproducibile e probabilmente anche importante dal punto di vista clinico.

Infine se è stretto ma vicino alla linea di parità (per esempio 0.88-0.99 oppure 1.02-1.12) significa

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

che il risultato è riproducibile ma clinicamente forse poco importante.

Insomma possiamo dire che tanto più l'IC95% è stretto tanto più elevata è la riproducibilità del risultato mentre tanto più è lontano dalla linea di parità tanto più cresce l'importanza clinica del dato. Al contrario tanto più una delle estremità dell'intervallo di confidenza è prossima alla linea di parità probabilmente tanto meno importante è, dal punto di vista clinico, il risultato. Infatti se troviamo che con un certo farmaco si è avuta la riduzione degli infarti del 30% (RR = 0,70) ma l'IC95% varia da 0,50 a 0,99 potremo certamente dire che il risultato è statisticamente significativo ma in realtà non sappiamo se la riduzione sia del 50% o dell'1%. Per motivi precauzionali conviene tener per buona l'ipotesi peggiore e considerare l'impatto clinico del farmaco, al più, modesto.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 7
L’odds ratio (OR)

Ragazzi preparatevi perché adesso viene un argomento bello tosto, il famoso (o famigerato, fate voi) **odds ratio** (conosciuto come OR). Intanto cominciamo a definire cosa è **l'odds.** Io traduco il concetto con "probabilità" anche se agli esperti probabilmente si rizzeranno i capelli e diranno che non è così. Ma per i nostri scopi basta e avanza. L'odds si richiama in qualche modo alla passione tutta anglosassone delle scommesse ed è praticamente la probabilità che un evento succeda rispetto alla probabilità che non succeda.

Ricordate cos'è il rischio assoluto? E' il numero di eventi rispetto al numero totale dei trattati. Nell'esempio che avevamo fatto all'inizio del nostro viaggio il rischio assoluto era del 3% per il farmaco e del 6% per il controllo. Con il farmaco c'erano 3 eventi per 100 trattati, con il controllo c'erano 6 eventi per 100 trattati. Come ricorderete il rischio assoluto si calcola dividendo il numero di eventi per numero dei trattati. Se ho 250 trattati con un farmaco e 22 eventi il rischio assoluto si calcola facendo 22 diviso 250 = 0.088 (= 8.8%). Ciò significa che si verificano 8.8 eventi ogni 100 trattati.

L'odds assomiglia un poco al rischio assoluto, con la differenza che invece di considerare tutti i trattati **si considerano quelli che non hanno avuto l'evento**. Nell' esempio precedente su 250 trattati si avranno 22 eventi e 228 non eventi. L'odds si calcola facendo 22 diviso 228 = 0.096 (= 9.6%). Significa che c'è una probabilità del 9.6% che l'evento succeda rispetto alla probabilità che non succeda (infatti 22 è il 9.6% di 228).

Come si può vedere **l'odds è sempre più grande di AR (cioè del rischio assoluto)** e questo è ovvio perché al numeratore va lo stesso numero ma al denominatore va un numero più piccolo (il numero dei non eventi al posto del numero totale dei trattati). Si capisce perché l'odds sia legato al mondo delle scommesse: un conto è dire che la probabilità di vincere è dell'8.8%, un altro far

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

credere, con l'escamotage della riduzione del denominatore, che è del 9.6%.

Naturalmente in uno studio, come si possono calcolare due rischi assoluti (uno per il braccio trattamento e uno per il braccio controllo) così si possono calcolare due odds. Lo so, purtroppo è un poco complicato, ma con un briciolo di pazienza e di attenzione la faccenda potrà diventare più comprensibile.

La tabella che segue riassume con un esempio questi concetti.

	Numero trattati	Numero eventi	Rischio assoluto	Odds
Trattamento	100	5	5/100 =0,05	5/95 = 0,052
Controllo	100	8	8/100 = 0,08	8 /92 = 0,086

Definito cos'è e come si determina l'odds veniamo all' odds ratio (abbreviato con OR): esso non è altro che il corrispondente del Rischio Relativo, solo che per calcolarlo si usa l'odds invece che il rischio assoluto.

Così mentre il Rischio Relativo, nell'esempio citato, viene dalla divisione 5% (Rischio Assoluto trattamento) diviso 8% (Rischio Assoluto controllo) = 0.625, il corrispondente OR viene dalla divisione 5,2% (odds farmaco) diviso 8,6% (odds controllo) = 0.60.

Come si vede l'OR è, in questo caso molto simile come grandezza, al Rischio Relativo. Così come RR anche OR può essere minore di 1 (l'intervento è favorevole), maggiore di 1 (l'intervento è sfavorevole) o uguale a 1. Nello stesso modo OR avrà un suo IC95% e le cose dette per l'intervallo di confidenza di RR valgono anche per l'intervallo di confidenza di OR.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Possiamo approssimare quindi che RR ed OR siano equivalenti, ma questo vale solo se il numero di eventi è piccolo rispetto al numero dei trattati perché allora quest'ultimo è molto simile, come grandezza, al numero dei non eventi.

Ma cosa succede se il numero di eventi è grande rispetto al numero dei trattati? Lo vediamo con l'esempio riassunto nella tabella seguente.

	Numero trattati	Numero eventi	Rischio Assoluto	Odds
Trattamento	100	30	30/100 = 0,3	30/70 = 0,42
Controllo	100	50	50/100 = 0,5	50/50 = 1

Rischio relativo = 0,3/0,5 = 0,6
Odds Ratio = 0,42/1 = 0,42

Come si vede in questo caso **OR è molto più piccolo di RR**. Possiamo dire quindi che OR ed RR praticamente corrispondono quando il numero degli eventi è piccolo rispetto al numero dei pazienti arruolati mentre OR diventa più piccolo di RR quando il numero degli eventi è grande rispetto al numero dei trattati. L'OR funziona quindi come una specie di *lente di ingrandimento* dei risultati ottenuti più potente di RR quando ci sono molti eventi rispetto ai trattati: un conto è dire RR = 0,60, un altro OR = 0,42. Al contrario quando gli eventi sono pochi rispetto ai trattati ci accorgiamo meno di questo potere di enfatizzazione e possiamo approssimare che RR ed OR quasi si equivalgano. A questo punto mi chiederete: perché complicare le cose e usare l'OR? non bastava l'RR? In realtà l'OR possiede delle proprietà matematico-statistiche (che qui non vale assolutamente la

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

pena di analizzare, anche perché avrei delle difficoltà a farlo) che l'RR non ha e che lo rendono particolarmente utile nelle meta-analisi. Tanto per dire, basti pensare che mentre il Rischio Assoluto può andare da 0 a 1 (vale a dire da 0% a 100%), l’odds può andare da 0 a infinito. Infatti se ci sono 100 trattati con zero eventi sia il Rischio Assoluto che l’odds si calcolano con la divisione 0/100 = 0, mentre se ci sono 100 trattati con 100 eventi il Rischio assoluto sarà 100/100 = 1 mentre l’Odds sarà 100/0 = infinito.

Mi fermo qui per non abusare della pazienza del lettore, anche perché, per i nostri scopi, le nozioni che ho fornito circa l’Odds Ratio sono più che sufficienti.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 8
Gli end-point composti

I ricercatori spesso non usano end-point singoli (per esempio il numero di decessi totali oppure il numero di infarti non fatali, ecc.) ma degli ***end-point composti*** (detti anche combinati). Di cosa si tratta? Un end-point composto è formato dalla somma di una serie più o meno numerosa di end-point singoli: in pratica è una specie di mega-evento costituito da vari eventi singoli. Un end-point composto per esempio può mettere insieme il numero di decessi totali, di stroke, di infarti non fatali e di interventi di rivascolarizzazione coronarica.

Perché si usano gli end-point composti? Una prima ragione è che in questo modo ***si migliora la potenza statistica dello studio***: per valutare un singolo evento bisognerebbe reclutare un numero molto elevato di pazienti, aumentando il numero di eventi con l'escamotage dell'end-point composto si possono reclutare meno soggetti. Inoltre l'end-point composto permette di solito un ***follow-up più breve***. Ancora: l'end-point composto consente di ***evidenziare differenze statistiche*** che non sarebbero state rilevate valutando un end-point singolo.

Tuttavia questa metodologia può creare problemi interpretativi. E' quindi necessario che gli autori che costruiscono l'end-point composto agiscano secondo canoni ben stabiliti. Per prima cosa i singoli componenti dell'end-point composto devono essere ***importanti dal punto di vista clinico***.

Per esempio un end-point composto che contenga i decessi da ogni causa + gli stroke + gli infarti del miocardio presenta tutti eventi clinicamente rilevanti. Invece un end-point composto da infarto + riduzione della PCR + riduzione del fibrinogeno contiene da una parte un evento importante come l'infarto ma dall'altra due end-point surrogati di scarsa importanza ai fini clinici.

In secondo luogo l'end-point composto deve essere ***stabilito prima*** di iniziare lo studio e non essere cambiato in itinere o a posteriori. Così in uno studio sugli antipertensivi si decide di valutare un end-point composto da infarto + stroke ma durante lo svolgimento del trial gli autori pensano di

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

aggiungere anche i ricoveri per scompenso cardiaco: questo modo di "cambiare le carte in tavola" può creare problemi di interpretazione dei risultati.

Un altro punto importante è che, oltre a riportare i dati complessivi dell'end-point composto, si devono specificare anche i ***risultati dei singoli end-point***, in modo da poter valutare l'effetto del trattamento su di essi. Questo è essenziale perché se si trova una riduzione dell'end-point combinato ed una contemporanea riduzione degli end-point singoli che lo compongono il risultato è attendibile, se invece non vi è corrispondenza tra l'end-point composto e i singoli elementi può essere più difficile interpretare correttamente i risultati. In altri termini se i vari elementi di un endpoint composto hanno un comportamento simile, cioè vanno tutti nella stessa direzione, questo significa che la loro scelta è stata probabilmente corretta ed è a favore del risultato trovato. Per esempio in uno studio l'end-point composto da infarto + angina risulta ridotto e parimenti risultano ridotti sia gli infarti sia gli episodi di angina: questo è attendibile. Se invece in uno studio si predispone un end-point composto da infarto miocardico non fatale + stroke + ricoveri per scompenso + interventi di by-pass coronarico e al termine del follow-up si trova che l'end-point composto è risultato ridotto ma solo grazie ad una riduzione cospicua degli interventi di rivascolarizzazione coronarica mentre gli altri end-point singoli non risultano ridotti o addirittura sono aumentati è lecito chiedersi quanto importante sia il risultato, tanto più se si considera che la decisione se sottoporsi o meno ad un intervento di rivascolarizzazione spesso dipende dalle scelte del medico (o anche del paziente), quindi è per certi versi soggettiva.

Vediamo un altro esempio di end-point composto di non semplice interpretazione: viene sperimentato contro placebo un farmaco antitrombotico che dovrebbe ridurre il rischio di tromboembolismo venoso dopo intervento di artroprotesi dell'anca. Si decide di valutare un endpoint composto da trombosi venosa sintomatica + trombosi venosa asintomatica riscontrata

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

con una flebografia praticata a tutti i pazienti arruolati nello studio da 10 a 15 giorni dopo l'intervento. L'end-point composto risulta ridotto in maniera statisticamente significativa ma andando ad esaminare i due singoli end-point si vede che non c'è nessuna riduzione delle trombosi venose clinicamente sintomatiche mentre vi è una riduzione significativa di quelle asintomatiche, diagnosticate grazie all'esame strumentale. Ora, è noto che dopo un intervento di artroprotesi all'anca le trombosi venose asintomatiche possono essere frequenti ma, pur non negando che possa essere utile ridurre anche questo tipo di evento, è evidente che lo scopo principale del trattamento è quello di ridurre gli episodi clinicamente sintomatici.

Un esempio reale ci viene da uno studio di prevenzione primaria in cui venne somministrata aspirina oppure placebo a donne sane. L'end-point primario era composto da mortalità cardiovascolare + infarto miocardico + stroke. Alla fine dello studio l' outcome primario composto non risultò ridotto nel gruppo trattato rispetto al placebo; analizzando i singoli elementi dell'endpoint composto si notò che mentre non risultavano ridotti né la mortalità cardiovascolare né l'infarto si aveva una riduzione dello stroke. Gli autori nelle loro conclusioni diedero una importanza eccessiva a questo risultato passando un po' sotto silenzio che in realtà lo studio aveva dato esito negativo per l'end-point primario composto, cioè quello su cui dovrebbe essere giudicato il trial. La riduzione dello stroke potrebbe essere reale oppure un semplice capriccio del caso visto che non è in linea rispetto agli altri due sottoelementi e all'end-point composto nel suo complesso. Il risultato quindi avrebbe dovuto essere interpretato con maggior cautela.

In alcuni casi l'uso di determinati end-point composti crea addirittura dei ***paradossi,*** come l'esempio che segue dimostra. Supponiamo di avere un nuovo farmaco contro l'osteoporosi e vogliamo dimostrare che funziona meglio di quelli classici usati finora. Decidiamo di valutare un end-point composto da fratture + numero di pazienti che mostrano un decremento della densità

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

ossea (BMD) misurata tramite densitometria. Alla fine dello studio il nuovo farmaco dimostra di ridurre l'end-point composto, ma andando a vedere i singoli end-point notiamo che la significatività viene raggiunta grazie all'endpoint surrogato (BMD) mentre il numero di fratture risulta addirittura più alto nel gruppo trattato che nel gruppo di controllo. In questo caso il ricorso ad un endpoint combinato permette all'outcome surrogato di ***"nascondere"*** un esito peggiore in quello che più ci interessa, cioè ridurre il rischio di fratture.

Un altro esempio, volutamente provocatorio, chiarirà ancora di più le idee: in uno studio su pazienti nefropatici si vuol valutare se un trattamento riduce un endpoint composto da raddoppiamento della microalbuminuria + aumento del 50% dei valori di creatinina + comparsa di uremia terminale definita come necessità di ricorrere alla dialisi + decessi totali. L'end-point composto risulta ridotto ma se si guardano i singoli endpoint si vede che a determinarlo sono solo i primi due mentre la comparsa di uremia terminale non si è ridotta e addirittura la mortalità totale risulta aumentata. Che valore dare alla riduzione dell'endpoint composto?

Altra confusione viene generata quando i ***componenti singoli molto simili hanno comportamenti tra loro conflittuali***: per esempio in un end-point composto da ictus fatale e non fatale si ha una riduzione dell'ictus non fatale e per contro un aumento di quello fatale. In questo caso anche se l'end-point composto dovesse essere ridotto come si deve valutare questo risultato?

Esaminare nel dettaglio questioni così complesse è compito di studiosi precipuamente esperti in "critical appraisal" ed è evidente che i medici pratici non possono avere né le competenze né il tempo per fare questo tipo di analisi. Purtroppo anche negli editoriali e nei commenti agli studi pubblicati dalle riviste più prestigiose non sempre si presta la dovuta attenzione a queste

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

problematiche. Né si può sempre fare affidamento sugli autori degli studi che di solito sono più interessati a presentare gli aspetti positivi del loro lavoro che le criticità presenti.

Qualcuno è arrivato ad affermare di non leggere più gli studi in cui ci sono endpoint composti. Senza giungere a decisioni così drastiche si può comunque consigliare di valutare con estrema cautela i risultati di un end-point composto, specialmente se questo contiene anche esiti surrogati oppure se i risultati dei componenti tra loro o quelli di un componente rispetto all'end-point composto in toto sono tra loro discordanti o ancora se gli autori non forniscono i dati relativi ai singoli componenti, rendendo in tal modo impossibile un'analisi separata.

E' interessante un'analisi di 114 RCT cardiovascolari pubblicata dal BMJ [BMJ 2007 Apr 14; 334:786] da cui risulta che l'uso di end-point composti è praticamente la regola in questo tipo di trials. Tuttavia in un terzo circa degli studi gli autori non riportano gli effetti del trattamento per i singoli end-point. Inoltre troppo spesso vengono mescolati, insieme ad end-point clinicamente importanti (come per esempio i decessi, gli infarti, gli stroke), anche end-point meno importanti dal punto di vista clinico (per esempio ricoveri per scompenso o per angina che tra l'altro possono essere influenzati dal comportamento dei medici e dei pazienti). In molti studi questi end-point di minor importanza sono gli unici ad essere ridotti mentre non lo sono quelli che maggiormente contano per i pazienti. Questo può portare il lettore ad interpretare in maniera sbagliata lo studio e ad attribuire al trattamento un'efficacia eccessiva estendendo i benefici visti per gli end-point meno importanti anche a quelli più importanti. Gli autori invitano quindi i medici a leggere con molta cautela gli studi ove sono previsti end-point composti e a valutare bene i risultati ottenuti su ogni singolo sotto-elemento.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 9
Pazienti persi al follow-up, studi di equivalenza e SAEs

In questo capitolo ci occuperemo di alcuni aspetti particolari degli RCT: cosa fare quando ci sono troppi pazienti persi al follow-up, le problematiche legate agli studi di equivalenza e la valutazione dei SAEs (che dovrebbe essere sempre prevista negli RCT, ma purtroppo non lo è).

Pazienti persi al follow-up

Quello dei pazienti persi al follow-up è un aspetto molto delicato da considerare quando si valuta uno studio. Mi limiterò a riportare l'esempio dello studio ISIS 2 (che ha analizzato l'efficacia dell'aspirina nell'infarto): in questo studio più di un quarto dei pazienti era persa al successivo follow-up. In altre parole non si sapeva che fine avessero fatto, se fossero vivi o morti, se avessero avuto o meno un infarto, ecc. Nessuno sa esattamente cosa bisognerebbe fare in questi casi ma, secondo alcuni autori, sarebbe necessario considerare i drop-outs (cioè i soggetti persi al follow-up) nel gruppo in trattamento come deceduti e quelli del gruppo controllo come viventi. Se si applicasse un tale approccio (troppo drastico?), i benefici dell'aspirina dimostrati dallo studio scomparirebbero.

Ma allora come si fa a risolvere il problema dei persi al follow-up e dei quali non si conosce quindi il destino? Viene comunemente accettato che perdite minori, dell'ordine del 2-3% dei pazienti, non vadano ad inficiare i risultati e non sia pertanto necessario ricorrere a particolari artifici. Per perdite superiori, in genere fino al 10%, si può ovviare ricorrendo alla cosidetta "sensitivity analysis".

Si tratta di un artificio che cerco di spiegare subito. Supponiamo di avere uno studio che prevedeva come end-point principale il numero di infarti e che abbia avuto una perdita al follow-up del 10% dei pazienti arruolati.

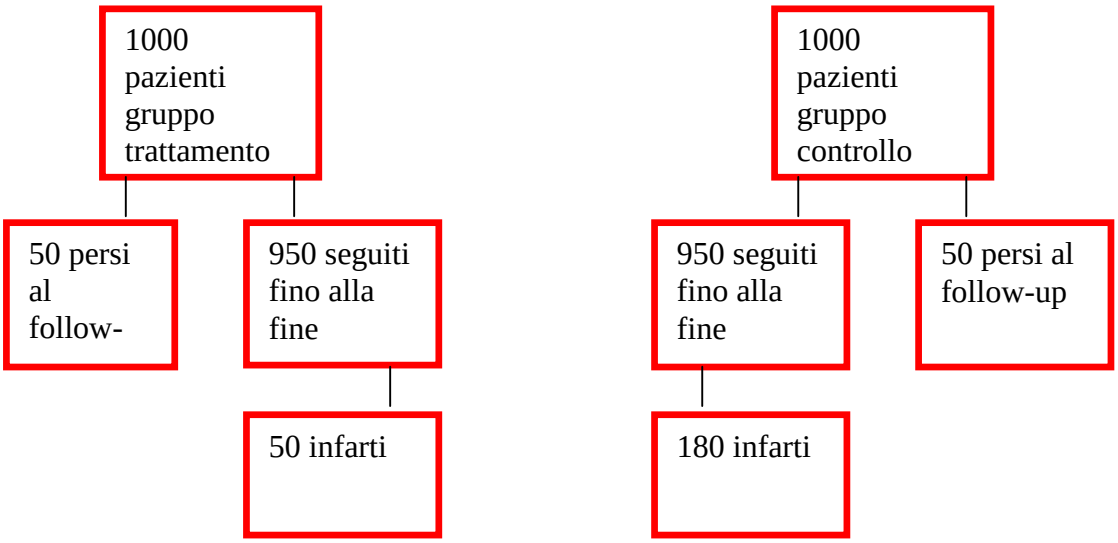
Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Si può “fingere”:

- che tutti i pazienti persi appartenenti al gruppo trattamento abbiano avuto un infarto e quelli del gruppo controllo no
- che tutti abbiano avuto un infarto, sia quelli del gruppo trattamento che quelli del gruppo controllo
- che nessuno di questi pazienti abbia avuto l'infarto
- che l’infarto l’abbiano avuto solo quelli del gruppo controllo e non quelli del gruppo trattamento.

Le cose ovviamente cambiano a seconda dello scenario ipotizzato: se per esempio supponiamo che tutti i pazienti persi del gruppo trattamento siano andati incontro ad un infarto e quelli del gruppo controllo no e, nonostante questo artificio estremo, il trattamento si dimostra ancora efficace possiamo essere tranquilli sulla bontà dei risultati trovati.

