Introduzione all'approccio critico alla decisione clinica

Giovanni Casazza
e Giorgio Costantino

Milano University Press

Giovanni Casazza e Giorgio Costantino

INTRODUZIONE ALL'APPROCCIO CRITICO ALLA DECISIONE CLINICA

Milano University Press

Introduzione all'approccio critico alla decisione clinica / Giovanni Casazza e Giorgio Costantino. Milano: Milano University Press, 2024.

ISBN 979-12-55101-02-4 (print)

ISBN 979-12-55101-03-1 (PDF)

ISBN 979-12-55101-04-8 (EPUB)

DOI 10.54103/milanoup.164

Pubblicato da:
Milano University Press

Via Festa del Perdono 7 – 20122 Milano

Sito web: https://milanoup.unimi.it

e-mail: redazione.milanoup@unimi.it

L'edizione cartacea del volume può essere ordinata in tutte le librerie fisiche e online ed è distribuita da Ledizioni (https://www.ledizioni.it/)

Indice

Premessa

L'idea di scrivere questo volume nasce dall'esperienza di attività didattica e di formazione nell'ambito della metodologia clinica che abbiamo svolto nel corso degli ultimi anni, che ha coinvolto, oltre che studenti di scuole di medicina di vari livelli, anche giovani medici all'inizio del proprio percorso clinico e professionale. Ed è proprio a loro che vogliamo rivolgerci. Nel nostro immaginario il lettore tipico di questo volume è ben rappresentato dal dottor CR, un medico di 28 anni che si è laureato da poco e sta facendo la scuola di specializzazione in Medicina Interna (ma potrebbe anche essere Cardiologia, Neurologia, Medicina d'Urgenza, Chirurgia...). Ogni giorno in reparto incontra medici più esperti che non di rado fanno e dicono cose completamente diverse e spesso CR si trova a sostenere le opinioni dell'uno rispetto all'altro, più per simpatia o autorevolezza, che per convinzione. Alla sera, nel buio della sua cameretta, qualche volta si chiede: "Ma è possibile che in medicina ognuno faccia cose diverse? È vero che secondo alcuni la medicina è un'arte, ma forse così è eccessivo…". Come può CR orientare il proprio futuro clinico e professionale?

Partendo da questa riflessione, la nostra idea è che questo volume possa un po' aiutare tutti i medici durante il loro percorso formativo, fornendo qualche strumento metodologico essenziale per orientarsi nella giungla della letteratura scientifica. Non vuole essere un testo di statistica, ma un testo che mostra come alcuni concetti statistici, o più in generale metodologici, possano essere d'aiuto anche nel migliorare la pratica clinica e nel gestire l'incertezza insita nella professione medica. Per rimanere collegati alla realtà clinica quotidiana abbiamo deciso di iniziare ogni capitolo con un caso clinico e con esempi presi da vari articoli di interesse clinico (di volta in volta potrete trovare i riferimenti bibliografici). Il nostro obiettivo è cercare di coniugare competenze metodologiche con esperienze cliniche e di farlo, quando possibile, in modo poco formale, e cercando comunque sempre di privilegiare l'aspetto di interpretazione critica dell'evidenza.

Partiremo dalla ricerca bibliografica, e, passando attraverso la descrizione delle diverse tipologie di studi e di alcuni fra i principali aspetti metodologici e statistici, arriveremo al ragionamento clinico e alle decisioni, per concludere con alcune considerazioni sull'incertezza che permea la quotidianità clinica. Al termine di ogni capitolo abbiamo inserito i "Punti chiave", cioè i concetti principali che pensiamo debbano essere acquisiti. Inoltre, sempre al termine di ogni capitolo, potete trovare la "Bibliografia consigliata", un elenco di alcuni articoli o volumi che riteniamo possano essere utili per approfondire i concetti trattati di volta in volta. Per non appesantire eccessivamente il testo, alcuni degli argomenti più ostici, trattati solo superficialmente, saranno approfonditi

in appendice. Infine, al termine del testo, potrete trovare un ricco glossario che dovrebbe facilitare la comprensione dei termini più tecnici utilizzati nei vari capitoli.

Buona lettura.

1. Ricerca di letteratura

1.1 Formulazione del quesito

Il signor PGD, degente al letto numero 15 del vostro reparto, ha 85 anni, è affetto da fibrillazione atriale permanente, vasculopatia cerebrale, insufficienza renale cronica moderata, pregressi episodi di scompenso cardiaco ed è al suo secondo ricovero nell'arco di tre mesi per anemia da carenza marziale. Vi chiedete se valga la pena continuare la terapia anticoagulante orale o sia meglio sospenderla.

Nella pratica clinica quotidiana ogni giorno ci troviamo di fronte a quesiti clinici di cui non è facile trovare le risposte, a volte perché le risposte non sono disponibili, a volte perché è difficile persino formulare la domanda appropriata, a volte infine perché la strategia di ricerca è complessa.

Obiettivo di questo capitolo è quello di suggerire un metodo per formulare domande che possano trovare risposta nella letteratura medica disponibile.

In particolare analizzeremo:

1.1.1 la domanda

a. il tipo di quesito
b. i modi di formularlo

1.1.2 le fonti di informazione

1.1.3 la strategia di ricerca della risposta.

Vedremo quindi come eseguire praticamente una ricerca utilizzando PubMed.

1.1.1 La domanda

a. Il tipo di quesito

I quesiti si possono suddividere in quesiti background e quesiti foreground.

Quesiti background

Sono quelli centrati su di un argomento (es. malattia) e non su uno specifico paziente. Riguardano quindi conoscenze di base, sono posti in genere da medici con minor esperienza e la risposta ad essi è facilmente reperibile in libri di testo, review, linee guida. Esempi di quesiti background sono: "Qual è l'iter diagnostico della polmonite acquisita in comunità?"; "Qual è la terapia da intraprendere in caso di trombosi venosa superficiale?"; "Come studiare un tumore dell'intestino?"; "Quale può essere la causa di una dispnea o la dose terapeutica della ticlopidina?".

Quesiti foreground

Sono centrati su uno specifico paziente. Poiché sono indirizzati ad una diagnosi, prognosi o terapia particolare per quel paziente, solitamente sono posti da medici con maggiore esperienza e difficilmente trovano risposta adeguata in libri di testo. Esempi di quesiti foreground sono: "Bisogna sospendere o continuare la terapia anticoagulante in un paziente con valvola meccanica mitralica che si presenti con un'emorragia maggiore?"; "In un paziente con versamento pleurico essudatizio e pleuroscopia negativa, è indicata la PET per escludere una patologia neoplastica?".

b. I modi di formulare il quesito

Tornando al nostro paziente: se tentassimo di rispondere alla domanda "È indicata la terapia anticoagulante in un paziente di 85 anni con scompenso cardiaco, insufficienza renale, fibrillazione atriale ed anemia?" inserendo in PubMed o altri database tutti questi termini o cercando in un libro di testo un paragrafo dedicato ad un tale caso, sicuramente il risultato sarebbe deludente. Dobbiamo riformulare il quesito in modo utile ad ottenere una risposta.

Per prima cosa è bene scomporre la domanda nei singoli elementi che la costituiscono, sapendo che è buona norma ricercare una risposta per volta. Nel caso di PGD ad esempio le domande saranno:

- Qual è l'aumento di rischio emorragico in un paziente con le caratteristiche del signor PGD che assume terapia anticoagulante rispetto al rischio emorragico di base di un paziente, con le stesse caratteristiche, che non assume nessuna terapia?
- Qual è la riduzione di rischio di ischemia che lo stesso paziente sperimenterebbe con il trattamento?
- Come possiamo calcolare entrambi i rischi per il nostro paziente?

Verosimilmente dovremo cercare delle scale di rischio, validate, con le quali esprimere il punteggio di rischio totalizzato dal nostro paziente in base ai fattori di rischio che lo caratterizzano.

Per formulare correttamente un quesito è utile utilizzare l'acronimo **PICOS** dove:

- **P** indica **Paziente** o Patologia o Problema;
- **I** indica **Intervento** (provvedimento da prendere) che può consistere in una terapia o in un esame diagnostico o uno score;
- **C** indica l'intervento di **Confronto** (provvedimento suggerito dalla pratica comune che si sospetta non essere ottimale nel caso specifico);
- **O** indica **Outcome**, cioè esito (effetto atteso clinicamente rilevante in base al quale decidere);
- **S** indica il tipo di studio che potrebbe rispondere meglio al nostro quesito. Vedremo in seguito, ad esempio, che, se il quesito sarà l'efficacia di una

terapia, lo studio randomizzato controllato (RCT) sarà il disegno ottimale, mentre, per una domanda diagnostica, sarà più indicato un disegno trasversale (anche detto cross-sectional). Molto spesso potremmo comunque partire controllando che non siano già presenti revisioni sistematiche. In questo caso, potremmo sicuramente iniziare da lì per guadagnare tempo e permetterci di avere una risposta maggiormente affidabile.