Per capire meglio la questione facciamo questo esempio:



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Come si vede la perdita al follow-up è del 5% in entrambi i gruppi. Il rischio assoluto di infarto nel gruppo trattamento è del 5% (attenzione perché si deve fare un'analisi intention to treat e quindi bisogna considerare la totalità dei pazienti assegnati a quel gruppo), nel gruppo controllo è del 18%, il Rischio Relativo è di 0,28 con IC95% di 0,21-0,33. Si conclude quindi che il trattamento è efficace nel ridurre il rischio di infarto in maniera significativa. Rimane però il problema dei pazienti persi al follow-up, quindi queste conclusioni sono affidabili? Ipotizziamo allora che tutti i 50 pazienti del gruppo trattamento abbiamo avuto un infarto (scenario peggiore ed in verità estremamente improbabile) e che nessuno dei 50 pazienti persi del gruppo controllo abbia avuto un infarto. Si ottiene:

- 1000 pazienti del gruppo trattamento: 50 + 50 infarti = 100
- 1000 pazienti del gruppo controllo: 180 infarti

Il rischio assoluto con il trattamento diventa del 10%, quello del gruppo controllo rimane del 18%. Si ha un RR = 0,56 con un IC95% = 0,44 - 0,70.

Come si vede, nonostante questa ipotesi estrema la riduzione del rischio ottenuta dal farmaco rimane significativa dal punto di vista statistico. Si può quindi dire che, nonostante la perdita di pazienti del 10% al follow-up (5% nel gruppo controllo e 5% nel gruppo trattamento), i risultati dello studio sono robusti perché se la significatività statistica rimane nello scenario più sfavorevole al trattamento a maggior ragione rimarrebbe nelle altre ipotesi.

Questo modo di procedere può essere accettato, come abbiamo visto, per perdite fino al 10% della casistica arruolata (alcuni autori portano questa percentuale al 15-20%). Per perdite superiori il trial, secondo stretti criteri metodologici, non fornisce più risultati attendibili. Infatti quando queste perdite sono notevoli esse possono essere molto squilibrate tra un braccio e l'altro: per esempio potrebbe essere che nel gruppo placebo i pazienti persi sono più a rischio o più malati di quelli

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

persi nel gruppo trattamento (o viceversa). Il fenomeno finisce comunque per contrastare i benefici della randomizzazione che, come sappiamo, ha lo scopo di distribuire in maniera eguale fattori di rischio noti e non noti tra i due bracci dello studio. Un numero troppo elevato di drop-outs può quindi creare un vero e proprio “bias”. Anche l'escamotage della sensitivity analysis va sempre considerato con prudenza.

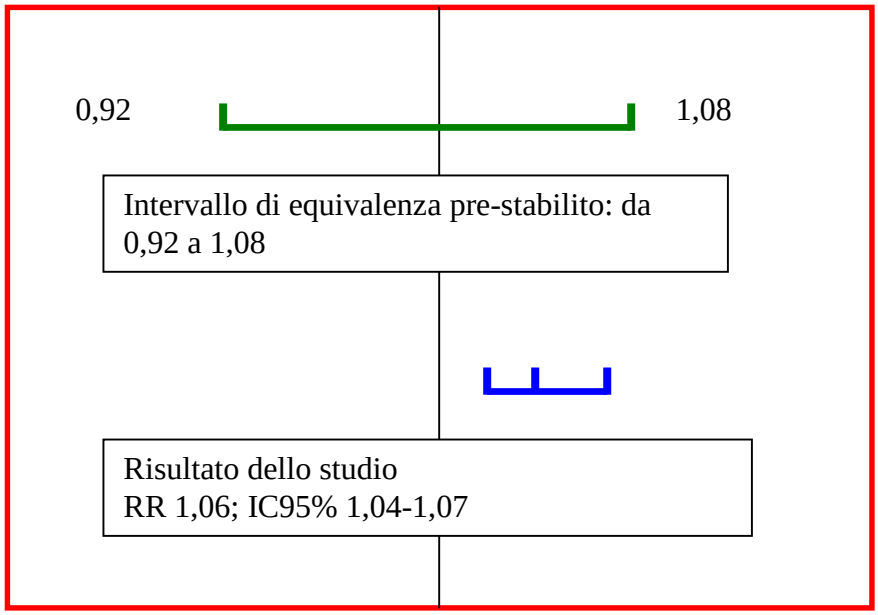
Studi di equivalenza

Gli studi possono essere disegnati con lo scopo di dimostrare che un trattamento è superiore ad un altro (studi di superiorità). Tuttavia i ricercatori spesso decidono di accontentarsi di dimostrare che un trattamento è equivalente ad un altro (studi di equivalenza). Ci sono anche gli studi di "non inferiorità" che non sono altro che un sottotipo particolare degli studi di equivalenza. Mi direte: dove sta la differenza? Negli studi di equivalenza i ricercatori, nel protocollo, stabiliscono chiaramente di considerare equivalenti due trattamenti se la differenza trovata varierà entro un certo intervallo detto **"intervallo di equivalenza**", stabilito a priori e in modo arbitrario. Spiegare come funziona la faccenda non è semplice, perché la statistica e la tecnica la fanno da padroni, cercherò di farlo a costo di commettere delle inesattezze.

In uno studio di equivalenza si parte da questo assunto: voglio dimostrare che un determinato farmaco è "equivalente" ad uno di confronto e quindi stabilisco un "intervallo di equivalenza" per esempio + 8% e - 8%. In altre parole: ho un farmaco antipertensivo e voglio vedere se è equivalente al farmaco di riferimento nel ridurre l'ictus e se trovo che l' intervallo di confidenza al 95% del rischio relativo (o di qualsiasi altra misura di efficacia considerata) cade entro l'intervallo di equivalenza che ho stabilito precedentemente dirò che i due trattamenti sono equivalenti. Così se

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

troverò un IC95% dell'RR compreso tra - 6 e + 7 dirò che di due trattamenti sono equivalenti, perché l'IC95% cade nell'intervallo di equivalenza che a priori ho stabilito essere compreso tra + 8 e - 8. E' interessante notare che se trovo un IC95% dell'RR compreso tra + 4 e + 7 posso sempre dire che i due trattamenti sono equivalenti perché l'IC95% cade nell'intervallo di equivalenza prestabilito, però se si trattasse di uno studio normale si dovrebbe concludere che il farmaco testato rispetto al controllo provoca un aumento dell'ictus e che il risultato è statisticamente significativo perché l'IC95% è sempre di segno positivo. Ora, questa cosa sarà anche accettata dagli studiosi, avrà tutti crismi e le benedizioni della statistica, e gli esperti mi daranno dell'ignorante, ma mi convince poco.



Come si deduce dalla tabella lo studio ha portato ad un RR di 1,06 con IC95% che va da 1,04 a 1,07; formalmente questo intervallo cade entro l'intervallo di equivalenza pre-stabilito quindi si può concludere che i due trattamenti sono equivalenti.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 9 – Pazienti persi al follow-up, studi di equivalenza e SAEs

Di solito quando si stabiliscono i limiti dell'intervallo di equivalenza si scelgono delle differenze piccole che si ritiene clinicamente non rilevanti: sarebbe difficile far accettare che due trattamenti sono equivalenti con una differenza di esiti clinici in più o in meno del 50%! Ovviamente negli studi di “non inferiorità” viene stabilito a priori un intervallo di non inferiorità: per esempio si definisce non inferiore ad un trattamento standard un nuovo farmaco se gli eventi che provocherà in più saranno compresi entro un determinato valore.

La questione diventa abbastanza complicata, pane per i denti di chi si dedica ai critical appraisal, tuttavia questo tipo di studi solleva numerose domande sia di tipo metodologico che etico. Bisogna che gli autori giustifichino il fatto di aver scelto uno studio di equivalenza con motivazioni cliniche ineccepibili, l'intervallo di equivalenza deve essere predefinito (in caso contrario lo studio non può essere considerato di equivalenza ma di superiorità), la numerosità del campione deve essere tarata su questa differenza, ecc.

Come fa notare un editoriale pubblicato sul Bollettino di Informazione sui Farmaci [BIF 2006; n. 3, pagg. 97 e seguenti] gli studi di equivalenza e non inferiorità presentano molti aspetti ambigui, vanno attentamente valutati per vedere se non si tratta in realtà di studi mascherati, cioè studi che all'origine erano di superiorità ma che, visti i risultati negativi, sono stati trasformati a posteriori in studi di equivalenza. Come dice Aristotele nella Retorica, si è oscuri se non si dice fin dall'inizio ciò che si vuole.

Un aspetto cruciale è poi la **corretta informazione dei pazienti che partecipano al trial**, i quali devono chiaramente sapere che stanno provando un farmaco che noi abbiamo deciso di dichiarare equivalente a quello di paragone anche se provoca un aumento degli eventi di una certa entità che abbiamo stabilito, in modo comunque arbitrario, essere clinicamente poco importante.

Mi domanderete: perché tutte queste complicazioni, non sarebbe più semplice fare solo studi di superiorità? Qui casca l'asino. Intanto la dimostrazione di "equivalenza" viene accettata dalle

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

autorità regolatorie per registrare un nuovo farmaco. Inoltre ormai la ricerca farmaceutica ha raggiunto un tal grado di perfezione che spesso è difficilissimo dimostrare la superiorità di un nuovo farmaco rispetto a quello di confronto in termini di riduzione degli eventi clinici. A dirla francamente però viene il sospetto che degli studi di equivalenza e di non inferiorità si tenda ad abusare, quasi si avesse il timore di confrontarsi "a viso aperto" con i concorrenti.

In conclusione consiglierei di considerare sempre con prudenza i risultati di uno studio di equivalenza o non inferiorità perché è stato visto che in molti casi la loro qualità metodologica è discutibile.

I SAEs

SAEs è l'abbreviazione di Serious Adverse Events: si tratta di tutti gli eventi avversi che si verificano in uno studio e che hanno portato a morte il paziente, che hanno costretto al ricovero o che hanno prodotto una disabilità grave. Conoscere i SAEs che si sono verificati durante uno studio è fondamentale.

Infatti solo il bilancio tra i benefici ottenuti dal trattamento e i suoi possibili rischi permette di capire se il farmaco è utile o meno. Purtroppo negli studi randomizzati e controllati, per quanto ben disegnati, con ampia casistica, e di lunga durata, spesso questi aspetti non vengono considerati (per la verità oggi meno che in passato) perché ai ricercatori interessa di più analizzare specifici end-point legati al trattamento oggetto dello studio.

Un' analisi interessante sui SAEs è stata compiuta da autori canadesi sugli studi che riguardano le statine. In questa analisi sono stati accorpati, separatamente, gli studi di prevenzione primaria e quelli di prevenzione secondaria. Ne risulta che mentre negli studi di prevenzione secondaria il beneficio delle statine è evidente sia sulla morbidità e sulla mortalità cardiovascolari che sulla

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

mortalità totale, le cose per la prevenzione primaria sono meno chiare. Infatti in questo caso le statine si mostrano utili a ridurre gli eventi cardiovascolari ma non la mortalità totale, inoltre negli studi in cui è stato possibile calcolare i SAEs (purtroppo non in tutti gli studi questo si è potuto fare) risulta che essi non sono ridotti dalla terapia. I canadesi si domandano quindi: perché le statine riducono gli eventi cardiovascolari ma non i SAEs (che comprendono **tutti** gli eventi gravi e i decessi verificatisi nello studio, quindi anche quelli cardiovascolari)? Forse perché ad una riduzione degli eventi cardiaci fa da controcampo un aumento di altre patologie? La risposta non è semplice ma secondo gli autori il beneficio globale delle statine in prevenzione primaria sulla salute dei pazienti sarebbe troppo enfatizzato.

Un altro esempio interessante di analisi del rapporto benefici/rischi di un trattamento ci viene dallo studio VIGOR in cui il rofecoxib venne paragonato al naproxene: ogni 1000 pazienti trattati si ebbero 4 ulcere complicate con il rofecoxib e 9 con il naproxene, ma nello stesso tempo si verificarono 4 infarti del miocardio con rofecoxib e 1 con il FANS non selettivo. Ciò vuol dire che trattando 1000 pazienti con il rofecoxib evito 5 ulcere complicate a scapito di 3 infarti del miocardio in più. Qual è l'importanza clinica del beneficio ottenuto sul versante gastrico?

Un esempio addirittura clamoroso è quello dello studio WHI sulla terapia ormonale sostitutiva in post-menopausa, ormai notissimo: a fronte di una riduzione statisticamente significativa del rischio fratturativo del 24% e del rischio di cancro del colon del 37% la TOS produceva, in questo storico RCT, un aumento del rischio di tumori mammari, infarti miocardici, ictus ed embolie polmonari, tanto che i ricercatori ritennero di interromperlo anticipatamente perché i benefici erano inferiori ai rischi: ogni anno per 10.000 donne trattate si avevano 19 eventi avversi in più con la TOS. Potrebbe sembrare un aumento del rischio molto piccolo, ma tenendo conto che la terapia veniva proposta a donne sane, ogni cautela è d'obbligo. I risultati del WHI sono stati così importanti che numerose società scientifiche hanno emanato delle linee guida in cui viene sconsigliato l'uso della

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

terapia ormonale per la prevenzione di patologie croniche come la cardiopatia ischemica e l'osteoporosi.

In conclusione sarebbe auspicabile che gli RCT, più che limitarsi a misurare singoli end-point, registrassero e riportassero tutti i SAEs avvenuti durante il follow-up. Solo in questa maniera, infatti, si può avere un quadro completo e tracciare un bilancio dei reali benefici (e degli eventuali danni) che un trattamento può provocare.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 10
Alcune considerazioni sugli RCT

Abbiamo visto nei capitoli precedenti vari aspetti degli RCT che richiamo brevemente:

1) La bontà della randomizzazione: se adeguata garantisce la confrontabilità dei due bracci; se di scarsa qualità può portare a sovrastimare l'efficacia del trattamento. Oserei dire che gli uomini sono nati con uguali diritti, ma negli studi solo la randomizzazione crea uguali i due gruppi.
2) La qualità del mascheramento: la mancanza di cecità o una cecità non adeguata possono portare a sovrastimare l'efficacia del trattamento fino al 10-15%; purtroppo non tutti gli studi riportano la qualità del mascheramento
3) Gli outcomes o end-point, che possono essere hard o surrogati, primari o secondari, singoli o composti
4) La completezza e la qualità del follow-up con le problematiche legate ai drop-out (pazienti persi al follow-up) e alla analisi effettuata secondo l'intenzione a trattare

Trattando gli end-point avevo accennato al problema della misclassificazione, citando il caso del cancro prostatico. Lo riprendo per approfondire il discorso. Se ricordate in quell' occasione richiamavo uno studio sulla efficacia della prostatectomia nel trattamento del tumore localizzato della prostata. In quello studio si trovò che la mortalità dovuta alla neoplasia era ridotta dall'intervento chirurgico rispetto alla vigile attesa, tuttavia la mortalità totale non differiva tra i due gruppi. Questo potrebbe dipendere dal fatto che il campione arruolato nello studio era troppo piccolo (circa 700 pazienti) per avere una potenza statistica tale da evidenziare variazioni nella mortalità globale. Ma potrebbe anche dipendere dal fatto che alcuni decessi verificatisi nel gruppo sottoposto a intervento chirurgico sono stati attribuiti ad altre cause e non al tumore prostatico. Potrebbero essersi verificati dei decessi da complicanze post-operatorie tardive (per dire: embolie

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

polmonari) che non sono stati ascritti al cancro prostatico. Questo esempio viene citato unicamente a scopo didattico perché qualche anno dopo venne pubblicato un aggiornamento dello studio con un follow-up più lungo che dimostrava che protraendo l'osservazione si evidenziavano i benefici dell'intervento pure sulla mortalità totale (anche se questi erano evidenti solo per chi aveva meno di 65 anni).

Questo possibile errore nelle attribuzioni delle cause di morte è stato ipotizzato soprattutto negli studi sugli screening oncologici. Uno studio che ha esaminato in dettaglio otto RCT sullo screening oncologico sottolinea che mentre lo screening riduce la mortalità specifica, non sembra influenzare quella globale. Ciò potrebbe essere dovuto a due tipi di bias: il primo (detto in termini complicati sticky-diagnosis bias) attribuisce erroneamente decessi da altre cause al tumore che deve essere oggetto dello studio (questo errore si verifica per solito nel gruppo di controllo non screenato), il secondo (detto slippery-linkage bias) attribuisce decessi dovuti a trattamenti aggressivi non alla neoplasia ma ad altre cause (questo errore per solito si verifica nel gruppo screenato). Gli autori di questa revisione giungono alla conclusione che l'efficacia degli screening non può essere affermata se non in presenza di una riduzione della mortalità totale (e non solo di quella specifica), perché la mortalità totale tout court è evidentemente un end-point non soggetto a valutazioni potenzialmente errate.

Ha destato meraviglia (e anche aspre polemiche) una meta-analisi (vedremo in seguito cosa si intende con questo termine) sullo screening mammografico, pubblicata dal Lancet nel 2000, ad opera di due autori danesi. Esaminando tutti gli RCT disponibili i due studiosi arrivano a concludere che lo screening mammografico, sulla base dei dati disponibili, non sembra portare benefici in quanto non riduce la mortalità totale, mentre il beneficio sulla mortalità da cancro mammario è comunque piccolo. In ogni caso, secondo lo studio, molti degli RCT esaminati sono contestabili dal punto di vista metodologico e sono quelli che mostrano un beneficio dello Capitolo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

screening. Al contrario se si esaminano i due RCT più affidabili l'utilità dello screening scompare. Queste conclusioni portarono ad un dibattito infuocato nelle maggiori riviste internazionali, con prese di posizione a favore e contro, con meta-analisi che contraddicevano la meta-analisi danese. Un altro esempio riguarda la terapia con aspirina, un caposaldo dei trattamenti cardiovascolari che nessuno si metterebbe in mente di contestare. Ebbene, vi sono voci fuori dal coro, molto critiche, che fanno notare come la terapia con aspirina nei maggiori trials riduca gli eventi cardiovascolari (fatali e non) ma non la mortalità globale. Questo fatto suggerisce che l'aspirina possa trasformare, piuttosto che ridurre, gli eventi cardiovascolari. L'asa può produrre epigastralgie, d'altra parte circa il 25% degli infarti non fatali è asintomatico: non è difficile pensare che l'asa, essendo tra l'altro dotato di effetto analgesico, aumenti la proporzione degli infarti silenti dal 25% al 30% e questo potrebbe spiegare l'apparente beneficio sugli infarti non fatali riscontrato dagli studi. L'asa non fa altro che trasformare gli infarti sintomatici in forme silenti. Ma non è finita. In tutti i trias in cui l'asa è stato usato long-term e che abbiano riportato la morte improvvisa come end-point si verifica che questo outcome è aumentato nel gruppo trattato con asa.

Questi due esempi illustrano bene quanto complessa e difficile possa essere l'analisi dei dati della letteratura. La cosa va naturalmente al di là degli scopi che ci siamo prefissi ma è opportuno almeno esserne a conoscenza. Il medico pratico cosa dovrebbe fare? In realtà sarà bene abituarsi all'incertezza, riconoscendo che lo studio e la meta-analisi perfetta non sono ancora stati portati a termine e che sempre più spesso saremo di fronte ad interpretazioni diverse che i vari esperti daranno degli stessi studi.

Un altro aspetto che mi preme qui esaminare è quello della **trasferibilità degli RCT nella pratica clinica quotidiana.** Un primo punto riguarda i criteri di arruolamento, cioè le caratteristiche dei pazienti inclusi nello studio: spesso gli RCT escludono gli anziani con pluripatologie, mentre le donne tendono ad essere poco rappresentate. Così l'esclusione di pazienti più gravi o con Capitolo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

determinate patologie può portare a selezionare una popolazione che risponde meglio al trattamento. D'altra parte l'inclusione di determinati pazienti con comorbidità può influire sull'end-point misurato. Per esempio se in un trial si includono pazienti con pregressa cardiopatia ischemica questo potrebbe influire sulla efficacia di un trattamento. La selezione di un certo tipo di pazienti non deve essere vista come un risvolto negativo: anzi questo aspetto determina una miglior individuazione dei soggetti a cui i risultati possono essere trasferiti. Il rovescio della medaglia comporta che un determinato studio può andar bene solo per una ristretta fascia di soggetti. Negli studi sull'impiego dei beta-bloccanti nello scompenso cardiaco erano stati esclusi, in genere, i pazienti con più di 65 anni, quelli con patologie importanti associate e quelli con valori di creatininemia superiori a 2,5-3 mg/dL; erano poco rappresentati anche i soggetti con fibrillazione atriale. Infine in quasi tutti gli studi erano arruolati pazienti con frazione di eiezione superiore al 40% (cosiddetto scompenso da disfunzione sistolica) mentre erano esclusi quelli con frazione di eiezione conservata (scompenso da disfunzione diastolica) che invece rappresentano circa il 40% e più di quelli visti nella pratica. Tutto questo può limitare molto la trasferibilità dei risultati degli studi sui beta-bloccanti nello scompenso cardiaco. Nello studio VALIANT il valsartan si è dimostrato non inferiore al captopril , mentre la loro associazione non era più efficace dei singoli farmaci. Si trattava però di scompenso post-infartuale con disfunzione sistolica, quindi questi risultati non sono necessariamente trasferibili a tutti i pazienti con scompenso cardiaco.