Per esemplificare, se dobbiamo valutare quale sia la riduzione del rischio di ischemia cerebrale in un paziente in fibrillazione atriale trattato con anticoagulanti, la nostra domanda sarà così formulata:

- **P:** in un paziente con **fibrillazione atriale**
- **I:** la terapia con **anticoagulanti**
- **C:** rispetto al **placebo** o all'**aspirina** (inizialmente il confronto può essere omesso)
- **O:** di quanto riduce il rischio di **ischemia cerebrale**?
- **S: revisione sistematica** o **RCT**

Seguendo questo schema, è più semplice identificare le parole chiave per eseguire una ricerca che possa dare una risposta adeguata. È bene ricordare che un buon suggerimento per aiutarci a fare una ricerca in modo adeguato è quello di pensare al tipo di studio che potrebbe rispondere alla nostra domanda arruolando il paziente in questione.

1.1.2 Le fonti di informazione

Il passo successivo è trovare le risposte. Attraverso Internet possiamo accedere alle fonti di informazione primarie (database primari) e secondarie (database secondari).

I database primari contengono citazioni di pubblicazioni scientifiche relative a studi primari. Contenendo molte citazioni, la ricerca attraverso questi siti avrà il vantaggio di essere molto sensibile, ma poco specifica. Il tempo necessario a selezionare fra le risposte ottenute quella più appropriata risulterà così piuttosto lungo.

I database primari più importanti sono:

- MEDLINE (https://pubmed.ncbi.nlm.nih.gov/) prodotto dalla National Library of Medicine (NLM), è il database bibliografico più completo e diffuso. PubMed è il portale di accesso a MEDLINE gratuito più usato.
- EMBASE (www.embase.com) prodotto da Elsevier Science, contiene, rispetto a MEDLINE, una maggiore quota di letteratura europea. La sovrapposizione dei due database è circa del 30%, ma EMBASE non è accessibile da un portale gratuito.
- CINAHL (https://www.ebsco.com/academic-libraries): è il più importante database dedicato alle scienze infermieristiche, ma anch'esso non è accessibile da un portale gratuito.

I database secondari sono costruiti "filtrando" un considerevole volume di letteratura primaria, contengono meno citazioni rispetto ai database primari e la ricerca attraverso di essi è meno sensibile, ma più specifica.

I database secondari più utilizzati sono:

- COCHRANE LIBRARY (https://www.cochranelibrary.com/), prodotta dalla Cochrane (http://www.cochrane.org/), un network internazionale creato con l'obiettivo di «preparare, aggiornare e disseminare revisioni sistematiche degli studi clinici controllati sugli effetti dell'assistenza sanitaria e, laddove non siano disponibili studi clinici controllati, revisioni sistematiche delle evidenze comunque esistenti». La Cochrane Library include i seguenti database:
  - *The Cochrane Database of Systematic Reviews* (Cochrane Reviews);
  - *The Database of Abstracts of Reviews of Effects* (DARE);
  - *The Cochrane Central Register of Controlled Trials* (CENTRAL);
  - *The Cochrane Database of Methodology Reviews* (Methodology Reviews);
  - *The Cochrane Methodology Register* (Methodology Register);
  - *Health Technology Assessment Database* (HTA);
  - *NHS Economic Evaluation Database* (NHS EED).

  Il prodotto più importante della Cochrane è il Cochrane Database of Systematic Reviews, che contiene migliaia di revisioni sistematiche di letteratura e meta-analisi su numerosi argomenti di interesse clinico.
- SITI DI LINEE GUIDA: le linee guida sono uno strumento esplicitamente nato per orientare la pratica clinica. Rappresentano, nel migliore dei casi, «raccomandazioni di comportamento clinico prodotte attraverso un processo sistematico di ricerca delle evidenze, coerenti con le conoscenze sul rapporto costo/beneficio degli interventi sanitari, intese a facilitare a medici e pazienti la scelta delle modalità di assistenza più appropriate in specifiche circostanze cliniche». Hanno l'obiettivo di condensare un notevole volume di conoscenze in un formato facilmente consultabile ed utilizzabile dal medico nel singolo paziente (vedi capitolo Linee guida).

  L'accesso, gratuito, a linee guida è possibile da vari siti:
  - *National Institute of Clinical Excellence (NICE)*
  - (https://www.nice.org.uk/guidance/published) banca dati di linee guida inglesi.
  - *Scottish Intercollegiate Guidelines Network* (SIGN)
  - (https://www.sign.ac.uk/our-guidelines/) banca dati di linee guida scozzesi.
  - *Sistema nazionale per le linee guida* (SNLG) (https://www.iss.it/snlg-consultazione): linee guida in italiano, sito dell'*Istituto Superiore di Sanità* (ISS). Sono di buona qualità, ma coprono un numero limitato di argomenti.

Un discorso a parte meritano i motori di ricerca. Ne esistono moltissimi, ma riportiamo i tre che secondo noi sono i più utili in ambito biomedico:

- GOOGLE (http://www.google.it/): è il principale motore di ricerca; indicizza articoli tramite MEDLINE, altre banche dati e anche direttamente dai siti degli editori. Risulta essere uno strumento utile soprattutto a fini diagnostici.
- GOOGLE SCHOLAR (http://scholar.google.it/): è il portale di Google dedicato alla ricerca in campo scientifico; indicizza molti dei contenuti di MEDLINE, delle librerie delle più prestigiose università e dei più importanti editori scientifici. Tra i suoi pregi vi è la ricerca anche tra libri e materiale audio-video. Non è però noto come indicizzi gli articoli né con quale frequenza venga aggiornato. Attualmente è possibile eseguire anche una ricerca avanzata, limitandola per autore, anno di pubblicazione, rivista, campo in cui cercare la parola chiave. Può essere molto utile per reperire i full text di alcuni articoli (compare un link di fianco all'articolo se presente in rete)
- TRIPDATABASE (http://www.tripdatabase.com/): è un motore di ricerca disegnato per fornire rapidamente risposte basate sulle evidenze. Ricerca articoli simultaneamente attraverso altri siti web che dovrebbero essere "evidence-based". Presenta risultati suddivisi in base al tipo di articolo trovato (ad esempio linee guida, revisioni sistematiche, sinossi basate sulle evidenze).

1.1.3 La strategia di ricerca della risposta

Con tutti questi siti e tutte le risorse disponibili, diventa a volte difficile capire dove e come effettuare la migliore ricerca delle informazioni per il nostro paziente.

Il nostro suggerimento è di impratichirsi con qualcuno di questi siti e di usare prevalentemente quelli che si conoscono meglio. Inizialmente è utile eseguire la stessa ricerca in differenti siti, in modo da verificare quale risponda meglio al tipo di quesito posto.

PubMed rimane comunque il database più comunemente utilizzato per qualsiasi tipo di ricerca in ambito medico.

1.2 PubMed

È l'interfaccia di MEDLINE, consultabile gratuitamente sul sito http://www.ncbi.nlm.nih.gov/pubmed; in MEDLINE sono indicizzati gli articoli pubblicati in migliaia di riviste di ambito medico di tutto il mondo, ad opera della National Library of Medicine. Prima dell'avvento della "rete", gli articoli erano indicizzati su cartaceo, il cosiddetto "Index Medicus", inventato dal mitico John Shaw Billings nel 1879.

PubMed è in continua evoluzione, quindi, di seguito daremo informazioni generali per eseguire una ricerca. Queste informazioni dovrebbero essere valide indipendentemente dalle future modifiche dell'interfaccia grafica di PubMed. Per questo motivo, abbiamo deciso di dare consigli sulla strategia di ricerca in generale senza addentrarci nella grafica di PubMed che potrebbe cambiare nel tempo.

1.2.1 Come eseguire una ricerca partendo da un quesito clinico

Il primo passaggio per eseguire la ricerca bibliografica su PubMed abbiamo detto essere quello di avere ben chiara la domanda da porci, utilizzando l'approccio suggerito dall'acronimo PICOS.