Un pratica abbastanza spesso usata dai ricercatori è quella di selezionare prima i pazienti da sottoporre poi al trial. Nello studio TNT una statina ad alto dosaggio ha ridotto gli eventi cardiovascolari rispetto a dosi più basse della stessa statina (senza vantaggi sulla mortalità totale) in pazienti con coronaropatia normocolesterolemici . Nella prima fase dello studio vennero arruolati oltre 15.000 pazienti affetti da malattia coronarica stabile, con valori di colesterolo LDL compresi tra 130 mg/dL e 250 mg/dL, trattati in aperto per otto settimane con 10 mg/die della statina in

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

esame. Successivamente , tra questi, furono scelti quelli (10.000) in cui il colesterolo LDL, dopo le 8 settimane di trattamento iniziale, risultava inferiore a 130 mg/dL, destinati a partecipare allo studio vero e proprio che prevedeva la somministrazione della statina alle dosi di 10 mg/die oppure 80 mg/die. Questo potrebbe aver portato alla selezione di pazienti che rispondono meglio al farmaco e aver scartato quelli più resistenti. Si tratta di una procedura prevista dalla ricerca clinica ma che va, in qualche modo, a interferire con la scelta casuale dei pazienti, e ciò potrebbe costituire una sorta di bias di selezione. Bisogna dire che questa pratica è molto usata (forse troppo): se ne trova un esempio anche in un altro studio molto citato come l'HPS, in cui da più di 32.000 pazienti arruolati inizialmente si passò a circa 20.500.

Nel valutare la trasferibilità dei trials bisogna poi considerare che, in genere, la ***compliance*** ottenuta negli studi sperimentali è maggiore rispetto a quella ottenuta nella medicina pratica perché pazienti che accettano di partecipare agli RCT sono pazienti più motivati.

Questo spiega come mai risultati che si possono ottenere nel mondo reale sono spesso diversi da quelli ottenuti nel contesto sterilizzato degli RCT. Il ***contesto*** in cui lo studio è stato realizzato ha un'importanza fondamentale. Così se un RCT viene realizzato in centri di ricerca universitaria è probabile che i pazienti siano sottoposti a follow-up ed accertamenti più intensivi rispetto a quanto avviene nella pratica. Al contrario studi portati a termine sul territorio da medici di famiglia rispecchiano più fedelmente la realtà clinica di ogni giorno. Due ampi studi canadesi sullo screening mammografico, in cui non c'è stata una riduzione della mortalità nel gruppo screenato, sono stati accusati di aver coinvolto centri radiologici con qualità inferiore rispetto ad altri RCT che avevano invece dimostrato l'efficacia dello screening mammografico nel ridurre la mortalità specifica. Ma questo, eventualmente, dimostra che non è detto che quanto si ottiene negli RCT sia poi tout court raggiungibile nella pratica di tutti i giorni, quindi non si tratta di una debolezza ma di una fotografia più aderente alla realtà. Il contesto in cui viene effettuato lo studio è importante

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

anche nei trials chirurgici, in cui possono essere riportati risultati migliori se sono coinvolti operatori ad elevata professionalità. Per esempio uno studio ha dimostrato che le complicanze post-prostatectomia sono meno frequenti se ad operare è un chirurgo con esperienza che fa molti interventi, anche se opera in un piccolo ospedale, piuttosto che un chirurgo con meno esperienza che opera in un ospedale più importante.

Un ulteriore esempio ci viene dai risultati emersi dall'esame del registro GRACE, secondo cui non sempre, nelle sindromi coronariche acute, l'approccio aggressivo (angioplastica o by-pass) è superiore alla terapia medica, in contrasto con i risultati di precedenti RCT. Il registro GRACE è un buon esempio che quanto si riesce ad ottenere negli studi randomizzati e controllati non è poi completamente trasferibile nel mondo reale.

Un altro elemento che condiziona la trasferibilità di un trial è la ***caratteristica degli interventi***. Per esempio nel gruppo di controllo può essere effettuato un trattamento placebo oppure la cosiddetta “usual care” o ancora un trattamento di riferimento. Numerosi studi hanno dimostrato l'efficacia dei triptani nella crisi emicranica, ma spesso il gruppo di controllo assumeva placebo. Negli studi in cui i triptani sono stati confrontati con farmaci alternativi (paracetamolo o FANS) la differenza tra i due trattamenti non è risultata così eclatante come viene spesso immaginato. In uno studio che confrontava due beta-bloccanti nello scompenso cardiaco (carvedilolo e metoprololo) è risultato più efficace il beta-bloccante di più recente immissione in commercio ma l'altro, che funzionava da controllo, venne usato ad un dosaggio inferiore di quello che era stato validato in studi precedenti. Anche il tipo di intervento effettuato nel gruppo in trattamento attivo va attentamente esaminato. Per esempio nello studio HOPE è stato dimostrato un beneficio del ramipril in soggetti a rischio cardiovascolare, anche non ipertesi. Tuttavia bisogna considerare che nello studio venivano usati 10 mg/die di rampril, mentre è esperienza comune osservare, nella pratica di tutti i giorni, la

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

prescrizione del farmaco a dosaggi inferiori. A dosaggi più bassi il ramipril ottiene gli stessi risultati?

Per valutare la trasferibilità e l'importanza clinica di un trial bisogna considerarne, infine, i ***risultati e la loro interpretazione***. Per esempio, una volta stabilito che un risultato è significativo statisticamente il nostro lavoro non è terminato. Dobbiamo stabilire se esso ha una qualche rilevanza clinica. I due concetti, significatività statistica e clinica, non necessariamente coincidono e se stabilire la prima è semplice, valutare la seconda lo è meno.

Un esempio ci farà comprendere meglio: supponiamo un RCT in cui vengono arruolati 40.000 pazienti, 20.000 trattati con il farmaco “X” e 20.000 con il farmaco “Y”. L’end-point dello studio sia la comparsa di scompenso cardiaco che rende necessario il ricovero. Dopo un follow-up di 5 anni si hanno 15 ricoveri per scompenso cardiaco nel gruppo “X” e 30 ricoveri nel gruppo “Y”. Si può correttamente dire che il farmaco “X” riduce il rischio di scompenso del 50% (RR = 0,50) e che il dato è significativo dal punto di vista statistico (IC95% = 0,27-0,93). Ma qual è l’impatto clinico? Se si calcola l’NNT notiamo che è necessario trattare 1333 pazienti per ben 5 anni per evitare un ricovero da scompenso cardiaco. In altre parole trattiamo inutilmente 1332 pazienti, che non riceveranno nessun beneficio dalla terapia mentre saranno esposti alla possibile comparsa di effetti collaterali potenzialmente gravi. L'esempio che ho fatto ovviamente è portato all'eccesso ma mi serve per far vedere che significatività statistica non vuol dire automaticamente studio clinico importante. A questo riguardo l’NNT fornisce informazioni più utili del Rischio Relativo e della P. Un esempio reale di quanto detto si può trovare negli studi sulle statine. Esaminiamo due di questi studi e prenderemo confidenza con il metodo che si può usare per valutare l’efficacia di un intervento farmacologico.

Nello studio denominato 4S (studio di prevenzione secondaria) dopo 5 anni di trattamento con una statina in soggetti con cardiopatia ischemica si ebbero 8,2 decessi ogni 100 pazienti nel gruppo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

statina e 11,5 nel gruppo placebo. Si ottenne una riduzione del rischio assoluto (ARR) del 3,3%, una riduzione del rischio relativo (RRR) del 29% e un NNT di 30.

Nello studio WOSCOPS (studio di prevenzione primaria), dopo 4,9 anni di trattamento si ebbero 3,2 decessi ogni 100 pazienti nel gruppo statina e 4,1 nel gruppo placebo. Si ottenne una ARR dello 0,9%, un RRR del 22% e un NNT di 111 (tra l'altro statisticamente non significativo, ma facciamo finta che lo sia).

L'esame di questi dati permette di affermare che, pur avendosi un beneficio in entrambi gli studi (riduzione dei decessi del 20-30%), la terapia è molto più efficace in prevenzione secondaria perché si trattano meno soggetti per evitare un evento: in caso di risorse economiche limitate è ovvio che la precedenza venga data al trattamento dei pazienti a rischio più elevato perché il rapporto costi/benefici è più favorevole.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 11
Sopravvivenza e curve di Kaplan

Si supponga di avere uno studio in cui ci sono 200 pazienti arruolati, 100 si sottopongono ad uno screening oncologico e 100 funzionano da gruppo di controllo. Dopo 5 anni si contano i decessi e si vede che nel gruppo screening ci sono stati 15 decessi e nel gruppo di controllo ci sono stati 15 decessi. Possiamo esprimere i dati dicendo che la mortalità nei due gruppi è stata del 15%, oppure, in maniera speculare, che la sopravvivenza è stata dell'85%. In altre parole in tutti e due i gruppi alla fine dello studio ci sono 85 soggetti ancora vivi mentre 15 sono deceduti. In questo caso, sia che si usi la mortalità sia che si usi la sopravvivenza, ci si riferisce sempre al numero totale dei soggetti arruolati nei due bracci.

Supponiamo ora di avere uno studio sullo screening del cancro polmonare e che nei 100 soggetti del gruppo intervento diagnostichiamo 30 tumori mentre nei 100 del gruppo di controllo ne diagnostichiamo 20. In entrambi i gruppi però alla fine si abbiano 15 decessi per cancro polmonare. A questo punto invece di riferirci al numero totale dei soggetti arruolati riferiamoci solo al numero dei malati di cancro polmonare, che sono, come s'è detto, 30 nel gruppo screening e 20 nel gruppo controllo. Possiamo dire che la sopravvivenza è stata di 15 su 30 in un gruppo (= 50%) e di 15 su 20 (= 25%) nell'altro. Oppure, usando la mortalità, che la mortalità è stata di 15 su 30 (= 50%) nel gruppo screening e di 15 su 20 (=75%) nel gruppo controllo. Quindi, se invece di riferirci al totale dei pazienti arruolati per ogni gruppo, ci si riferisce ai casi trovati di cancro polmonare apparentemente lo screening aumenta la sopravvivenza a 5 anni nei malati dal 25% al 50% (oppure, il che è lo stesso, riduce la mortalità nei malati dal 75% al 50%). Si noti però che se noi andiamo a contare quanti sono stati i morti per cancro polmonare in entrambi i gruppi la mortalità è sempre del 15% e non viene ridotta dallo screening (o se si vuole la sopravvivenza è sempre dll'85% e non viene aumentata dallo screening). Come è possibile una cosa del genere? Dipende da un fenomeno ben noto degli screening che va sotto il nome di **“sovradiagnosi**”: nel gruppo

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

screenato sono stati diagnosticati 10 tumori in più perché siamo andati a cercarli ma è probabile che si tratti di forme indolenti di tumore, che non sarebbero mai diventate clinicamente evidenti. Per questo motivo negli studi di screening l'end-point che si dovrebbe misurare è la mortalità (o il suo speculare, la sopravvivenza) riferiti al totale dei pazienti studiati e non solo a quelli malati. Purtroppo bisogna stare bene attenti quando si legge uno studio per capire in quale modo sono state determinate mortalità e sopravvivenza. Per esempio nello studio ELCAP hanno seguito oltre 31.000 pazienti screenati per cancro del polmone con TAC spirale. Sono stati trovati 484 cancri del polmone, di questi 412 avevano un tumore in stadio I: la loro sopravvivenza media a 10 anni (peraltro stimata) è stata dell'88%. E' stata poi confrontata questa sopravvivenza con quella media del cancro polmonare nella realtà (dove non si attua lo screening) che, a 10 anni, è del 10%. Gli autori hanno quindi concluso che la TAC permette una diagnosi precoce e un aumento della sopravvivenza fino a circa il 90% per le forme in stadio I. Sono affidabili queste conclusioni? Molto poco sia perché lo studio era di tipo osservazionale e mancava un gruppo di controllo, sia perché la sopravvivenza è stata calcolata sui malati, esponendosi al bias della sovradiagnosi. Quando si legge di aumento della sopravvivenza o di riduzione della mortalità, soprattutto in uno studio di screening, bisogna stare molto attenti a cosa queste misure si riferiscono e come sono state calcolate. Un altro esempio mostrerà come possa essere usato il calcolo della sopravvivenza, espresso in termini di tempo. Si tratta dello studio della Mayo Clinic sullo screening del cancro polmonare mediante radiografia del torace ed esame dell'escreato. Lo studio era di tipo randomizzato e confrontava due modalità di screening: ogni 4 mesi e ogni anno. Alla fine dello studio si vide che i pazienti a cui era stato diagnosticato un cancro polmonare nel gruppo screening ogni 4 mesi avevano una sopravvivenza media di 16 anni mentre quelli con cancro polmonare del gruppo screening ogni anno avevano una sopravvivenza media di 5 anni. Tuttavia la mortalità per cancro polmonare alla fine dello studio era uguale nei due gruppi. Come si spiega

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

questo apparente paradosso? Si spiega anche qui col fatto che la mortalità è stata calcolata sul totale dei soggetti arruolati per ogni braccio mentre la sopravvivenza, espressa in termini di tempo, è stata calcolata sui soggetti malati. Vi può essere quindi la distorsione legata alla sovradiagnosi.

Negli studi di screening potrebbe entrare in gioco un altro bias, quello della **anticipazione diagnostica**. Prendiamo due gruppi di 10 pazienti ciascuno sottoposti a screening oppure no per cancro del polmone. Supponiamo che in ognuno dei due gruppi ci sia un paziente di 45 anni che ha un cancro polmonare iniziato da 5 anni, quindi all'età di 40 anni, non visibile alle odierne tecniche di indagine. Per ogni gruppo ci sono 10 pazienti di cui 9 sani e uno con cancro polmonare "invisibile", iniziato da 5 anni. Siccome lo screening scopre prima il tumore, nel paziente appartenente al gruppo screenato la neoplasia viene scoperta dopo un anno che è iniziato lo studio, a 46 anni. Il paziente viene trattato e poi muore a 52 anni, 6 anni dopo la diagnosi. Invece nel paziente non sottoposto a screening il tumore viene scoperto più tardivamente, a 50 anni, mentre il decesso avviene sempre a 52 anni. La sopravvivenza dopo la diagnosi nel primo caso è di 6 anni e nel secondo di 2. Lo screening aumenta la sopravvivenza di 4 anni? In realtà la mortalità per cancro polmonare, alla fine dello studio, è uguale nei due gruppi e pari al 10% (muore di cancro del polmone 1 sui 10 arruolati per ogni gruppo). Quindi lo screening ha scoperto prima il tumore, ma questo non si è tradotto in una riduzione della mortalità totale (vedi tabella).

	Nascita	Inizio tumore	Entrata studio	Diagnosi	Decesso
Paziente screenato	1940	1980	1985	1986	1992
Paziente non screenato	1940	1980	1985	1990	1992

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 11 – Sopravvivenza e curve di Kaplan

Mi direte: bravo furbo, tu la sopravvivenza l'hai misurata dalla diagnosi e in questo modo c'è effettivamente una distorsione, ma comincia a fare i conti da quando i due pazienti sono entrati nello studio e vedrai che entrambi vivono 7 anni. Tu fai i conti a seconda di come ti conviene per giustificare la tua tesi. Purtroppo non funziona nemmeno così: anche cominciando a contare a partire dall'inizio dello studio non siamo sicuri che le cose tornino. Per spiegarmi farò un altro esempio schematizzato in tabella in cui due pazienti, con un cancro polmonare "invisibile" iniziato per entrambi a 40 anni, entrano nello studio ad un'età diversa.

	Nascita	Inizio tumore	Entrata studio	Diagnosi	Decesso
Paziente screenato	1940	1980	1985	1986	1992
Paziente non screenato	1937	1977	1985	1987	1989

Come si vede, se anche andiamo a misurare la sopravvivenza a partire dall'entrata nello studio la troveremmo maggiore nel paziente screenato mentre in realtà sappiamo che in entrambi, da quando è iniziato il tumore, è stata di 12 anni.

Non è detto ovviamente che sia così, ma siccome non possiamo sapere in quale momento il tumore ha cominciato a svilupparsi, è giocoforza affidarsi ad un end-point che non abbia queste potenziali distorsioni, vale a dire la **mortalità calcolata rispetto al totale dei soggetti arruolati.**

Le cose però non sono semplici come le abbiamo fin qui illustrate. Pensiamo per esempio ad uno studio in cui sono arruolati 100 pazienti per parte e che, dopo 3 anni, registri 15 decessi per braccio. Ovviamente si può concludere che la mortalità è del 15% in entrambi i casi e che

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

l'intervento studiato (sia esso un farmaco oppure un intervento di screening o altro) non ha portato beneficio rispetto al controllo. Però supponiamo che nel braccio intervento si siano registrati 2 decessi in tutto nel primo anno, altri 5 decessi nel secondo anno, 8 decessi nel terzo. Invece nel braccio controllo si siano registrati: 6 decessi nel primo anno, 6 decessi nel secondo anno e 3 decessi nel terzo anno. Pur essendo vero che il numero di decessi totali alla fine del terzo anno è identico nei due gruppi, il rischio di morte è diverso nei tre diversi periodi: è minore per il primo gruppo nel primo e secondo anno mentre nel terzo anno risulta avvantaggiato il secondo gruppo. Affidarsi, allora, solo alla mortalità alla fine dello studio fa correre il rischio di non tener conto dei benefici dell'intervento nei primi due anni. Probabilmente ognuno di noi preferirebbe affidarsi all'intervento perché, almeno per i primi due anni, avrebbe una mortalità minore rispetto al controllo, anche se poi, alla fine del terzo anno, il vantaggio scompare.

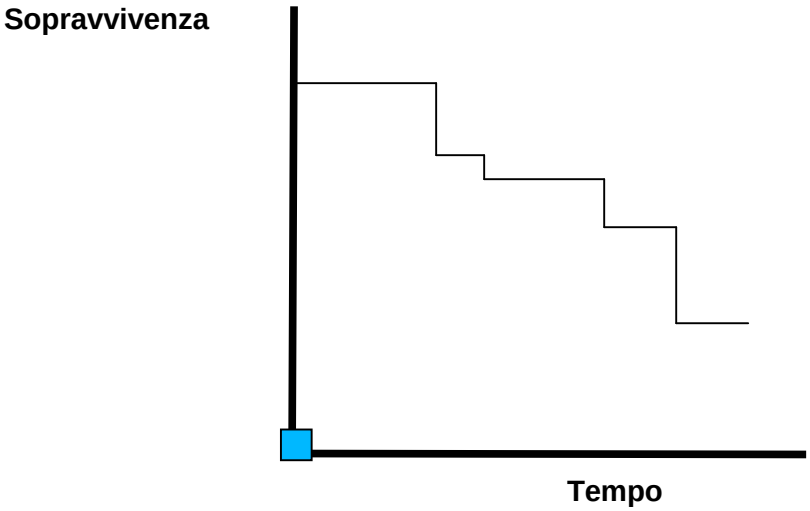
Per ovviare a queste difficoltà vengono costruite delle curve di sopravvivenza (dette **curve di Kaplan-Meier**). Si costruisce un sistema cartesiano in cui l'asse delle ordinate rappresenta la sopravvivenza e l'asse delle ascisse il tempo. La costruzione della curva dipende dal timing nel quale si verificano gli eventi. In altre parole non si stabilisce un intervallo fisso a priori (per esempio contare gli eventi ogni 2 settimane oppure ogni 2 mesi, oppure alla fine di ogni anno, ecc.) ma è la comparsa di ciascun evento che determina la durata degli intervalli.

Si supponga di voler determinare la sopravvivenza fino a 90 giorni di un gruppo di 100 pazienti con ictus: il primo decesso avvenga il giorno 9, il secondo il giorno 12, il terzo il giorno 18, il quarto il giorno 26, il quinto il giorno 44, il sesto il giorno 57, il settimo il giorno 68, l'ottavo il giorno 72, mentre dal giorno 73 al 90 non avviene alcun decesso. Al giorno 9 avviene il primo decesso, quindi dei 100 pazienti iniziali restano 99, la sopravvivenza alla fine del primo periodo (giorni 0-9) sarà di 99/100 = 0,99. Al giorno 12 avviene il secondo decesso, restano 98 pazienti degli iniziali 99 del secondo periodo (giorni 10-12), la sopravvivenza nel periodo sarà di 98/99 = 0,9898. La

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

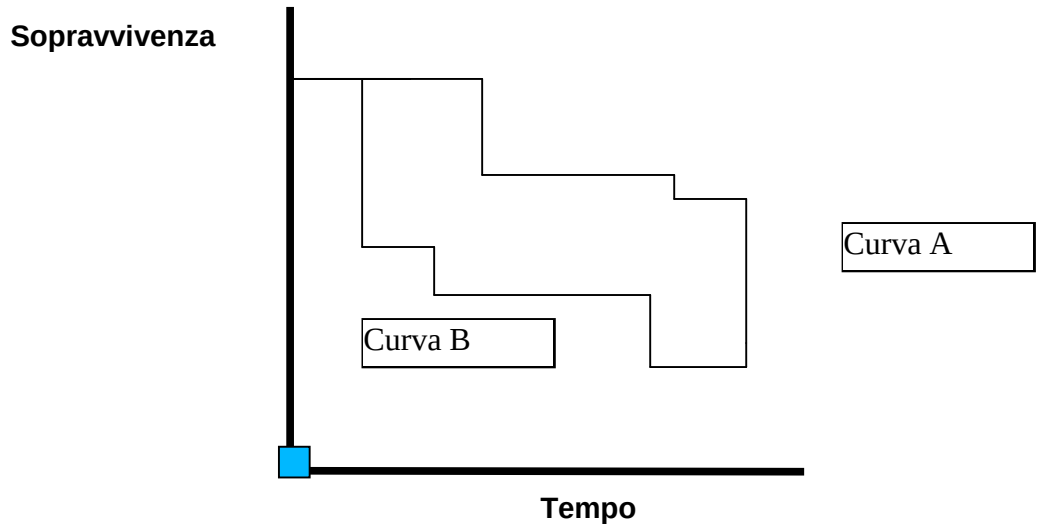
sopravvivenza cumulativa dei due periodi si ottiene moltiplicando quella del primo per quella del secondo (0,99 x 0,9898 = 0,9799). Al giorno 18 avviene il terzo decesso, degli iniziali 98 pazienti restano 97, la sopravvivenza nel periodo (giorni 13-18) sarà di 97/98 = 0,9897. La sopravvivenza cumulativa si trova moltiplicando la sopravvivenza cumulativa precedente per quella trovata nel terzo periodo (0,9799 x 0,9897 = 0,9698). Così si prosegue fino alla fine del follow-up previsto dallo studio.