Il secondo passaggio sarà quello di costruire in PubMed la nostra strategia di ricerca. Qualcuno potrebbe dire che creare una buona ricerca bibliografica è come cucinare: per cucinare dobbiamo valutare ogni ingrediente singolarmente e solo in seguito li dobbiamo unire sapientemente. Così anche nella ricerca bibliografica in PubMed è fondamentale ricercare un termine alla volta e solo successivamente unire i vari termini.

Quando inseriamo un termine nella stringa di ricerca, PubMed lo ricerca automaticamente come text word (cioè testo semplice) e come MeSH term. I MeSH term sono parole chiave standardizzate riconosciute da PubMed e utilizzate per indicizzare gli articoli. Questa funzione è fondamentale perché differenzia PubMed da altri motori di ricerca in cui non c'è un'indicizzazione così accurata. Per fare un esempio, se vogliamo cercare uno studio di accuratezza diagnostica e scriviamo "diagnosis" nella nostra stringa di ricerca, PubMed identificherà anche tutti gli articoli che abbiano scritto nell'abstract o nel titolo "sensitivity" o "specificity". Questo permette alla nostra ricerca di essere molto più sensibile!

Qual è però il rovescio della medaglia? Il tempo che può passare dalla pubblicazione dell'articolo all'indicizzazione in PubMed può essere elevato, anche di mesi. Se inseriamo nella nostra stringa di ricerca la parola ricercandola solo come termine MeSH potremmo quindi perdere tutti gli articoli che non sono stati ancora indicizzati. Questi articoli, invece, verranno identificati tramite l'utilizzo della parola semplice (per esempio "diagnosis", ma nell'articolo dovrà essere presente esattamente il termine "diagnosis").

Una volta ricercato ogni termine singolarmente, è utile controllare come PubMed abbia ricercato il termine, se abbia quindi identificato un MeSH term e se il MeSH term sia davvero coerente con la nostra ricerca. Per questo, possiamo andare nella ricerca avanzata o nella history e vedere la stringa di ricerca tramite i "Details". In questo modo vedremo come PubMed abbia ricercato il termine e se ha riconosciuto un MeSH term specifico.

Se vogliamo, possiamo anche effettuare una ricerca solo nel database dei MeSH term per vedere il significato del MeSH term identificato, gli "entry

Terms" (cioè i termini che PubMed riconosce sinonimi della nostra parola chiave) e la posizione nella quale il nostro termine è situato. Si possono infine selezionare anche dei subheadings, cioè delle sottocategorie del termine (per esempio "therapy" o "diagnosis"). In questo caso, PubMed cercherà solo gli articoli che sono stati riconosciuti come sottocategoria del termine chiave.

Una volta ricercata ogni parola singolarmente e valutato come PubMed ha eseguito la ricerca, è importante unire le parole tra loro con il termine AND o con il termine OR. Questa azione si potrà fare nella history, dove la nostra ricerca viene salvata per alcune ore. Possiamo anche decidere di ricercare dei sinonimi o eseguire diverse strategie di ricerca per capire quale possa dare migliori risultati.

Se vogliamo essere più specifici, possiamo inserire dei limiti alla ricerca (per esempio legati al tipo di studio, alla rivista all'anno di pubblicazione). Dobbiamo essere consapevoli che ogni limite che inseriamo è come usare un MeSH term, quindi, non ci farà vedere gli articoli più recenti non ancora indicizzati. Per questo è buona norma usare i limiti solo tardivamente.

A volte potrebbe essere utile ricercare un singolo articolo, per esempio, potremmo ricordarci l'autore e l'anno di pubblicazione di uno studio o la rivista. In questo caso, potrebbe essere molto utile usare la griglia "Single citation matcher" che, di solito, si trova nella prima pagina e ci fornisce una griglia che semplifica nettamente la nostra ricerca.

Ci sarebbero molte altre cose da dire, ma il rischio è di perdersi nel mare di PubMed senza riuscire a surfare!

Il consiglio, allora, è di guardare i tutorial e di provare a navigare in PubMed. Infine, ci si può iscrivere gratuitamente e chiedere a PubMed di inviare alla nostra casella di posta i risultati della ricerca che vogliamo con cadenza temporale definita (per esempio, ottenere ogni settimana una ricerca con stringa "sincope"). Ciò permette di essere sempre aggiornati sulle ultime pubblicazioni.

1.2.2 Eseguiamo una ricerca in PubMed

A questo punto, siamo pronti per provare ad eseguire una ricerca utilizzando PubMed.