In questa maniera, riportando la sopravvivenza cumulativa che si trova per ogni periodo nell'asse delle ordinate, si costruisce una curva che ha un **andamento a scalini**. Un esempio è la figura sottostante in cui viene rappresentata una classica curva di sopravvivenza di Kaplan-Meier.



Naturalmente se abbiamo uno studio a due bracci si può costruire una curva di Kaplan-Meier per il braccio intervento e una per il braccio controllo (vedi figura).

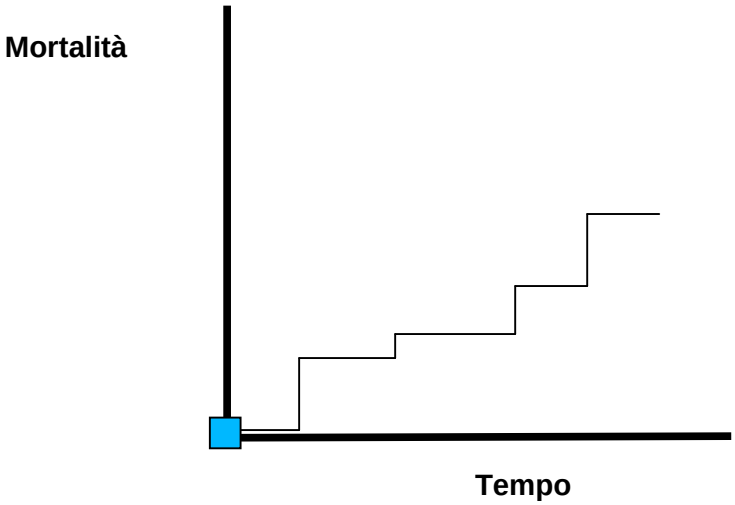
Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati



Come si può facilmente vedere, anche se alla fine dello studio la sopravvivenza è la stessa per entrambe le curve, si nota chiaramente che ai tempi intermedi essa è superiore per la curva A rispetto alla curva B. Se la prima rappresenta la curva ottenibile con uno screening e la seconda quella nel gruppo senza screening, anche se al termine dello studio non c'è differenza, è evidente che nei tempi intermedi lo screening otterrebbe dei benefici, purché il tempo guadagnato sia ovviamente degno di essere vissuto e la qualità di vita decente.

Naturalmente come si costruisce una curva di sopravvivenza si può costruire una curva di mortalità. In questo caso nell'asse delle ascisse va messo ancora una volta il tempo mentre nell'asse delle ordinate va posta la mortalità. La curva risultante avrà un aspetto del tutto speculare rispetto alla curva di sopravvivenza, come mostrato nella figura che segue.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati



Nella stessa maniera si possono costruire curve di Kaplan per ogni altro outcome: infarti, stroke, interventi di rivascolarizzazione coronarica, fratture di femore e così via.

In realtà quando si paragonano due curve di Kaplan le cose non sono mai così lineari come nell'esempio che ho fatto: spesso le curve si intersecano o si sovrappongono tra loro e confrontarle diventa impossibile. Anche nel caso comunque le due curve fossero così distinte come quelle che ho disegnato, bisogna poi stabilire se la differenza trovata nella sopravvivenza è o meno significativa. A tal fine si ricorre al **log-rank test.** Tuttavia questo test permette solo di rifiutare l'ipotesi nulla, vale a dire dice solo se la differenza tra le due curve è significativa, ma non fornisce informazioni circa l'entità di tale differenza o l'intervallo di confidenza. Per questo viene usato il **modello a rischi proporzionali di Cox**. L'argomento è molto complicato. In pratica possiamo dire questo: viene determinato l'Hazard Rate, che esprime l'inclinazione di una curva, mentre il confronto fra due curve si fa calcolando l'Hazard Ratio (HR), che si ottiene dal rapporto tra due Hazard Rate.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Per semplificare: l'HR esprime qualcosa di simile al rischio relativo o RR con la differenza che questo è dato dal rapporto tra due rischi assoluti mentre l'HR è dato dal rapporto fra due Hazard Rate. Il vantaggio di questo metodo rispetto al log-rank test è chiaro: l'HR con il relativo IC95% permette di **stimare l'entità della differenza trovata** (in pratica è la stessa differenza che passa tra "P" ed RR con il suo IC). Un altro vantaggio del modello di Cox rispetto al log-rank test è che quest'ultimo confronta due curve per una variabile (per esempio due gruppi di pazienti infartuati che assumono una statina a dosaggi diversi) ma non considera la possibile influenza di altre variabili che possono influire sulla mortalità. Invece il modello di Cox permette di confrontare le due curve per quella variabile (statine a dosaggio diverso) a parità di altre variabili che possono interferire (per esempio sesso, età, presenza di comorbidità, stato economico, assunzione di aspirina, ecc.). In gergo tecnico si parla di **analisi multivariata**. Così in un gruppo di infartuati si possono paragonare le curve a seconda se i pazienti assumono o meno una statina, considerando contemporaneamente l'influenza di altre condizioni che si pensa possano impattare sugli outcomes: se vi è o meno uno scompenso cardiaco o una BPCO, se si eseguono o meno controlli cardiologici frequenti e così via.

Per concludere questo capitolo riassumo i concetti che devono rimanere:

1) la mortalità o la sopravvivenza dovrebbero essere determinate sul totale dei soggetti studiati e non solo sui malati, soprattutto se si tratta di studi di screening, in quanto quest'ultima modalità espone a possibili bias di sovradiagnosi e di anticipazione diagnostica

2) nel valutare i risultati di uno studio si devono giudicare gli end-point non solo sulla base dei dati "bruti" registrati alla fine del follow-up ma anche sul confronto fra curve di Kaplan-Meier.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 12.
Gli studi osservazionali

Fino a questo momento abbiamo parlato degli studi di intervento che sono, in genere, randomizzati e controllati. Passiamo ora agli studi osservazionali. Abbiamo già detto che essi sono caratterizzati dal fatto che i ricercatori si limitano ad "**osservare**" quello che succede nella realtà, senza intervenire nel somministrare attivamente un trattamento e nel dividere in modo randomizzato i gruppi. Ovviamente anche in questo tipo di studi può esistere un gruppo di controllo (e allora si parla di studi osservazionali controllati) o può non esistere (studi osservazionali non controllati). I primi sono senz'altro i più comuni.

Ma allora, se anche negli studi osservazionali può esserci un gruppo di controllo, qual è la differenza con gli RCT, potrà chiedere qualcuno. La differenza è che *in nessun caso la suddivisione dei gruppi avviene con metodica randomizzata*. Quindi in questo tipo di studi si verifica uno sbilanciamento nella selezione del campione, che in gergo tecnico viene detto ***bias di selezione.*** Ne abbiamo già accennato in precedenza ma ora è arrivata l'occasione per approfondire l'argomento.

Per capirci e per rendere più chiare le cose prendiamo uno studio osservazionale molto conosciuto e citato, il cosiddetto ***Studio delle Infermiere***. In questo studio vennero seguite per molti anni oltre 60.000 infermiere americane. I ricercatori registrarono il tipo di terapia eventualmente praticata dalle donne, il loro stile di vita, le abitudini alimentari, le malattie che si verificarono, ecc. Come vedete si tratta di uno studio osservazionale classico perché i ricercatori non programmarono né decisero alcun tipo di trattamento e, se qualche terapia veniva effettuate dalle donne arruolate nello

studio, questa era decisa da altri (i medici curanti, le donne stesse), ma mai dai ricercatori. L'analisi dei dati, man mano che venivano raccolti col passare degli anni, evidenziò che le donne in

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

post-menopausa trattate con terapia ormonale sostitutiva avevano una percentuale di eventi cardiovascolari inferiore alle donne che non assumevano ormoni. La differenza di eventi era molto evidente, tanto che si giunse ad affermare un effetto benefico rilevante della terapia ormonale sostitutiva (cosiddetta TOS) sulla prevenzione della cardiopatia ischemica. Per il vero molti avevano fatto notare che ciò avrebbe dovuto essere dimostrato da uno studio di intervento perché lo Studio delle Infermiere molto probabilmente era viziato da un bias. Cosa intendevano dire questi bastian contrari? In poche parole volevano dire che non era la TOS a proteggere dall'infarto ma erano le donne, già in partenza più sane, con uno stile di vita più corretto, di un ceto sociale più abbiente (quindi a maggior controllo sanitario) che sceglievano la TOS. In effetti alcune analisi avevano dimostrato, per esempio, che le donne che sceglievano gli estrogeni avevano già di base valori di pressione arteriosa più bassi rispetto alle donne che non assumevano ormoni. Questo **squilibrio** nel gruppo di donne in TOS viene detto **bias di selezione del campione** e rende non corretto il paragone con le donne che non assumono la terapia, perché diverse. Insomma, come abbiamo detto altre volte, rende "sleale" il paragone perché le donne che non sceglievano la TOS partivano già svantaggiate. Tuttavia la differenza riscontrata nella frequenza di eventi cardiovascolari era di tale entità che la maggior parte degli studiosi la riteneva incompatibile con qualsiasi bias di selezione. Il loro discorso era questo: per quanti bias di selezione ci siano l'azione protettiva degli ormoni sul cuore è così evidente che non può essere dovuta solo ad essi.

Il famosissimo WHI, garantendo la confrontabilità dei due gruppi di donne, essendo randomizzato, ha permesso di stabilire che in realtà quello che veniva suggerito dallo Studio delle Infermiere era del tutto ingiustificato e anzi la TOS aumenta la frequenza degli eventi trombotici.

Facciamo un altro esempio per capire bene cosa siano questi tanto citati bias. Bias è un termine inglese che significa asimmetria, sbilanciamento, pregiudizio. Con lo studio che citerò ora dovrebbe essere del tutto comprensibile che cosa significa in questo contesto. Lo studio è questo:

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

analizzando un registro che conteneva i dati clinici di oltre 57.000 pazienti ricoverati per sindrome coronarica acuta i ricercatori si accorsero che i pazienti che venivano trattati con morfina per il dolore toracico avevano una mortalità più elevata di coloro che non venivano trattati con tale farmaco. E' corretto dire che la morfina aumenta la mortalità nelle sindromi coronariche acute? Prima di rispondere alla domanda consideriamo il tipo di studio: si tratta di uno studio osservazionale in quanto i ricercatori non hanno somministrato alcun tipo di trattamento ma si sono limitati a "osservare" i dati (cioè le cartelle cliniche) riportati in un registro. Esiste ovviamente un gruppo di controllo (quelli che non venivano trattati con morfina) ma la suddivisione nei due gruppi (trattati e non trattati) non è stata decisa dai ricercatori né è avvenuta con metodo randomizzato. La decisione se usare o meno la morfina veniva infatti presa dai medici che curavano i malati. E' possibile quindi che anche qui vi sia un bias di selezione, nel senso che i medici adoperavano l'oppioide nei casi con dolore toracico più grave ed è verosimile che questi pazienti avessero anche una forma di sindrome coronarica acuta più importante. L'aumento della mortalità osservato nei trattati potrebbe quindi essere dovuto al fatto che questi erano pazienti più gravi e compromessi e quindi più a rischio di morte. L'unica conclusione che si può trarre è che lo studio ha dimostrato una **"associazione**" tra uso di morfina e aumento della mortalità, ma non è detto che questa associazione sia del tipo causa-effetto. Per dimostrarlo in modo più convincente bisognerebbe disegnare un RCT, peraltro difficile da immaginare perché eticamente non si potrebbe usare il placebo nel gruppo di controllo.

I bias di selezione sono sicuramente quelli più importanti ma negli studi osservazionali se ne possono trovare di altro tipo. Li passeremo rapidamente in rassegna. Un bias comune è dovuto al fatto che i medici che raccolgono i dati di solito non sono in cieco e quindi possono essere inconsapevolmente influenzati dal conoscere il tipo di terapia effettuata. Si tratta del cosiddetto ***performance bias***.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Inoltre spesso i dati vengono raccolti basandosi sui ricordi retrospettivi dei pazienti e quindi possono essere del tutto inaffidabili (**recall bias**). Vediamo con il solito esempio se si riesce a capire meglio. Uno studio vuol stabilire se l'assunzione di vegetali tre volte alla settimana può ridurre l'incidenza di cancro del colon. Vengono così selezionati da un database oncologico un certo numero di pazienti affetti dalla neoplasia e poi da una popolazione generale si scelgono altrettanti soggetti (paragonabili per età e sesso) non malati. Si fa compilare un questionario ad ogni partecipante allo studio chiedendo di specificare quante volte alla settimana negli ultimi cinque anni hanno assunto vegetali o fibre. Si capisce bene quanto poco valore possano avere i dati raccolti, essendo legati unicamente alla memoria e alla valutazione del singolo soggetto, tutte cose impossibili da controllare per i ricercatori. Qualsiasi sia il risultato ottenuto dallo studio (sia che dimostri un effetto protettivo di vegetali e fibre che nessun effetto) deve essere preso con beneficio di inventario e richiede una conferma da parte di rigorosi studi sperimentali. In realtà questo esempio non è stato scelto a caso: i dati di letteratura sull'argomento
sono contrastanti, proprio perché derivano unicamente da studi osservazionali. Per dire, uno studio osservazionale svedese, contrariamente a studi simili precedenti, ha suggerito che il tipo di dieta ha poca influenza sullo sviluppo del cancro del colon mentre un effetto preventivo sembrano avere il controllo del peso corporeo e l'attività fisica. Anche quest'ultimo è comunque un dato discutibile per il fatto stesso di derivare da uno studio non sperimentale, in cui i bias prima ricordati possono esercitare un effetto confondente.

Non va dimenticato un'altra distorsione che si può annidare negli studi non sperimentali, dovuta alla ***non uniformità dei criteri diagnostici***. Supponiamo per esempio che si voglia determinare la capacità di un immunostimolante di ridurre le frequenza delle riacutizzazioni della BPCO. Vengono quindi selezionati dei medici di famiglia, ognuno dei quali arruola un certo numero di pazienti affetti da BPCO. Si chiede poi ai medici di recuperare dalle cartelle cliniche dei singoli

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

pazienti i trattamenti effettuati e il numero di episodi acuti negli ultimi due anni. E' evidente che qualsiasi fosse il risultato non sarebbe attendibile in quanto non erano pre-specificati i criteri usati per definire cosa si intende per riacutizzazione (un medico potrebbe aver registrato come riacutizzazione solo gli episodi caratterizzati da febbre e tosse con escreato, un altro gli episodi in cui, pur in assenza di febbre, l'escreato diventava di tipo purulento, un altro ancora potrebbe non aver registrato tutti gli episodi, ecc). Si potrebbe addirittura obiettare che siccome non erano pre-definiti neppure i criteri diagnostici per la BPCO la casistica potrebbe riguardare soggetti con malattie molto diverse (un medico potrebbe aver classificato come BPCO anche asme bronchiali, bronchiectasie, ecc.). Come al solito, anche in questo caso, bisognerebbe disegnare un RCT in cui un gruppo viene trattato con immunostimolanti, un altro con placebo e definire in anticipo i criteri di esclusione dallo studio, i criteri diagnostici usati per dire che uno ha una BPCO, i criteri diagnostici usati per dire che uno ha una riacutizzazione.

La conclusione che possiamo trarre da quanto si è detto finora è di prendere sempre con le molle i risultati che derivano da studi osservazionali, i quali dovrebbero, se possibile, essere confermati da RCT.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 13
I vari tipi di studi osservazionali

Lasciato il problema dei bias (il cui concetto a questo punto dovrebbe essere abbastanza chiaro, almeno spero) passiamo ad esaminare la tipologia dei vari studi osservazionali. Ne esistono sostanzialmente di tre tipi:

- gli studi prospettici o longitudinali
- quelli caso-controllo
- quelli cross-sectional o trasversali.

Lo ***studio prospettico o longitudinale*** parte dalla esposizione ad un trattamento per arrivare al risultato. Detto così risulta difficilmente comprensibile ma il solito esempio diraderà le nebbie. Supponiamo di voler determinare se i folati, riducendo i livelli di omocisteina, producono effetti favorevoli a livello cardiovascolare. A questo scopo selezioniamo da una popolazione di un comune i soggetti maschi di età compresa tra i 40 e i 70 anni (ho scelto questo criterio, ma potrebbe essere usato un qualsiasi altro criterio). Una volta avuto l'elenco li contattiamo uno per uno e li dividiamo in base al fatto se assumano o meno supplementi di folati. Dopo 5 anni andiamo a valutare quanti eventi cardiovascolari ci sono stati nei due gruppi. In questo modo possiamo vedere se chi assume folati ha una frequenza di eventi cardiovascolari inferiore, superiore o simile a chi non li assume. Siamo quindi partiti dalla esposizione (assunzione di folati e vitamine) per arrivare ai risultati (eventi cardiovascolari).

Lo ***studio caso-controllo*** parte invece dal risultato per arrivare alla esposizione. Anche qui converrà fare il solito esempio. Poniamo di voler determinare se il fumo di sigaretta aumenta il rischio di cancro della mammella. Selezioniamo da un registro tumori le donne di una data regione affette da neoplasia mammaria e come gruppo di controllo prendiamo donne (paragonabili per età, stato sociale ed economico, ecc.) scelte dall'anagrafe della stessa regione (o anche di una regione diversa). Successivamente andiamo a chiedere quante sigarette hanno fumato negli ultimi 5 anni

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

sia ai casi (cioè le donne con pregresso cancro mammario) che ai controlli (cioè le donne scelte come paragone, senza storia di cancro mammario). Sulla base dei risultati ottenuti possiamo determinare se vi è una correlazione tra il fumo e il cancro mammario. Come si vede in questo caso siamo partiti dal risultato (pazienti affette da neoplasia mammaria) per arrivare alla esposizione (numero medio di sigarette fumato negli ultimi 5 anni).

Lo ***studio cross-sectional (detto anche trasversale)*** esamina esposizione e risultato contemporaneamente, fornendo, per così dire, una fotografia istantanea di una certa popolazione. Per esempio dalla anagrafe di un comune prendiamo tutti i soggetti di età compresa tra 40 e 80 anni e determiniamo quanti sono i diabetici (esposizione) e nello stesso tempo quanti hanno sofferto di ictus (risultato). Possiamo così vedere se la patologia che si vuol studiare (in questo caso il diabete) è associata o meno ad un aumentato rischio di ictus semplicemente valutando se l'ictus risulta più frequente tra i diabeteci rispetto ai non diabetici.

Un altro esempio di studio trasversale è il seguente: vengono selezionate dal database di un gruppo di medici tutte le donne che hanno eseguito una densitometria ossea e nello stesso tempo si registra l'assunzione di statine. Si può determinare, in tal modo, se l'uso delle statine è associato o meno ad un aumento della massa ossea. Faccio notare che in questi casi si parla sempre di "associazione" e non di relazione causale. Così uno studio di tipo trasversale tra i pazienti asmatici troverà che in questi è più frequente l'uso degli steroidi inalatori rispetto ai non asmatici, ma non si può certo concludere che gli steroidi inalatori causano l'asma!

Come abbiamo visto gli studi osservazionali presentano numerose limitazioni. Ci si potrebbe allora chiedere perché vengano fatti e se non sia meglio abbandonarli del tutto. In verità molto spesso e per numerose questioni essi rimangono l'unica fonte disponibile di informazioni e non è realistico aspettarsi ulteriori dati da studi sperimentali. Intanto gli RCT costano molto di più in termini economici e di tempo, poi sono difficili da fare nel caso di malattie rare (è difficile reclutare

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

abbastanza pazienti); gli studi osservazionali invece possono servire per evidenziare effetti collaterali rari di un farmaco oppure possono permettere di formulare delle ipotesi che poi andranno confermate da RCT disegnati ad hoc.

Per esempio alla domanda se l'uso dei contraccettivi orali aumenti o meno il rischio di cancro mammario è possibile rispondere solo rifacendosi a studi di tipo caso-controllo o comunque osservazionali, i quali rimarranno probabilmente l'unica evidenza disponibile in quanto è difficilmente pensabile uno studio randomizzato e controllato. Per citare un esempio ricorderò il famoso Million Women Study, un lavoro di tipo prospettico, che ha dimostrato l'associazione tra terapia ormonale sostitutiva e neoplasia mammaria. Nello stesso modo studi caso-controllo recentemente hanno dimostrato che non vi è alcuna associazione tra vaccinazione MMR ed autismo.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 14
Le meta-analisi e le revisioni sistematiche

Per molte condizioni cliniche esistono svariati RCT in cui i trattamenti hanno dato ora risultati negativi ora positivi ora dubbi e c'è il pericolo che il medico venga a conoscenza solo di alcuni di essi. Si realizza così un effetto distorsivo a cui cercano di porre rimedio le meta-analisi. Si tratta di veri e propri studi che mettono insieme i risultati di più RCT (possono anche essere metanalisi di studi osservazionali) su un dato argomento, permettendo una visione sintetica e più completa del problema. E' evidente il vantaggio di un simile approccio: con un solo studio il medico può disporre di conoscenze aggiornate in cui si tiene conto dei risultati di molti lavori che egli non avrebbe mai avuto l'opportunità di esaminare. Facciamo un esempio molto semplice. Supponiamo che esistano 3 RCT che hanno valutato l'efficacia di due farmaci (che chiameremo A e B) nel ridurre l'incidenza di ulcere complicate in pazienti trattati cronicamente con FANS. In ogni RCT sono stati arruolati 1000 pazienti nel braccio A e 1000 pazienti nel braccio B. La tabella che segue mostra i risultati per ognuno dei 3 RCT sia in termini di eventi che di RR e relativo IC95%, mentre alla fine si vede il risultato ottenuto sommando insieme i 3 studi.

	Farmaco A Trattati/ulcere complicate	Farmaco B Trattati/ulcere complicate	RR IC95%
1° RCT	1000/50	1000/75	0,76 (0,47-0,94)
2° RCT	1000/22	1000/30	0,73 (0,43-1,26)
3° RCT	1000/30	1000/28	1,07 (0,64-1,78)
Totali	3000/102	3000/133	0,77 (0,60-0,99)

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

A questo punto un medico che venisse a conoscenza solo del primo RCT penserebbe che il farmaco A è migliore e sceglierebbe quello, un medico che conoscesse solo il secondo e/o il terzo RCT reputerebbe i due farmaci equivalenti. Il merito delle meta-analisi è considerare insieme i tre RCT come fossero uno studio unico. In questo caso risulta che ci sono stati 3000 pazienti trattati con il farmaco A con 102 ulcere complicate e 3000 trattati con il farmaco B e 133 ulcere complicate. Il farmaco A riduce il rischio del 23% con un IC95% compreso tra 0,60 e 0,99. La meta-analisi suggerisce quindi che il farmaco A è preferibile ma probabilmente la superiorità clinica rispetto a B è minima perché l'estremo di destra dell'IC95% è quasi a ridosso dell'unità. Ovviamente le meta-analisi vengono effettuate con un sistema molto più complesso di quanto ho cercato di mostrare. Di sicuro gli esperti mi boccerebbero senza possibilità di appello perché nella realtà le tecniche di assemblaggio dei risultati degli studi sono molto più complicate e sofisticate di quanto potrebbe sembrare dal mio esempio, che si è limitato a mettere insieme i 3 RCT facendone una semplice somma matematica.

Però penso che il concetto sia chiaro e del tutto sufficiente per gli scopi che mi propongo, anche perché non siamo noi materialmente che dobbiamo fare i calcoli.

Le meta-analisi sono considerate uno degli strumenti più potenti della medicina basata sulle evidenze, ma sono sempre affidabili? In realtà vi sono alcune criticità. Intanto il lettore non sa se sono stati considerati **tutti** gli studi disponibili su quell'argomento oppure solo alcuni; inoltre le meta-analisi possono non tener conto dei risultati degli studi non pubblicati, studi che spesso hanno dato esiti negativi. Questo in gergo tecnico si definisce *pubblication bias* ed è un punto che merita un approfondimento. Non sono rari i casi di studi che avendo dato esito negativo, cioè non essendo riusciti a dimostrare l'utilità di un farmaco rispetto al placebo o ad un farmaco di

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

riferimento (quando addirittura non ne risulti una inferiorità), vengono dimenticati nel cassetto perché è ovvio che non c'è molto interesse da parte dello sponsor ad una loro pubblicazione. Gli editori delle riviste mediche più importanti hanno pubblicato congiuntamente quello che in gergo viene detto "position paper" richiedendo che tutti gli studi debbano essere iscritti in un registro consultabile da chiunque lo voglia fare. In tal modo si viene a creare una enorme banca dati che permette di sapere se vi sono studi di cui si sono perse le tracce. Ovviamente quello degli studi non pubblicati è un punto cruciale e non è facile per chi si appresta a fare una meta-analisi accedere ai dati di questi studi perché questo presuppone anzitutto di esserne a conoscenza e poi di contattare personalmente i vari autori.

Va considerato, ancora, che nel fare una meta-analisi bisognerebbe compiere una ricerca il più completa e sistematica possibile su molte banche dati, in modo da non lasciarsi sfuggire studi importanti su quel determinato argomento. Insomma ci vuole alle spalle un know-how ed una organizzazione coi fiocchi: le meta-analisi non sono per tutti e non le si può improvvisare. Per lo stesso motivo, come si vedrà in seguito, anche la loro interpretazione non è affare di tutti i giorni. Ritornando alle nostre meta-analisi vi è da dire inoltre che se gli studi considerati sono di scarsa qualità anche i risultati complessivi potranno essere meno affidabili. La validità di una meta-analisi dipende pure dalla casistica dei vari studi che la compongono: in genere se si tratta di studi con casistica limitata le conclusioni potrebbero risentirne. Un esperimento molto elegante è quello riportato dal British Medical Journal [Counsell CE et al. The miracle of DICE therapy for acute stroke: fact or fictional products od subgroup analysis? BMJ 1994; 309:1677-1681] : ad un corso sullo stroke a cui partecipavano 2256 persone si pensò di dividere i partecipanti in 44 gruppi, ogni gruppo composto da un minimo di 5 persone e da un massimo di 100. Successivamente ciascuno di questi gruppi veniva suddiviso in due bracci: in pratica ogni gruppo simulava un RCT con un

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

braccio trattamento e un braccio controllo. Ad ogni persona del gruppo veniva dato un dado da gettare: se usciva il numero "sei" si segnava quella persona deceduta, tutti gli altri numeri volevano dire sopravvivenza. E' ovvio che i risultati sarebbero stati del tutto casuali, mentre il risultato teorico atteso della percentuale di decessi era di 100 diviso 6 = 16,7% perché sei sono le facce di un dado.

La tabella sottostante riassume i risultati ottenuti nell'esperimento.

Tutti i gruppi	Decessi trattamento	Decessi controllo	RR e IC95%
N° partecipanti 2256	16%	17,6%	0,8 (0,5-1,1)

Come si vede non c'era differenza statisticamente significativa tra trattamento e controllo; questo è logico perché trattandosi di pura casualità il risultato era atteso in questi termini. I ricercatori suddivisero, allora, i gruppi in base al numero dei partecipanti (più o meno di 40 persone) e trovarono delle sorprese. La tabella sottostante riporta i risultati suddivisi a seconda della numerosità del gruppo.

	Decessi trattamento	Decessi controllo	RR e IC95%
Gruppi con più di 40 persone (11 gruppi per 1190 partecipanti)	19,5%	17,8%	1,1 (0,9-1,4)
Gruppi con meno di 40 persone (33 gruppi per 1066 partecipanti)	12%	17,1%	0,7 (0,53-0,94)

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Assemblando i risultati dei gruppi con più di 40 persone la differenza tra trattamento e controllo non era significativa, mentre lo diventava miracolosamente mettendo insieme i risultati degli studi con pochi partecipanti. In altre parole facendo una specie di meta-analisi separata si ottiene una apparente significatività statistica per i risultati ottenuti nei gruppi con pochi partecipanti, mentre sappiamo che la differenza trovata è del tutto casuale e dovuta al gioco dei dati. Quando gli studi assemblati dalla meta-analisi hanno tutti pochi pazienti arruolati si può correre quindi un pericolo di questo tipo, di scambiare per significativo un risultato che invece non lo è. Però è anche vero che talora si preferisce disporre di vari studi con poca casistica ma ben fatti piuttosto che di pochi studi ampi ma fatti male. Lo scopo della meta-analisi infatti è quello di assemblare insieme molti studi in modo da aumentare la potenza statistica del campione. Tuttavia va sempre ricordato l'acronimo anglosassone GIGO (garbage in, garbage out), come suggerisce Tom Jefferson nel suo "Attenti alle bufale": se in una meta-analisi butti dentro spazzatura non può che uscire spazzatura, ancorché rivestita e nobilitata da abiti matematici. Insomma, l'abito non fa il monaco.

Anche le modalità con cui viene eseguita una meta-analisi e la scelta degli studi da assemblare influenzano i risultati finali. Un esempio molto istruttivo ci permetterà di capire i termini della questione. Nel numero del 9 dicembre 2000 della rivista "The Lancet" vennero pubblicate due meta-analisi sui calcio-antagonisti che si proponevano di valutare se questi farmaci sono altrettanto efficaci degli altri trattamenti nel ridurre gli eventi clinici avversi dell'ipertensione. Pur prendendo in considerazione praticamente gli stessi RCT, le due meta-analisi giunsero a conclusioni in gran parte contrastanti. Nelle prima delle due meta-analisi [Blood Pressare Lowering Treatment Trialists Collaboration. Lancet 2000; 356: 1955-1964] si decise di fare un doppio confronto: da un parte mettere insieme tutti gli studi in cui un calcio-antagonista era stato confrontato con una terapia a base di diuretici/betabloccanti (5 RCT) e dall'altra gli studi che

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

avevano paragonato un calcio-antagonista con un aceinibitore (2 RCT). I risultati di questo modo di procedere furono i seguenti: rispetto a diuretici e betabloccanti i calcioantagonisti riducono in modo significativo lo stroke, mentre non ci sono differenze per gli altri outcomes (mortalità totale, cardiopatia ischemica, ecc.); paragonati agli aceinibitori i calcio-antagonisti aumentano in maniera significativa il rischio di cardiopatia ischemica, mentre non vi sono differenze per gli altri eventi cardiovascolari e per la mortalità. Gli autori quindi concludevano che i calcio-antagonisti sono sostanzialmente equivalenti agli altri trattamenti antipertensivi, riducono il rischio di stroke rispetto a diuretici e betabloccanti e aumentano il rischio coronario rispetto agli aceinibitori. Nella seconda delle due meta-analisi [Pahor M et al. Lancet 2000; 356:1949-1954] si decise una metodologia diversa: vennero raggruppati tutti gli studi (9 RCT) in cui un calcio-antagonista era confrontato con un altro trattamento, per cui il confronto fu calcio-antagonista versus altri trattamenti (diuretici, betabloccanti, aceinibitori). In questa seconda analisi risultò che i calcio-antagonisti aumentavano in modo significativo il rischio di infarto, eventi cardiovascolari maggiori e scompenso cardiaco rispetto agli altri farmaci presi nel loro complesso, mentre riducevano in maniera non significativa il rischio di stroke; la mortalità totale, come nella prima meta-analisi, non differiva tra calcio-antagonisti e altri trattamenti. In questo caso gli autori concludevano che i calcio-antagonisti sono meno efficaci degli altri farmaci antipertensivi.

Volete un altro esempio? Nel 2003 furono pubblicate a breve distanza di tempo, prima su JAMA e poi su Lancet, due ampie meta-analisi sui farmaci antipertensivi. La prima, quella di JAMA [Psaty BM et al. JAMA 2003; 289:2534-2544] dimostrava che i tiazidici a basse dosi sono, per un end-point o per un altro, preferibili agli altri trattamenti; la seconda, quella di Lancet [Blood Pressare Lowering Treatment Trialists Collaboration. Lancet 2003; 362: 1527-1535], suggeriva che sostanzialmente tutte le varie classi di antipertensivi sono egualmente efficaci. Come è possibile

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

una tale differenza di conclusioni? La meta-analisi di JAMA adottava un sistema particolare, detto network meta-analysys, che permette dei confronti indiretti: per esempio se in un trial si confrontano amlodipina ed enalapril e in un altro enalapril e clortalidone questa metodologia permette di confrontare in modo indiretto amlodipina con clortalidone. Si tratta però di un metodo che viene ritenuto meno affidabile del classico confronto diretto. Sarebbe come dire che se il Milan batte la Juventus e la Juventus batte l'Inter, allora il Milan è più forte dell'Inter. Questa conclusione potrebbe andare bene ai tifosi rossoneri ma non certo a quelli nerazzurri. La meta-analisi di Lancet adottava invece la metodologia classica dei confronti diretti ma considerava diuretici e beta-bloccanti insieme, come un'unica classe. Questo è stato ritenuto sleale in quanto si finisce per penalizzare i diuretici tiazidici accorpandoli con i beta-bloccanti su cui esistono dubbi di efficacia nel ridurre le complicanze dell'ipertensione, perlomeno nei soggetti anziani.

Come di può vedere la faccenda non è per nulla semplice e, a seconda della metodologia adottata, una meta-analisi, pur condotta tecnicamente in modo corretto, può portare a una conclusione piuttosto che ad un'altra.

Un altro esempio (prometto che è l'ultimo) mostrerà quanto sia complicata la questione.

Una meta-analisi sui beta-bloccanti pubblicata da Lancet nel 2005 [Lindholm LH et al. Lancet 2005; 366: 1545-1553] concludeva che questi farmaci non dovrebbero essere usati come antipertensivi di prima scelta perché rispetto ad altri trattamenti portano ad un aumento di ictus del 16% (significativo) e della mortalità totale del 3% (non significativo).

Nel 2006 sul Canadian Medical Association Journal [Khan N et al. CMAJ 2006; 174:1737-1742] arriva una contro-metanalisi che muove alla precedente tre obiezioni:

1. ha tralasciato studi importanti, non includendoli nell'analisi
2. ha usato un end-point unico (l'ictus) invece di un end-point composto (ictus, infarto, morte) così

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

che se si può verificare quello che i tecnici chiamano "survival bias": se per esempio i beta-bloccanti riducono le morti coronariche, una maggior quantità di persone viene salvata e può andare incontro ad ictus che così apparentemente sembra più frequente rispetto ad un altro trattamento che non riduce le morti coronariche

3. negli studi vi era una forte eterogeneità: in questi casi può essere sbagliato mettere insieme studi in cui l'età dei pazienti arruolati era di 60-70 anni o più e studi in cui l'età era di 40-50 anni; è preferibile eseguire un' analisi separata degli studi accorpandoli in base all'età dei partecipanti.

In questo modo la meta-analisi del CMAJ arriva a dimostrare che i beta-bloccanti sono paragonabili agli altri trattamenti negli ipertesi giovani, mentre sono meno efficaci negli anziani, soprattutto per quanto riguarda la prevenzione dell'ictus. E però "chi di spada ferisce di spada perisce": infatti a mio modo di vedere è del tutto opinabile anche questo voler mettere insieme tutti i trattamenti antipertensivi, in un unico calderone, per confrontarli con i beta-bloccanti: i diuretici non sono uguali ai calcio-antagonisti né agli aceinibitori! Volerli considerare in un unico blocco (cosa peraltro comune a varie altre meta-analisi) mi sembra una forzatura artificiosa che può portare ad un duplice pericolo; penalizza i farmaci che funzionano di più ed enfatizza i benefici dei farmaci che funzionano di meno.

Insomma la valutazione critica della bontà metodologica di una metanalisi non è un giochetto da poco, richiede conoscenze e competenze specifiche e dimestichezza con la materia. Per dire: gli studi assemblati nella metanalisi possono essere tra loro eterogenei (cioè caratterizzati da una variabilità di risultati più o meno ampia). L'eterogeneità viene calcolata (e per fortuna non spetta a noi farlo) e poi espressa con un numero: si parla di eterogeneità quando questo numero è inferiore a 0,10. Esistono due metodi per fare i calcoli metanalitici, uno detto

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

"fixed- effect" e l'altro detto "random-effect". Non ci interessa ovviamente sapere come si effettuano questi calcoli ma basti dire che se il pool di trials assemblati tra loro è caratterizzata da eterogeneità elevata i calcoli fatti con il metodo del fixed-effect potrebbero dare risultati molto diversi da quelli ottenuti con il metodo "random-effect", che in casi di elevata eterogeneità è il modello consigliato. Come si capisce da queste che, almeno per noi, sono astruserie, valutare se è stato adottato il metodo più affidabile è compito non semplice, che è bene sia riservato agli esperti. I quali però talvolta non hanno vita facile perché può succedere che gli autori della metanalisi non riportino tutti i dati necessari (per esempio non riportano tutti gli eventi riscontrati nei singoli RCT per ogni gruppo ma solo la loro somma), rendendo di fatto molto arduo stabilire la correttezza delle conclusioni. Altri punti che gli esperti di "critical appraisal" considerano sono la sistematicità della ricerca cui abbiamo accennato all'inizio (cioè se la ricerca degli studi da assemblare è stata completa, prendendo in esame le varie banche dati), la presenza di possibili pubblication bias, la qualità degli studi ritrovati, la definizione dei criteri di inclusione nella metanalisi, Questo dovrebbe bastare per farci capire che la valutazione della qualità di una metanalisi non si può improvvisare ma deve essere lasciata a studiosi ed esperti ad hoc preparati.

Le revisioni sistematiche

Le revisioni sistematiche della letteratura sono studi in cui vengono passati al pettine fitto le banche dati esistenti (o almeno così si dovrebbe fare) in modo da ritrovare tutte le evidenze disponibili per un dato argomento. Una delle istituzioni più prestigiose che elabora revisioni sistematiche è la Cochrane Collaboration: in queste revisioni vengono presi in considerazione

anche studi non pubblicati o pubblicati solo in abstract e quando necessario vengono contattati gli autori degli studi per avere ulteriori dettagli. In tale modo si cerca di avere il maggior numero di

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

informazioni possibili e di evitare il bias di pubblicazione. Le revisioni Cochrane prevedono una metodologia di lavoro molto rigorosa e definita: la ricerca viene effettuata sulle principali riviste e sulle banche dati internazionali e gli studi vengono inclusi nell' analisi sulla base di criteri ben specificati. Le revisioni, disponibili a pagamento (gli abstract invece si possono consultare liberamente online), vengono periodicamente aggiornate. L'unico limite sta nel fatto che non tutto il sapere medico vi è rappresentato per cui non sempre si trova la revisione atta a rispondere ad una specifica domanda.

Le revisioni sistematiche possono essere di **tipo qualitativo** quando esprimono un giudizio globale e stringato, ma spesso presentano anche veri e propri calcoli matematici in modo da **effettuare una meta-analisi** degli studi ritrovati e da esprimere il risultato finale in termini di RR o meglio di OR, con relativo intervallo di confidenza. Per ogni studio viene anche riportato il peso che lo stesso ha nel pool. Non sempre dopo aver fatto una revisione sistematica è possibile produrre una meta-analisi perché per esempio gli studi hanno una eterogeneità troppo elevata.

Allora mi direte: ma che differenza c'è tra una meta-analisi e una revisione sistematica? La meta-analisi è semplicemente una metodica che assembla tra loro gli studi su un determinato argomento e, con particolari tecniche statistiche, ne fornisce un risultato di tipo matematico. La revisione sistematica è un processo molto complesso in cui dapprima si compie un esame completo (o così dovrebbe essere) della letteratura, poi si prendono gli studi selezionati sulla base di criteri espliciti e se ne fa una valutazione qualitativa ed eventualmente, se è possibile, una meta-analisi (valutazione quantitativa).

Per spiegarmi meglio passerò in rassegna i vari passaggi compiuti dalla Cochrane per valutare se lo screening del cancro del polmone sia o meno utile. In questo caso l'intervento, ovviamente, non era costituito dalla somministrazione di un farmaco ma dello screening stesso.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 14 – Le metanalisi e le revisioni sistematiche

Il punto di partenza è la domanda: lo screening (con esame dell'escreato, radiografia del torace o TAC spirale) riduce la mortalità da cancro polmonare? Per poter rispondere gli autori non si limitano a cercare gli studi in vari database (Cochrane Central Register of Controlled Trial, MEDLINE, PREMEDLINE, EMBASE) ma compiono anche una ricerca manuale su bibliografie e riviste mediche e discutono con esperti. Come si vede un lavoro imponente per cercare di trovare tutte le evidenze disponibili, compresi studi non pubblicati o presentati solo ai Congressi. Dopo aver ritrovato gli studi che interessano si analizzano i dati secondo l'intenzione a trattare, usando il metodo adatto di pooling (random effect oppure fixed effect) a seconda dell'eterogeneità trovata. Un momento importante della revisione è la valutazione della qualità metodologica degli studi: nel caso specifico gli autori giudicano deboli da questo punto di vista molti dei trials trovati con la loro ricerca. Si passa poi a quantificare l'end-point che interessa (vale a dire la mortalità da cancro polmonare) esprimendola sotto forma di RR e si trova che non c'è una differenza statisticamente significativa sia negli studi in cui lo screening frequente veniva confrontato con screening meno frequente, sia negli studi in cui lo screening con radiografia del torace + esame dell'escreato veniva paragonato alla sola radiografia del torace. Alla fine, dopo aver sottolineato che la ricerca non ha permesso di ritrovare studi in cui screening veniva paragonato a non screening e neppure RCT sullo screening tramite TAC spirale, gli autori chiudono il loro lavoro con le conclusioni che non sono favorevoli allo screening del cancro polmonare.

Come si può vedere una revisione sistematica di qualità è un processo complesso e difficile che richiede personale adeguatamente addestrato e in possesso di conoscenze specifiche. Non è insomma cosa da improvvisare.