GMP è un signore di mezza età, obeso, che fuma circa 40 sigarette al giorno, ha una BPCO nota con frequenti riacutizzazioni ed è alcolista attivo. Giunge in Pronto Soccorso per una grave insufficienza respiratoria. È la nostra prima guardia in Pronto Soccorso ed è il primo paziente critico che dobbiamo gestire in prima persona. Siamo in un piccolo ospedale di provincia e, praticamente, sono gli infermieri a dirci cosa fare, applicando il protocollo standard del luogo: ossigeno a bassi flussi, cortisone endovena, aerosol con broncodilatatori e magnesio ev. Tornati a casa, ci chiediamo se sia stato proprio corretto usare il magnesio in quel caso.

1.2.2.1 Dieci piccoli indiani... e non ne rimase nessuno

La strategia che vi proponiamo di seguire è quella messa in atto costantemente da EC, clinico di comprovata fama, ad ogni turno di guardia in Pronto Soccorso, per fare una ricerca bibliografica. Si ispira ad Agatha Christie, per non dimenticare nulla che possa essere d'aiuto ai pazienti.

Decimo indiano:

Identificare la domanda clinica (quesito background o foreground?).

In questo caso vogliamo sapere se, nel caso di GMP, paziente di mezza età, obeso, fumatore, con BPCO riacutizzata, il magnesio sia meglio del placebo. Il centro della nostra domanda è, quindi, il paziente e la domanda sarà di tipo foreground. Se avessimo voluto sapere la terapia della BPCO riacutizzata, saremmo stati di fronte, invece, ad un quesito di tipo background.

Nono indiano:

Identificare il tipo di studio che più facilmente può rispondere alla nostra domanda (Trial randomizzato controllato? Revisione sistematica con meta-analisi? Studio osservazionale? Linea guida? Revisione non sistematica o libro di testo?).

Se la domanda è di tipo foreground sarà più probabile che la nostra risposta possa essere reperita in uno studio originale, per esempio in un trial o in una revisione sistematica, piuttosto che in un libro di testo. Per quanto riguarda gli studi originali, dovremo tenere in mente la piramide dell'evidenza: se c'è almeno una revisione sistematica, probabilmente sarà un buon punto di partenza rispetto al singolo trial clinico. Nel caso del paziente GMP, possiamo pensare di individuare anzitutto le revisioni sistematiche.

Ottavo indiano:

Identificare la sorgente che più facilmente può dare una risposta (libro di testo, linee guida, PubMed, Google, Tripdatabase).

Una volta che abbiamo identificato il tipo di domanda e il tipo di studio che vogliamo cercare, dobbiamo chiederci dove sarà più facile cercare quello studio. Visto che vogliamo indagare le potenzialità di PubMed, proviamo a fare la nostra ricerca su questa banca dati.

Settimo indiano:

Scrivere in modo chiaro la nostra domanda, utilizzando l'acronimo **PICOS**:

Patient, **I**ntervention, **C**omparison, **O**utcome, **S**tudy. Questo ci aiuterà a scegliere le parole chiave appropriate.

Nel nostro caso, la domanda potrebbe essere formulata nel modo seguente: in un paziente con BPCO riacutizzata (P: COPD), l'utilizzo del magnesio (I: Magnesium) è meglio del placebo (C: placebo) per curarne la riacutizzazione (O: prognosi, mortalità, ospedalizzazione)?

Sesto indiano:

Iniziare la ricerca usando una parola chiave per volta.

Eseguire la ricerca di una singola parola per volta, come abbiamo già detto, ci permetterà di rifinire facilmente la nostra ricerca complessa in seguito. Potremmo subito accorgerci che un termine utilizzato non è il più adatto o che c'è un sinonimo da aggiungere alla ricerca. Anche se potrebbe sembrare che cercare una parola per volta sia una perdita di tempo, vi assicuriamo che, a lungo andare, vi accorgerete che così non è.

Quinto indiano:

Verificare nei details se i termini utilizzati sono utilizzati dal software in modo appropriato, così come intendevamo che fossero utilizzati.