Ritorniamo ora all'esempio che avevo fatto all'inizio del capitolo sui tre RCT che avevano confrontato due farmaci per la guarigione dell'ulcera. Sono possibili due situazioni: gli autori hanno

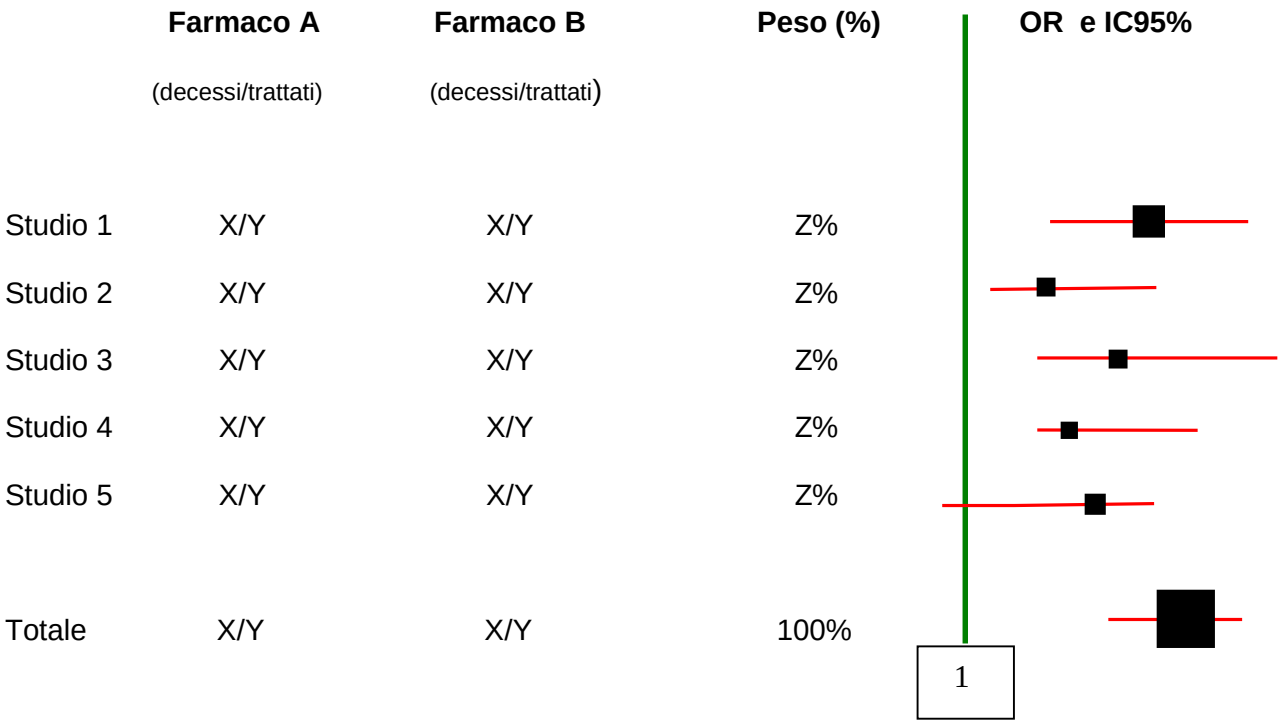
Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

eseguito una rigorosa ricerca bibliografica e una revisione della letteratura e i 3 RCT sono gli unici esistenti (in questo caso la meta-analisi eseguita è presumibilmente importante e di qualità) oppure, all'opposto, gli autori non hanno effettuato alcuna ricerca e si sono limitati a fare una pura operazione matematica mettendo insieme solo i tre studi che erano a loro conoscenza (in questo caso la meta-analisi è una semplice esercitazione tecnico-statistica che non può essere di alcun ausilio al medico perché non si sa se ci sono altri studi, magari più importanti, sull'argomento). Per usare un paragone, impreciso ma che rende l'idea, possiamo dire che nel primo caso è stata preparata una insalatona con "tutte" le verdure dell'orto (ma vengono scartate comunque le verdure marce), nel secondo solo un'insalata di verdure "scelte" a discrezione del cuoco, in cui però possono mancare quelle che più ci piacciono. Insomma per essere credibile una meta-analisi dovrebbe essere sempre preceduta da una revisione della letteratura, meglio se sistematica, altrimenti non è che un semplice esercizio di matematica, tipo quelle espressioni chilometriche che ci facevano fare al liceo.

Naturalmente in mezzo a queste due eventualità estreme vi è tutta una gamma intermedia di studi in cui la meta-analisi è preceduta da una analisi della letteratura più o meno di qualità e così come si passano al setaccio gli RCT per valutare la loro qualità metodologica, così si può fare con le revisioni sistematiche. Ma si tratta di un compito da lasciare agli esperti. Come linea generale si consiglia di fidarsi maggiormente di revisioni sistematiche pubblicate dalla Cochrane Collaboration oppure da Enti Governativi (come per esempio la U.S. preventive Service Task Force).

Nella figura che segue viene riportato, per concludere questo capitolo, lo schema generale con il quale si trovano raffigurati i risultati di una metanalisi. Nell'esempio sono considerati studi in cui erano confrontati il farmaco A e il farmaco B per l'end-point mortalità totale (ovviamente potrebbe trattarsi di qualsiasi altro outcome).

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati



Come si vede per ogni studio viene riportato il numero di decessi e il numero di pazienti arruolati sia per il farmaco A che per il farmaco B, oltre al peso che ogni singolo studio ha nella metanalisi. Il quadrato nero è la stima puntuale dell'OR (la dimensione del quadrato riflette la numerosità dei pazienti arruolati in ogni studio) mentre la linea rossa rappresenta graficamente l'IC95%. L'intervallo di confidenza ottenuto dalla metanalisi è inferiore all' intervallo di confidenza dei singoli studi e indica pertanto una maggior precisione nella determinazione del range entro cui può variare l'OR trovato. In questo caso la conclusione è in linea con quanto trovato da quasi tutti i singoli studi: il farmaco B riduce la mortalità totale rispetto al farmaco A. Naturalmente ho disegnato solo uno schema semplificato, nella realtà vengono riportati nella rappresentazione riassuntiva anche alcuni test statistici che valutano la coerenza degli studi, per esempio il test per l'eterogeneità, ma

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

per i nostri scopi penso sia sufficiente comprendere l'essenziale. Un ultimo appunto merita però **l'eterogeneità che può esserci nei vari trials**: questo aspetto è uno dei punti critici di una meta-analisi. L'eterogenità può essere calcolata in modo statistico, ma per noi può bastare **la tabella che riassume i dati dei vari studi.** Prendiamo la tabella precedente e notiamo che tutti i 5 trials hanno dato, chi più chi meno, un risultato positivo per il farmaco B. Non c'è quindi eterogeneità e il risultato globale può ritenersi ragionevolmente affidabile. Supponiamo ora che il primo e il secondo RCT abbiano dato risultato statisticamente positivo per il farmaco B, il terzo RCT abbia dato un risultato positivo per il farmaco B ma statisticamente non significativo, il quarto e il quinto RCT un risultato statisticamente a favore del farmaco A. Come si vede ci sono studi in controtendenza e, soprattutto se i loro IC95% non si sovrappongono, è probabile che l'eterogeneità sia tale per cui non sarebbe corretto neppure fare la metanalisi. Nell'esempio appena fatto anche se il pooling (cioè l'assemblaggio dei dati) desse un risultato statisticamente favorevole al farmaco B è possibile che per alcuni pazienti si ripetano i risultati negativi ritrovati nei due ultimi trial per questo farmaco. In questi casi più che fare una somma "matematica" sarebbe più utile cercare di capire il perché di una eterogeneità così elevata.

Ma perché si può ritrovare una elevata eterogeneità? Una delle cause è che le popolazioni arruolate nei vari studi sono troppo diverse. Poniamo per esempio che A e B siano due antibiotici per la polmonite e che, per assurdo, il farmaco A funzioni meglio nei giovani mentre il farmaco B funzioni meglio negli anziani. Se nei primi due RCT sono stati arruolati prevalentemente anziani con polmonite è evidente che i risultati saranno a favore di B; se negli ultimi due RCT sono stati arruolati prevalentemente giovani è ovvio che funzioni meglio il farmaco A. In questo esempio, volutamente paradossale ma didattico, la diversità della popolazione arruolata nei vari trias rende conto della eterogeneità dei risultati trovati.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 15
Il grado delle evidenze e le linee guida

Da quanto abbiamo detto finora risulta chiaramente che le prove scientifiche non sono tutte eguali, sono più o meno affidabili a seconda del tipo di studio che le ha generate. E' ampiamente accettata una classificazione delle evidenze secondo una scala che vede al primo posto meta-analisi , revisioni sistematiche ed RCTs di elevata qualità e all'ultimo posto l'opinione di esperti. A livello intermedio stanno gli studi osservazionali prospettici e quelli caso-controllo.

Lo schema sottostante riassume il livello delle evidenze secondo la Scottish Intercollegiate Guidelines Network (SIGN). Altre organizzazioni utilizzano una nomenclatura lievemente diversa, ma la sostanza del discorso non cambia.

Livello delle evidenze

1++: Meta-analisi di elevata qualità, review sistematiche di RCTs, RCTs con un rischio molto basso di bias

1+: Meta-analisi condotte in modo corretto, revisioni sistematiche di RCTs con un rischio basso di bias

1: Meta-analisi, revisioni sistematiche o RCTs con un rischio alto di bias

2++: Revisioni sistematiche di alta qualità di studi di coorte o caso controllo oppure studi di coorte o caso controllo con un rischio molto basso di bias e alta probabilità che la relazione sia causale

2+: Studi caso-controllo o di coorte ben condotti con un rischio basso di bias e probabilità moderata che la relazione sia causale

2: Studi caso-controllo o di coorte con un rischio elevato di bias e rischio significativo che la relazione sia non causale

3: Studi non analitici, per esempio case reports o serie di casi

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 15 – Il grado delle evidenze e le linee guida

Le linee guida

Le linee guida sono un concentrato di raccomandazioni sul comportamento da tenere in determinate situazioni. Vengono elaborate da Enti Governativi o da Società Scientifiche e la loro utilità è fuor di dubbio in quanto riassumono le conoscenze disponibili su un certo argomento (diabete, ipertensione, dispepsia, ecc). Di solito le raccomandazioni possiedono un grado più o meno elevato di forza in dipendenza delle evidenze su cui si basano. Ogni Società o Ente che elabora le linee guida adotta una propria classificazione delle raccomandazioni che però sono generalmente tra loro sovrapponibili.

Nello schema sottostante vengono schematizzate le raccomandazioni e il grado di forza adottato dalla SIGN.

Grado di forza delle raccomandazioni

Grado A: raccomandazioni che si basano su evidenze 1++, 1+ e 1

Grado B: raccomandazioni che si basano su evidenze 2++

Grado C: raccomandazioni che si basano su evidenze 2+ e 2

Grado D: raccomandazioni che si basano su evidenze di grado 3 e 4

Le linee guida non sono però una panacea che risolve d'incanto tutte le difficoltà della clinica. Una revisione su 227 linee guida pubblicata nel 2001 sul Canadian Medical Association Journal ha dimostrato che molte forniscono raccomandazioni non basate su prove scientifiche ma su pareri personali degli autori, che molte non vengono periodicamente aggiornate e che solo nel 9% dei casi potevano essere consigliate ai medici perché di buona qualità. Da non scordarsi poi gli

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

eventuali conflitti di interesse degli estensori delle linee guida, che sono esplicitati ma non eliminati.

Le linee guida inoltre solo raramente vengono applicate dai medici per una serie di ragioni che sono state ben studiate: sono troppo complesse, non tengono conto del contesto particolare in cui debbono essere trasferite, sono calate dall'alto e viste dagli operatori sanitari come una prevaricazione alla loro libertà clinica. Ciononostante esse sono divenute uno strumento indispensabile per i medici che se ne debbono servire con intelligenza , scegliendo quelle emanate da Enti Governativi (come la US Preventive Services Task Force) o da Società Scientifiche importanti. Però non debbono mai - e neppure possono - sostituirsi al giudizio clinico del medico. Il British Medical Journal scrisse qualche anno fa (cito a memoria): noi vi diamo le evidenze ma sta a voi prendere le decisioni. Se le Linee Guida possono essere di qualità variabile, mi potete chiedere, come facciamo a giudicare se sono un buon piatto o se vanno gettate dalla finestra? Esiste un metodo molto complesso (detto AGREE) che valuta una linea guida sotto ben 23 punti di vista. Naturalmente si tratta di un metodo complicato e che richiede molto tempo, sicuramente da lasciare agli esperti di critical appraisal. Però anche noi, medici pratici, possiamo fare la nostra parte utilizzando un **filtro a tre sole voci**, suggerito da due autori italiani (Grilli e Liberati) in un articolo apparso sul Lancet nel 2000. Vediamo di che cosa si tratta.

Per prima cosa andiamo a controllare se la linea guida specifica chiaramente da chi è stata elaborata e la qualifica di ogni autore. Siccome la linea guida deve essere rigorosa nelle sue raccomandazioni e poi anche applicata nella pratica, dovrebbe prevedere la presenza non solo di specialisti della materia, ma anche di esperti in metodologia della ricerca e della valutazione degli studi, di medici di famiglia, di associazioni di pazienti, di amministratori. Insomma dovrebbe essere

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

una specie di coro polifonico. In più ogni autore dovrebbe chiaramente dichiarare i suoi conflitti di interesse, anche se, come ho già detto, la dichiarazione non li elimina. Mi direte: ma perché anche amministratori e associazioni dei pazienti? Gli amministratori sono importanti perché va valutata anche la sostenibilità economica: è inutile stilare una linea guida che poi non può essere applicata perché enormemente dispendiosa in termini di risorse umane o monetarie. Le associazioni dei pazienti sono essenziali perché sempre più si tende a dare importanza al punto di vista di chi la linea guida se la deve poi caricare sulla spalle. La seconda cosa che dobbiamo andar a vedere è se nella linea guida esiste una apposita sezione in cui sono descritti i metodi che sono stati utilizzati per trovare e valutare le referenze bibliografiche (un po' come il metodo che abbiamo visto essere usato nelle revisioni Cochrane). Infine per ogni raccomandazione la linea guida dovrebbe chiaramente specificare se essa deriva da una o più revisioni sistematiche, da meta-analisi, da RCT, da studi osservazionali, da analisi a posteriori o di end-point secondari o surrogati, ecc. Insomma dovrebbe dire su quali evidenze si basa il tal consiglio e qual è la qualità di queste evidenze. Se una raccomandazione si basa sul consenso vuol dire che degli esperti si sono riuniti a tavolino ed hanno deciso tra di loro che la cosa da fare è quella e non un'altra. Si tratta sicuramente di un consiglio degno della massima attenzione, dato che viene dato da esperti, ma, come diceva quel tale: "In Dio crediamo, gli altri devono portare qualche prova a sostegno di quello che affermano". In conclusione se una linea guida non risponde a questi tre criteri di qualità possiamo tranquillamente saltarla a piè pari senza timore di perdere qualcosa d'interessante. Dimenticavo: controllate sempre che sia contenuta la data in cui la linea guida è stata stilata e quella in cui è prevista una revisione.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 16
Ancora statistica?

Per finire in bellezza affronteremo un campo che riguarda i test diagnostici. Cosa c'entra questo con gli studi clinici, mi chiederete? C'entra perché non è raro trovare studi in cui si parla di sensibilità e specificità di un test. Di che cosa si tratta? Cercherò di spiegarmi nella maniera più semplice possibile.

La **sensibilità** di un test nel diagnosticare una determinata malattia rappresenta la percentuale dei soggetti malati che il test riesce a scoprire. Se dico che un esame diagnostico ha una sensibilità del 90% nell'individuare una certa malattia vuol dire che su 100 soggetti **malati** di quella patologia, sottoposti al test, 90 saranno positivi e 10 saranno negativi (**falsi negativi**).

La **specificità** del test invece si riferisce ai soggetti sani e mi dice la percentuale di soggetti sani che risultano negativi al test. Se dico che un test ha una specificità dell'80% vuol dire che su 100 soggetti sani 80 avranno un test negativo ma 20 lo avranno positivo (**falsi positivi**).

Conoscere la sensibilità e la specificità di un test è importante perché permette di avere un'idea dell'**efficacia** del test stesso.

Un esempio, che riguarda il tanto contestato PSA, è illuminante. Questi dati derivano dall'analisi prelimiare su 8.575 uomini arruolati nel braccio placebo dello studio Prostate Cancer Prevention Trial, uno studio su circa 19.000 uomini di almeno 55 anni senza storia di cancro prostatico, con un valore iniziale di PSA inferiore o uguale a 3 ng/ml ed esplorazione rettale negativa. I pazienti sono stati seguiti per sette anni con dosaggio annuale del PSA ed esplorazione rettale. Veniva eseguita una biopsia prostatica se il PSA superava il valore di 4 ng/ml oppure se la palpazione della prostata diventava anormale. Usando un cut-off di PSA di 4 ng/ml la sensibilità del test era del 21% mentre la specificità era del 93.8%. Riducendo il cut-off a 1.1 ng/ml si riuscirebbe a diagnosticare 83.4% dei tumori ma i falsi positivi aumenterebbero al 61%. Se si pone la soglia

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

decisionale a 3.1 ng/ml la sensibilità del test è del 32% e la specificità dell'87% mentre per valori di 2.1 ng/ml si ha una sensibilità del 53% e una specificità del 73%. Se si decidesse quindi di abbassare il cut-off a valori attorno a 2 ng/ml si potrebbe scoprire poco più della metà dei tumori perché il 47% dei malati risulterebbe falsamente negativo, nello stesso tempo ben il 27% dei sani risulterebbe falsamente positivo; il che non è poco se si screena l'intera popolazione maschile attorno ai 50-55 e se si considera che per ora non ci sono dimostrazioni che lo screening riduca la mortalità da cancro prostatico.

Tuttavia se ci limitassimo a quanto detto sopra, potremmo essere tratti in inganno. Se ho un test che ha una sensibilità e una specificità del 90% potrei anche dirmi soddisfatto perché ci saranno solo un 10% di falsi positivi e di falsi negativi. Purtroppo la statistica non è così semplice. Infatti al medico pratico i valori di sensibilità e di specificità di un test interessano relativamente: quello che interessa di più è sapere qual è la probabilità che quella persona sia veramente malata in caso di test positivo e quale la probabilità che sia sana in caso di test negativo. Nel primo caso si parla di **Valore Predittivo Positivo di un test (VPP**) e nel secondo caso di **Valore Predittivo Negativo (VPN).**

Per poter calcolare il VPP e il VPN è necessario conoscere la **prevalenza** della malattia nella popolazione in esame. Userò il termine prevalenza sia nel senso classico (percentuale di soggetti malati in una popolazione generale) sia in senso lato (per esempio percentuale di soggetti malati in un gruppo di pazienti che presentano determinati sintomi; in questo caso si dovrebbe dire, più correttamente, “probabilità pre-test”).

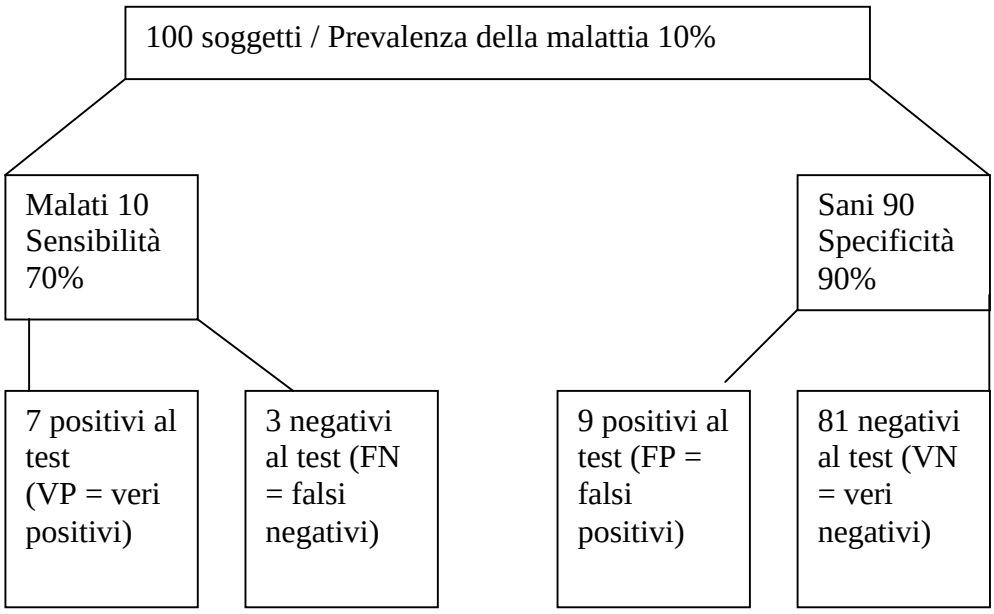
Ipotizziamo ora una malattia che abbia un prevalenza del 10% (cioè su 100 soggetti 10 siano malati e 90 siano sani) e un test per scoprirla che abbia una sensibilità del 70% e una specificità del 90%.

Come sappiamo la sensibilità andrà applicata ai 10 malati per cui, essendo del 70%, 7 saranno

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

positivi al test (veri positivi) e 3 saranno negativi (falsi negativi). Al contrario la specificità andrà applicata ai 90 soggetti sani per cui, essendo del 90%, 81 risulteranno negativi al test (veri negativi) e 9 saranno positivi (falsi positivi).

Tutto questo è schematizzato nella figura sottostante.



A questo punto è possibile calcolare i tre parametri che più interessano al medico: il valore predittivo positivo del test (VPP), il valore predittivo negativo (VPP) e l'accuratezza globale del test (overall accuracy).

Il **VPP** risponde alla domanda: su 100 soggetti con test positivo quanti sono quelli affetti dalla malattia (cioè i VP)? Basta dividere il numero dei VP trovato per il numero dei positivi totale (VP + FP) e moltiplicare per 100. Nel caso in esame 7/16 = 0,43 x 100 = 43%. **In altre parole ogni 100 test positivi ci si deve aspettare che 43 siano veri e 57 siano dei falsi positivi.**

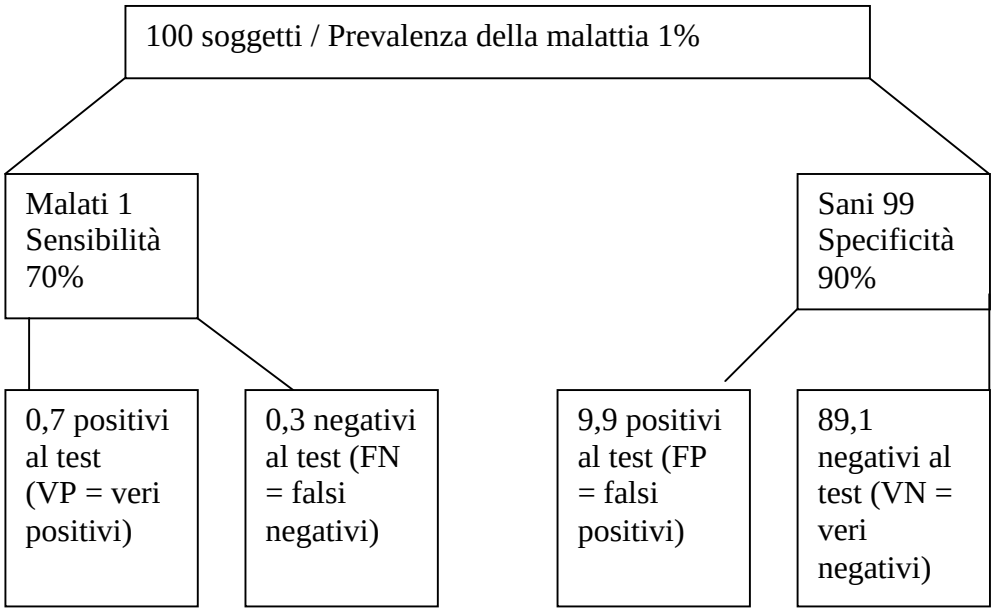
Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Il **VPN** risponde invece alla domanda: su 100 soggetti con test negativo quanti sono quelli che effettivamente non hanno la malattia (VN)? Basta dividere il numero dei VN trovato per il numero dei negativi totali (VN + FN) e moltiplicare il tutto per 100. Nel caso in esame 81/84 = 0,96 x100 = 96%. **Su 100 test negativi ci si deve aspettare che 96 siano corretti mentre 4 sono dei malati (FN).**

L'accuratezza totale del test (**overall accuracy**) risponde alla domanda: su 100 test eseguiti quanti sono quelli corretti? E' evidente che la risposta si ottiene sommando i VP e i VN. Nel caso in esame rispettivamente 7 + 81 = 88. Questo vuol dire che il test restituisce una risposta corretta nell'88% dei casi e in 12 casi sbaglia. Si può esprimere l'accuratezza o come percentuale (88%) oppure come corrispondente numero decimale (0,88). Si noti come il valore ottimale di accuratezza di un test sia 100% (= 1) in quanto in questo caso il numero dei veri positivi e dei veri negativi corrisponde al totale dei soggetti sottoposti al test e non vi sono né falsi positivi né falsi negativi. Questa si verifica, teoricamente, quando sensibilità e specificità sono entrambe del 100%. Purtroppo si tratta di una eventualità teorica che non si riscontra quasi mai nella pratica clinica, in quanto nessun test è perfetto fino a questo punto.

Come si può facilmente intuire il VPP e il VPN sono enormemente influenzati dalla prevalenza della malattia. Ipotizziamo per esempio che il test con sensibilità 70% e specificità 90% servisse per una malattia che ha una prevalenza non del 10% ma più bassa. Ovviamente cambierebbe tutto e i dati sarebbero completamente stravolti, come si può vedere dalla figura che segue in cui è riassunto il comportamento del test nel caso la prevalenza fosse dell'1%

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati



Su 100 soggetti solo 1 sarebbe malato (quindi i VP sarebbero 0,7 e i FN 0,3), mentre 99 sarebbero sani (89,1 sarebbero VN e 9,9 sarebbero FP).

Il VPP diventerebbe 6,6% e il VPN sarebbe 99,6%. In altre parole a parità di sensibilità e specificità se la prevalenza della malattia è alta aumenta il VPP e si riduce il VPN mentre se la prevalenza della malattia è bassa si riduce il VPP e aumenta il VPN.

L'accuratezza diagnostica del test, in questo caso, sarebbe data dalla somma dei VP e dei VN, quindi sarebbe 89,8%, che, come si vede, non cambia molto rispetto all'esempio precedente. Quello che cambia molto è soprattutto il VPP che si riduce: questo vuol dire che se si trova un test positivo in caso di bassa prevalenza della malattia si tratta probabilmente di un falso (e purtuttavia saranno necessari ulteriori accertamenti per confermare o escludere la patologia in esame) .

La tabella sottostante mostra alcuni valori predittivi positivi e negativi di un test al variare della

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

sensibilità, della specificità e della prevalenza della malattia.

	Prevalenza della malattia 1%	***Prevalenza della malattia 10%***	***Prevalenza della malattia 30%***
Sensibilità 80% Specificità 80%	VPP 3,8% VPN 99,7%	VPP 30,7% VPN 97,2%	VPP 63,1% VPN 90,3%
Sensibilità 90 Specificità 80	VPP 4,3% VPN 99,8%	VPP 33,3% VPN 98,6%	VPP 65,8% VPN 94,9%
Sensibilità 80% Specificità 90%	VPP 7,4% VPN 99,7	VPP 47,0% VPN 97,5%	VPP 77,4% VPN 91,3%
Sensibilità 90% Specificità 90%	VPP 8,3% VPN 99,8%	VPP 50,0% VPN 98,7%	VPP 79,4% VPN 95,4%
Sensibilità 70% Specificità 70%	VPP 2,3% VPN 99,5%	VPP 20,5% VPN 95,45%	VPP 50% VPN 87,5%
Sensibilità 70% Specificità 60%	VPP 1,7% VPN 99,4%	VPP 16,2% VPN 94,7%	VPP 42,8% VPN 82,3%
Sensibilità 60% Specificità 70%	VPP 2,0% VPN 99,4%	VPP 18,1% VPN 94,0%	VPP 46,1% VPN 80,3%

Come si può notare il valore predittivo positivo di un test è influenzato relativamente poco da sensibilità e specificità mentre risente molto di più dalla prevalenza della malattia nella popolazione testata: quando la prevalenza è bassa il VPP è molto basso pur in presenza di buoni valori i sensibilità e specificità.

Possiamo dire, come regola generale, che in caso di alta prevalenza della malattia un test positivo è probabilmente esatto e un test negativo ha buone probabilità di essere sbagliato mentre in caso di bassa prevalenza di malattia un test positivo è probabilmente sbagliato e uno negativo è

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

probabilmente esatto.

Primo scenario

Applichiamo ora quanto appena appreso ad un caso reale: il test “fattore reumatoide” e il test anticorpi anticitrullina” per la diagnosi di artrite reumatoide. Il fattore reumatoide ha una sensibilità media del 60% e una specicificità media del 79%.Gli anticorpi anticitrullina hanno una sensibilità media del 65% e una specificità media del 95%. La prevalenza dell’artrite reumatoide stimata nella popolazione italiana è dello 0,5%.

Si otterranno i seguenti risultati:

Fattore reumatoide:
VPP = 1,41%
VPN = 99%
Accuracy = 78,9%

Anticorpi anticitrullina:
VPP = 6,1%
VPN = 99,8%
Accuracy = 94,8%

Come si può vedere entrambi i test sono molto utili se negativi in quanto escludono quasi con certezza la presenza di artrite reumatoide. Invece se i test risultano positivi la probabilità di avere una artrite reumatoide è bassa, soprattutto con il test “fattore reumatoide”.

Tra i due test quello che performa meglio è il dosaggio degli anticorpi anticitrullina perché restituisce una risposta esatta 95 volte su 100 mentre il fattore reumatoide sbaglia 11 volte su 100.

Secondo scenario

Supponiamo adesso di somministrare i due test non ad una generica popolazione ma a soggetti che presentano determinati sintomi che lasciano sospettare un’ artrite reumatoide. In questo caso

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

ovviamente la prevalenza della malattia non sarà più dello 0,5%, ma molto più elevata perché si sta esaminando una popolazione selezionata. Si ipotizzi che la prevalenza in questa popolazione sia del 50%.

Si otterranno i seguenti risultati:

Fattore reumatoide:
VPP = 74%
VPN = 66%
Overall accuracy = 69,5%

Anticorpi anticitrullina
VPP = 92,8%
VPN = 73%
Overall accuracy = 80%

Come si può vedere in questo caso aumenta il VPP dei due test ma diminuisce il VPN.

Così se si applicasse il dosaggio degli anticorpi anticitrullina ad una popolazione indifferenziata un test positivo sarebbe probabilmente errato (VPP 6,1%) mentre un test negativo sarebbe esatto (VPN 99,8%). Invece applicandolo ad una popolazione con sospetta artrite reumatoide un risultato positivo è probabilmente vero (VPP 92,8%9) mentre un risultato negativo ha buone probabilità di essere sbagliato (VPN 73%).

Inoltre in entrambi gli scenari si riduce l'accuratezza diagnostica. Come si spiega questo fenomeno? A prima vista si potrebbe pensare che se il test viene somministrato ad una popolazione selezionata in cui la prevalenza dell'artrite reumatoide è più elevata il test dovrebbe avere una performance migliore. In realtà non è così: siccome i due test hanno una sensibilità non elevata si ha, rispetto allo scenario in cui il test viene somministrato ad una popolazione indifferenziata, un aumento cospicuo dei falsi negativi. Ovviamente questo va ad incidere sulla

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

performance del test. Se per ipotesi la sensibilità dei test fosse stata del 95% l'overall accuracy sarebbe stata simile nei due scenari.

L'apparente riduzione della overall accuracy nel secondo scenario non deve trarre in inganno. Per esempio riferendoci al test "anticorpi anticitrullina" è vero che quando viene somministrato ad una popolazione indifferenziata fornisce un risultato esatto nel 94% dei casi mentre se viene somministrato ad una popolazione selezionata i risultati esatti sono "solo" 80 su 100. Però ipotizziamo un medico che decida di screenare con il test tutti i suoi assistiti: dovrà richiedere moltissimi test e ne avrà quindi, in valore assoluto, molti di sbagliati; al contrario un medico che richiede il test solo a chi presenta sintomi sospetti di artrite reumatoide chiederà il test pochissime volte e di conseguenza il numero di test sbagliati sarà, in valore assoluto, molto minore.

Dal che consegue che gli esami, al di fuori di ben specifici programmi di screening, andrebbero richiesto solo in presenza di un motivato sospetto clinico.

Piccolo quiz finale

Questo test venne somministrato da un ricercatore ad un gruppo di medici: "La probabilità che una donna abbia un cancro al seno è dell'0.8%. In una donna affetta da cancro al seno la probabilità che la mammografia sia positiva è del 90%; se invece non ha il cancro c'è una probabilità del 7% che la sua mammografia sia positiva. Se una donna si sottopone alla mammografia e questa risulta positiva quanto è probabile che si tratti realmente di tumore?" Provate a rispondere senza leggere sotto. Poi controllate la risposta.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Risposta

Portiamo l'esempio a 1000 donne: la probabilità che abbiano un cancro del seno è 0.8%: ciò significa che su 1000 donne 8 hanno un cancro e 992 sono sane.In queste 8 donne con cancro la mammografia è positiva nel 90% dei casi, cioè in 7,2 e negativa in 0,8.

Nelle 992 che sono sane la mammografia è positiva nel 7% cioè in 69,44.

Trascurando le virgole si avranno, nelle 1000 donne esaminate, 69 (falsi positivi) + 7 (veri positivi) = 76 mammografie positive. Solo il 9,2% di esse avrà però un cancro! Infatti 7 corrisponde al 9.2% di 76 e il VPP del test in questo esempio è del 9.2%.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 17
Per gli appassionati

In questo capitolo verranno affrontati due argomenti (la likelihood ratio o rapporto di verosimiglianza e la curva ROC) che possono essere interessanti per gli appassionati, ma se non rientrate tra questi lo potete tranquillamente saltare senza correre il rischio di aver perso qualcosa di essenziale.

Abbiamo visto che un determinato test possiede una sua sensibilità e una sua specificità e che conoscendo questi due dati e la prevalenza della malattia in una determinata popolazione è possibile calcolare il numero dei veri positivi, dei falsi positivi, dei veri negativi e dei falsi negativi e da questi risalire al valore predittivo positivo e al valore predittivo negativo del test.

Però esiste un altro modo di mettere in relazione sensibilità e specificità tra di loro: si tratta del **likelihood ratio (o rapporto di verosimiglianza**), abbreviato di solito con la sigla LR. Si possono avere due tipi di LR, uno positivo (LR+) e uno negativo (LR -).

Il LR + mette in relazione la probabilità di trovare il test positivo nei malati con la probabilità di trovarlo positivo nei sani. Calcolarlo è facile: si divide Sensibilità per 100 - Specificità. Così se per esempio un test ha una sensibilità del 90% e una specificità del 90% si avrà LR + = 90 / 10 = 9. In pratica significa che se su 10 sani il test è positivo 1 volta, su 10 malati è positivo 9 volte.

Conoscere il LR + di un test permette di calcolare la probabilità post-test della malattia una volta che sia nota la probabilità pre-test. Il modo di procedere è un po' indaginoso e mi servirò del solito esempio. Supponiamo che in una determinata popolazione la prevalenza di una certa malattia sia del 10%. Ciò significa che la probabilità pre-test di trovare tale malattia nel campione esaminato è del 10% (in altre parole la probabilità pre-test non è altro che la prevalenza della malattia in quella

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

popolazione). Supponiamo anche che il LR + di un determinato test per tale malattia sia 4.

Per prima cosa bisogna trovare il cosiddetto **odds pre-test**. Come ho già detto, il concetto di "odds" è tipico del modo anglosassone e si richiama all'ambiente delle corse. Noi siamo abituati a dire che nella popolazione esaminata la probabilità di trovare un malato è del 10%. Invece l'odds è la probabilità di malattia rispetto alla non malattia. Nel caso in esame la non malattia riguarda 90 persone su 100. Quindi l'odds si calcola dividendo 10 per 90 (cioè per i non malati).

L'odds pre-test sarà quindi: 10/90 = 0,11.

Se si moltiplica l'odds pre-test per la LR + (che è 4) si trova **l'odds post -test**: 0,11 x 4 = 0,44.

Conoscendo l'odds post-test è possibile calcolare **la probabilità post-test** con la seguente formula: odds post test/odds post test + 1.

Sarà quindi: 0,44/1 + 0,44 = 0,44/1,44 = 30%.

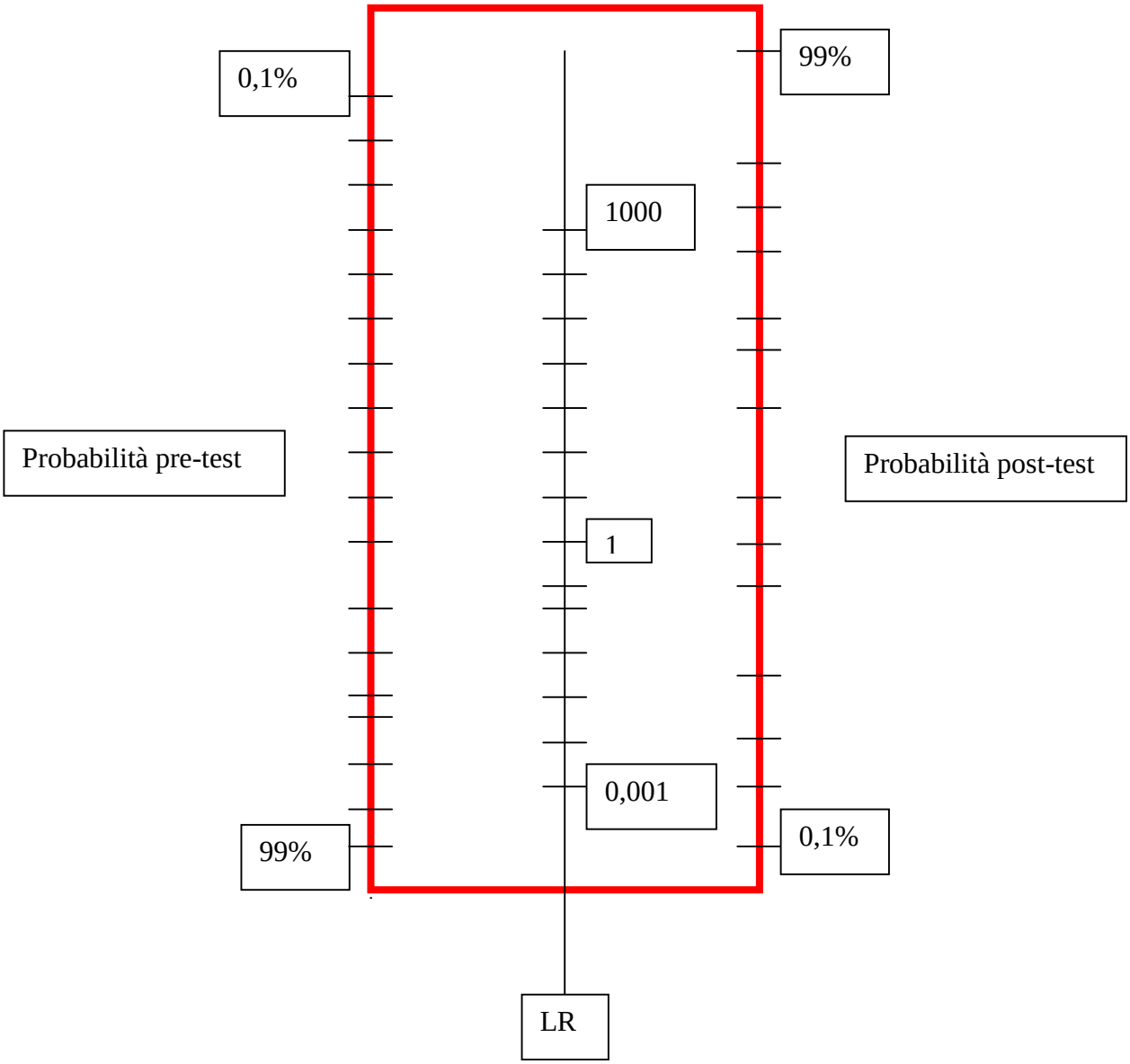
Possiamo quindi dire che se la probabilità pre-test della malattia è 10% e il LR + è 4 la probabilità post test sale al 30%. Si noti che **la probabilità post-test non è altro che il valore predittivo positivo** del test, che abbiamo già visto nel capitolo precedente.

E' ovvio che tanto maggiore è il LR + tanto più utile è il test: un valore appena superiore a 1 cambia di poco la probabilità pre-test, un valore compreso tra 5 e 10 provoca un aumento moderato della probabilità pre-test mentre la performance maggiore la ottiene un test con LR + superiore a 10 perché in questo caso si ottengono notevoli cambiamenti dal pre-test al post-test. Nel caso esemplificato la probabilità pre-test passa dal 10% al 30% del post-test.

Teoricamente il LR + potrebbe avere anche risultati più piccoli di 1: in questo caso la probabilità post-test diventa addirittura inferiore a quella pre-test. Un test con sensibilità del 55% e specificità del 30% avrebbe un LR + di 0,78. Se la probabilità pre-test è del 10%, dopo il test si riduce al 7,9%!

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Conoscendo la probabilità pre-test e il LR è possibile calcolare la probabilità post-test senza dover fare i calcoli prima descritti utilizzando un apposito nomogramma messo a punto da Fagan.



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

La figura mostra come funziona il nomogramma di Fagan: unendo con un righello la probabilità pre-test e il LR si ottiene la corrispondente probabilità post-test senza dover ricorrere ai calcoli.

Ovviamente come esiste un LR + esiste anche un **LR –** che mette in relazione la **probabilità di trovare il test negativo nei malati con la probabilità di trovarlo negativo nei sani**. Si può calcolare con la formula: 100 – Sensibilità diviso Specificità. Per esempio con sensibilità 90% e specificità 80% esso sarà dato dalla divisione 10/80 = 0,125.

Facciamo un esempio per capirci. Si supponga una malattia con probabilità pre-test del 10% e un test con LR – di 0,05. Si avrà:

- odds pre-test = 10/90 = 0,11
- odds post-test = 0,11 x 0,05 = 0,0055
- probabilità post-test = 0,005/1 + 0,0055 = 0,0054

In altre parole la negatività del test fa diminuire la probabilità dal 10% a poco più del 5 per mille!

Per **concludere** sul significato del Likelihood Ratio:

1) LR + si riferisce ai test trovati positivi e tanto più è grande tanto più aumenta la probabilità pre-test
2) LR – si riferisce ai test trovati negativi e tanto più è piccolo tanto più riduce la probabilità pre-test

Vediamo di spiegarci ancora con degli esempi.

Eseguo un test che ha un LR + di 40 e un LR – di 0,003. Se trovo il test positivo so che aumenta considerevolmente la probabilità pre-test.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 17 – Per gli appassionati

Se trovo il test negativo so che è molto verosimile si tratti di un vero negativo perché riduce considerevolmente la probabilità pre-test.

Le cose cambiano per valori diversi di LR: facciamo l'ipotesi di un test con LR + di 1 e LR – di 1. Sia che trovi il test positivo sia che lo trovi negativo so che questo non cambierebbe la probabilità pre-test, quindi si tratterebbe di un test inadatto a guidare le mie decisioni diagnostiche.

Tutto questo discorso ovviamente non può prescindere dal valore della probabilità pre-test di malattia perché, come abbiamo visto nel capitolo precedente, le cose possono essere diverse a seconda se questa è bassa o alta. Per esempio se applico un test con LR + di 40 ad una probabilità pre-test di 1% si ottiene una probabilità post-test del 28,5%: in ogni caso nel 71,5% dei casi un test positivo sarebbe comunque un falso positivo!

Ma perché si usa il LR? La maggiore utilità si ha per test non dicotomici. In questi casi ogni valore ha una sua sensibilità e una sua specificità: ricorrendo al LR per ogni valore è possibile fare un paragone diretto. In pratica si può paragonare sensibilità e specificità del test per i vari valori senza dover ricorrere alla costruzione della curva ROC (vedi in seguito) ma semplicemente confrontando i valori del LR. Si prendano per esempio la sensibilità e la specificità del PSA per vari valori: si può calcolare per ognuno di essi il LR , in modo da paragonarli, come mostra la tabella che segue.

	Sensibilità	***Specificità***	***LR +***
PSA 2 ng/mL	88	27	1,2
PSA 4 ng/mL	71	50	1,42
PSA 6 ng/mL	50	71	1,7
PSA 10 ng/mL	27	90	2,7

Come si può vedere il LR + più favorevole sarebbe quello del valore 10 ng/mL perché è quello che

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

più aumenta la probabilità pre-test. Tuttavia in questo caso si avrebbe un elevato numero di falsi negativi: ben il 73% dei malati sfuggirebbe al test. Il miglior compromesso tra sensibilità e specificità in questo caso si realizza scegliendo un valore compreso tra 4 ng/mL e 6 ng/mL .

Un esempio famoso che viene riportato in tutti i testi di statistica medica è quello della ferritina per la diagnosi di anemia ferropriva: valori di ferritina inferiori a 15 hanno un LR + di 52 e quindi il test conferma la carenza di ferro, valori compresi tra 35 e 64 hanno un LR + di 1 e quindi non confermano né escludono la diagnosi di sideropenia, valori superiori a 95 hanno un LR + di 0,08 e quindi il test è molto negativo ed esclude la carenza di ferro.

Un'altra utile applicazione del LR si può vedere nell'esempio che segue. Si avverte che i dati di probabilità pre-test e LR + riportati non corrispondono a dati reali e servono solo a scopo didattico.

Una paziente di 48 anni si presenta in uno studio medico riferendo crisi di dolore toracico saltuarie, prevalentemente dopo sforzi fisici, di breve durata. Il medico che la visita pensa subito ad una angina pectoris però sa che in una popolazione non selezionata di donne di 48 anni la prevalenza di cardiopatia ischemica (quindi la probabilità pre-test) è bassa, del 2% circa. Tuttavia il medico sa anche che quella specifica paziente è ipertesa da qualche anno e soffre di diabete, inoltre le caratteristiche del dolore toracico sono abbastanza suggestive di una origine ischemica. La presenza di questi tre segni insieme (ipertensione, diabete, caratteri del dolore suggestivi) possiede un LR + per cardiopatia ischemica di 20. La probabilità pre-test del 2% diventa allora una probabilità post-test del 28%. Il medico decide di fare eseguire un elettrocardiogramma da sforzo al cicloergometro e trova un sottoslivellamento del tratto ST di circa 2 mm. Questo segno possiede un LR + per cardiopatia ischemica di 22. Allora la probabilità che era del 28% prima di eseguire il test diventa una probabilità post-test dell' 89%. La paziente viene ricoverata e sottoposta a coronarografia che evidenzia in effetti una grave stenosi dell'arteria discendente anteriore.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

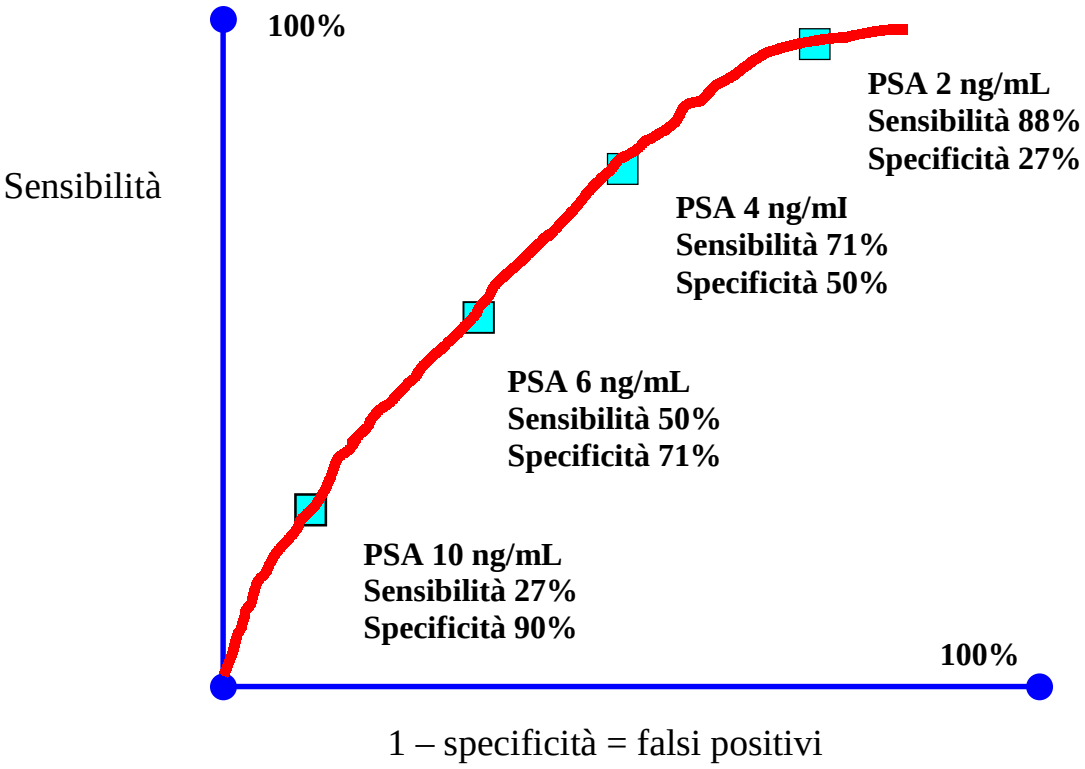
L'esempio seguente invece riguarda la diagnosi di polmonite e contiene probabilità pre-test e LR + tratte dalla letteratura [Grassi M. Diagnosticare il polmone malato. Occhio Clinico 2005; 7: 12-14].
Si presenta in ambulatorio un paziente con malattia respiratoria acuta. In una popolazione di soggetti con tale patologia la prevalenza di polmonite (probabilità pre-test) è del 5%. Ciò significa che 5 hanno una polmonite e 95 hanno una patologia respiratoria indifferenziata. Tuttavia vi sono alcuni segni e sintomi che se presenti possono aumentare la probabilità pre-test. Per esempio la presenza di febbre ha un LR + per polmonite medio di 1.9. Questo vuol dire che la probabilità post test diventa del 9% circa. Il segno clinico tachipnea ha un LR + medio di 2,4: la probabilità pre-test che era diventata del 9%, se è presente anche tachipnea, passa al 18% circa. La tachicardia ha un LR + medio di 1,9: se è presente anche tachicardia la probabilità post-test aumenta ulteriormente e arriva al 29% circa. Infine se sono presenti anche crepitii polmonari, il cui LR + medio è di circa 2, la probabilità post-test passa dal 29% al 44%. In conclusione quasi un paziente su due con una flogosi respiratoria in cui siano associati iperpiressia, tachicardia, tachipnea e crepitii polmonari ha una polmonite. Ovviamente a questa conclusione ci si può arrivare anche su base puramente clinica basata sull'esperienza ma la quantificazione della probabilità post-test è utile in quanto permette di dire che il bicchiere può essere mezzo pieno ma anche mezzo vuoto: un po' più di un paziente su due con le caratteristiche cliniche suddette non ha una polmonite, mentre sulla base dell'esperienza si sarebbe portati a giudicarla molto più probabile.

La curva ROC

E’ evidente che ogni test avrà una sua sensibilità e specificità in base al valore scelto per discriminare i malati dai sani. Ho già fatto l’esempio il PSA, usato per lo screening del cancro della prostata. Se si sceglie un cut-off di 4 ng/mL questo valore non garantisce di suddividere con un

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

taglio netto i sani dai malati: vi saranno dei sani con valori superiori a 4 ng/mL (falsi positivi) e dei malati con valori inferiori (falsi negativi). Se si decide di porre il cut-off ad un valore più basso (per esempio 2 ng/mL) sicuramente si avranno meno falsi negativi ma aumenteranno i falsi positivi. Al contrario se si pone il cut-off ad un valore più elevato la maggiore specificità sarà scontata da una riduzione della sensibilità. La rappresentazione grafica di tutto questo si può fare con un sistema di assi cartesiani in cui sull'asse delle ordinate si pone la sensibilità e su quello delle ascisse il numero dei falsi positivi (vale a dire 1 – specificità). Per ogni valore di PSA si avranno valori di sensibilità e di specificità diversi e in questo modo ogni valore di PSA sarà individuato da un punto derivante dalla intersezione della sensibilità e della specificità rispettive. Unendo i vari punti così determinati si costruisce la curva ROC (Receiver Operating Characteristics).



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Capitolo 17 – Per gli appassionati

Per PSA di 10 ng/ml la sensibilità è bassa (cioè si perdono molti tumori) ma la specificità è elevata (cioè vi sono pochi falsi positivi). Progressivamente aumenta la sensibilità e si riduce la specificità man mano che si abbassa il valore. Per valori di PSA di 2 ng/ml la sensibilità è massima (si identificano quasi tutti i casi di tumore) ma nello stesso tempo si avranno molti falsi positivi perché si riduce la specificità.

Nel decidere il cut-off di un esame conviene spesso scegliere un compromesso tra specificità e sensibilità, per esempio prendendo il punto della curva che più si avvicina all'angolo superiore sinistro del diagramma, in questo caso un valore compreso tra 4 e 6 ng/mL.

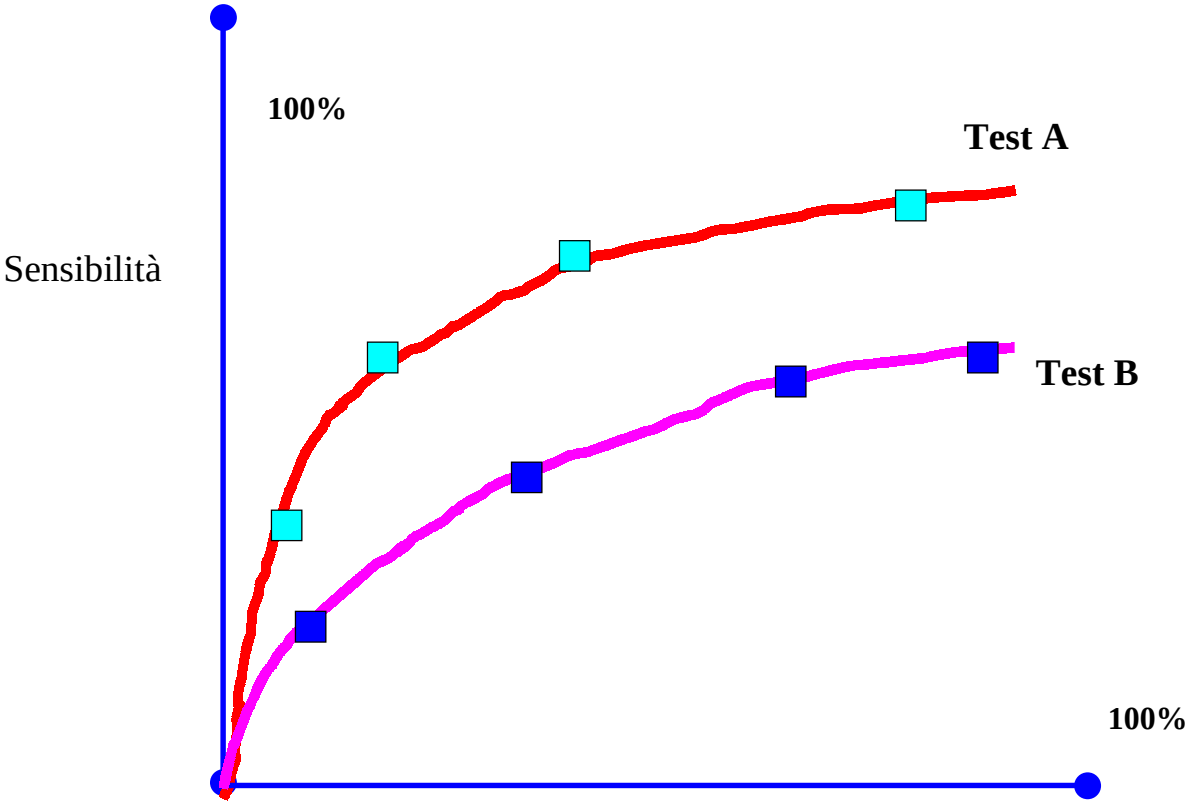
Si osservi che tanto più la curva si avvicina all'angolo superiore sinistro del diagramma tanto più ampia è l'area che essa sottende (area sotto la curva) e quindi tanto più efficace è il test.

Comunque non sempre, nella scelta del cut-off, si sceglie il punto che rappresenta il miglior compromesso tra sensibilità e specificità. Dipende anche dalla condizione che il test diagnostica. Per esempio nel caso del dosaggio delle troponine, che servono a discriminare se vi è o meno una necrosi miocardica, si può privilegiare la sensibilità a scapito della specificità e quindi scegliere un cut-off basso (parte in alto a destra della curva) che consente di avere pochi falsi negativi pur potendosi avere un maggior numero di falsi positivi.

Le curve ROC permettono anche di paragonare l'accuratezza di due test usati per la diagnosi di una determinata malattia.

Si supponga di avere due test per al diagnosi della malattia "X", il TEST A e il TEST B, le cui rispettive curve ROC sono esemplificate nella figura seguente.

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

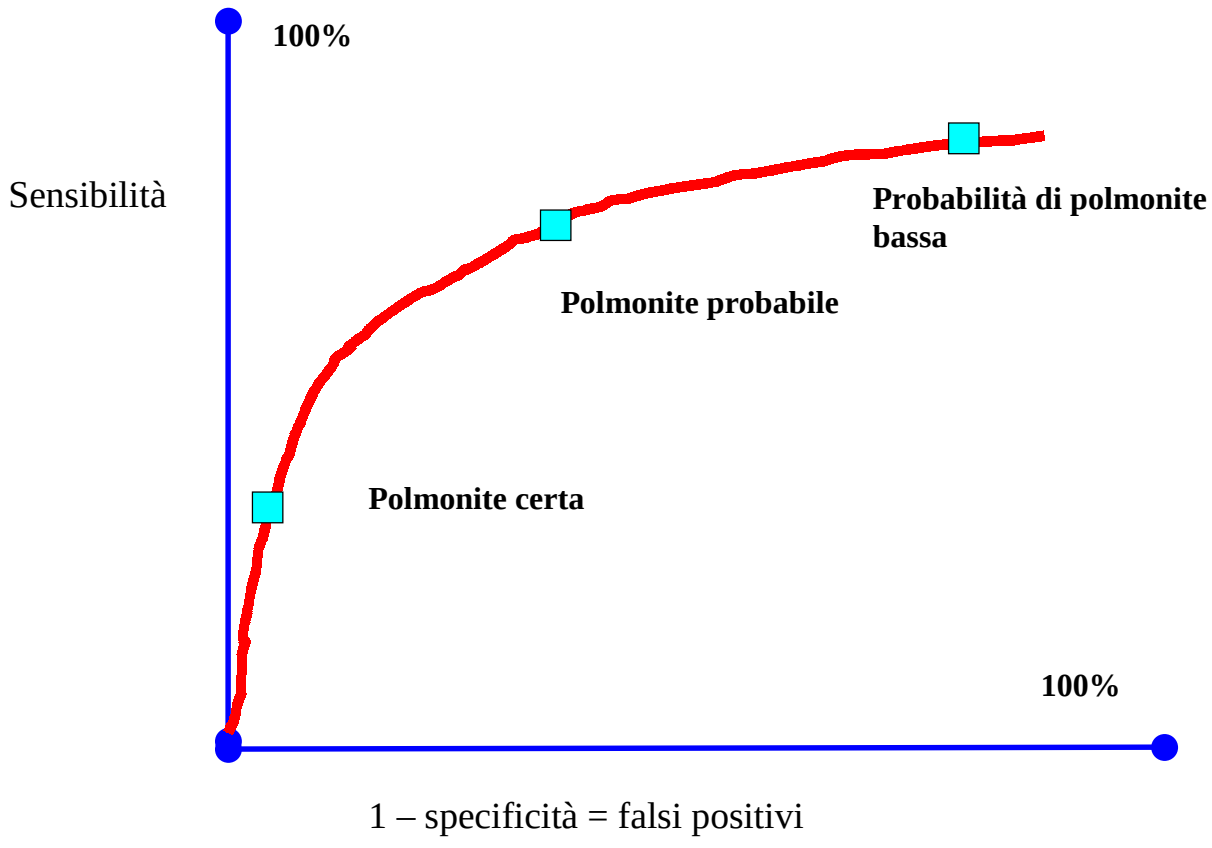


Il test A si avvicina di più all'angolo superiore sinistro e la relativa curva sottotende un'area maggiore rispetto a quella del test B. Il test A quindi avrà una performance superiore al test B.

Non necessariamente le curve ROC si costruiscono per test espressi con valori quantitativi, ma si possono disegnare anche per test che si esprimono in modo qualitativo come per esempio la radiografia del torace per la diagnosi di polmonite. Si possono prevedere vari cut-off decisionali per il trattamento del paziente. Se il quadro radiologico viene considerato dal radiologo come

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

probabilità “molto elevata di polmonite” e si trattano solo questi casi si avrà una elevata specificità ma si possono perdere tutti i casi di polmonite con quadro radiologico non patognomonico. Se si decide di trattare, oltre ai casi definiti certi e probabili, anche quelli descritti come “probabilità bassa di polmonite” si avrà una specificità ridotta (cioè si trattano anche quadri che non sono una polmonite) ma con sensibilità elevata (si tratteranno praticamente tutte le polmoniti). In un punto intermedio della curva si situeranno i casi definiti come probabilità intermedia.



Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Appendice
Riferimenti e links essenziali

Statistica di base

- http://www.nilesonline.com/stats : Un sito su probabilità e statistica che introduce i concetti fondamentali fornendo esempi pratici
- http://research.ed.asu.edu/msms/multimedia/multimedia.cfm : The Statistical Instruction Internet Palette (SIIP) - National Institutes of Health
- http://www.epibiostat.ucsf.edu/epidem/epidem.html : Epidemiology (indirizzi vari)
- http://www.unimaas.nl/~epid/ : Epidemiology - Maastricht University
- http://www.stats.gla.ac.uk/cti/links_stats/books.html : Statistics Resources - Publishers & Textbooks
- http://www.bayesian-initiative.com/Bay_Resources.htm : The Bayesian resource list. Compendio di risorse utili per la statistica Bayesiana.
- http://www.cne.gmu.edu/modules/dau/stat/index.html : Mappa statistica
- http://www.williamoslerhc.on.ca/ : Sito dedicato a William Osler e ai suoi insegnamenti circa la pratica clinica moderna.
- http://members.aol.com/johnp71/javastat.html : Interactive Statistical Calculation Pages. Pacchetto statistico di pubblico utilizzo.
- http://www.stat.unipg.it/iasc/: IASC home page. International Association for Statistical Computing. Servizio offerto dal Department of Statistics – Università di Perugia
- http://www.dst.unive.it/: Dipartimento di Statistica. Università Cà Foscari di Venezia
- http://lib.stat.cmu.edu/: StatLib Index. Hosted by the Department of Statistics at Carnegie Mellon University; sistema per distribuire software statistico, datasets, informazioni per email, FTP e WWW.
- http://www.nilesonline.com/stats/: Statistics Every Writer Should Know. La statistica base che ognuno dovrebbe conoscere
- http://davidmlane.com/hyperstat/index.html: HyperStat Online Contents. Ipertesto introduttivo alla statistica
- http://www.cne.gmu.edu/modules/dau/prob/probability_bdy.html: The Probability Line Approfondimento sulle probabilità.
- http://calculators.stat.ucla.edu/powercalc/ : Power Calculator. Offre la possibilità di calcolare il potere di uno studio
- http://molbio.info.nih.gov/molbio/: Computational Molecular Biology at NIH. Mantenuto da High Performance Scientific Computing Section, Center for Information Technology National Institutes of Health

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Statistica in Medicina

- http://www.paho.org/english/country.htm: Evoluzione e statistica su malattie in america
- http://www.census.gov/hhes/www/hlthins/hlthins.html: Health Insurance Statistics
- http://www.who.int/whosis: Health and health-related statistical information from the WHO Global Programme on Evidence for Health Policy. Statistiche aggiornate su molti campi medici. Motore di ricerca su pubblicazioni e siti

Software

- http://www.camcode.com: versione shareware di *StatDirect*, un software semplice da usare, che consente di calcolare statistiche, trattare regressioni multiple, logistiche, ecc.
- http://www.apl.it: offre la versione shareware di *Statgraphics*, un software suddiviso in pacchetti separati per lavorare con la statistica da quella elementare via via fino ad un livello professionale. Per informazioni si può accedere al sito http://www.statgraphics.com.
- http://www.spss.it: *SPSS* è un pacchetto ricco di test statistici, si interfaccia con excel e consente di analizzare i problemi più complessi.
- http://www.epiinfo.it/ : *l'Epinfo* programma gratuito reso disponibile dal Center for Disease Control.
- http://www.stata.com: Consente di diffondere i programmi statistici scritti nel linguaggio di programmazione *STATA*

Libri e riviste

- http://davidmlane.com/hyperstat/index.html : Enciclopedia on line con motore di ricerca, indici, glossari e links ad altri libri

L'Italia in numeri

- http://www.istat.it: I numeri sull'Italia riassunti dall'Istituto Nazionale di Statistica

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati

Con la speranza di avervi offerto un piccolo aiuto per decidere senza l'indovina…

Renato Rossi - Sopravvivere tra numeri e statistica © 2011. Edito on line da www. pillole.org - Tutti i diritti riservati