Quando inseriamo una parola nella griglia di ricerca di PubMed, questa viene automaticamente ricercata come parola libera (text word) o tradotta automaticamente nel MeSH term più affine. Dato che la procedura di traduzione è automatica, è importante verificare sempre se la traduzione ci soddisfa o se il software ha "mis-interpretato" la nostra indicazione. Inserire nella griglia di ricerca direttamente il MeSH term potrebbe risultare controproducente, dato che intercorre sempre un intervallo di tempo tra la pubblicazione dell'articolo in MEDLINE e la sua indicizzazione con assegnazione dei termini MeSH. Rischieremmo così di perdere gli articoli più recenti. Inoltre, se la nostra parola non venisse riconosciuta come MeSH term, perderemmo anche tutti gli articoli inerenti alla nostra ricerca che non sono stati indicizzati con il termine da noi utilizzato, ma con sinonimi (l'acronimo COPD e/o i termini chronic obstructive pulmonary disease).

Quarto indiano:

Unire le parole chiave tramite la history.

A questo punto cliccando su "Advanced search" vedremo la nostra "Search history" e dovremo unire i termini della nostra ricerca con **OR** (presenza di un termine oppure di un altro), selezionando così l'insieme di tutti gli articoli

trovati che utilizzano, ad esempio, l'acronimo COPD e/o chronic obstructive pulmonary disease (l'operatore OR rende più sensibile la ricerca), oppure con **AND** (presenza contemporanea di due termini), selezionando l'insieme di articoli che contengono la parola COPD insieme alla parola magnesium (l'uso dell'operatore AND rende più specifica la ricerca). Per unire i termini nella "History" si può cliccare con il tasto sinistro del mouse sul numero che individua la ricerca elementare che ci interessa combinare.

Terzo indiano:

Valutare se sia più efficace utilizzare i filtri o le clinical queries.

Arrivati a questo punto, potremmo decidere che i risultati ottenuti con la nostra ricerca sono soddisfacenti oppure che il numero degli articoli trovati è troppo elevato per consentirci di passarli tutti in rassegna. Per questo può essere appropriato utilizzare i limiti o le clinical queries. Ad esempio, volendo trovare revisioni sistematiche sul magnesio, potremmo inserire come limiti "systematic review" o entrare in "Clinical queries" nella sottoclasse "Systematic review".

Secondo indiano:

Analizzare l'articolo trovato e leggere eventuali **riferimenti bibliografici** o **related articles.**

Non sempre appare subito ciò che ci aspettavamo di trovare: per alcuni termini potrebbe non esistere un MeSH term adeguato oppure potremmo trovarci di fronte ancora a troppi articoli. Possiamo quindi provare a seguire i consigli di PubMed che, a fianco degli articoli trovati, ci indica anche altre possibili ricerche da condurre, insieme agli articoli più richiesti sull'argomento e agli articoli in ordine di rilevanza pratica. A volte, invece, è più ragionevole ripiegare su una revisione, anche non sistematica, di cui controllare i riferimenti bibliografici che potrebbero suggerire singoli articoli particolarmente interessanti.

Ultimo indiano: omelia finale

Pro e contro PubMed.

PubMed è il database biomedico più completo, è gratuito, aggiornato e trasparente nella sua organizzazione e gestione, relativamente facile e friendly per quanto riguarda l'utilizzo. I suoi limiti sono la necessità di pratica per poterlo usare al meglio, se non si vuole rimanere impantanati nella miriade di articoli che vi può elencare, senza seguire alcun criterio di rilevanza pratica e/o scientifica. La presenza in chiaro della data di pubblicazione ci consente sempre di sapere

se quello che abbiamo trovato è il più recente articolo o linea guida uscita, cosa che Google, ad esempio, non permette di conoscere.

Punti chiave

Davanti ad un problema clinico il modo in cui vi suggeriamo di procedere è il seguente:

- ✓ identificare il tipo di domanda clinica (quesito background o foreground?);
- ✓ identificare il tipo di studio che più facilmente può rispondere alla domanda (RCT, revisione sistematica, studio di coorte?);
- ✓ identificare la sorgente che può dare una risposta (libro di testo, linee guida, PubMed, Google, Tripdatabase);
- ✓ scrivere la domanda utilizzando l'acronimo PICOS: Patient, Intervention, Comparison, Outcome, Study. Questo ci aiuterà a trovare le parole chiave che ci serviranno per la ricerca;
- ✓ iniziare la ricerca di una parola chiave per volta e verificare se i termini utilizzati sono riconosciuti come MeSH terms tramite i details;
- ✓ unire le parole chiave tramite la history;
- ✓ identificare se è più utile utilizzare i filtri o le clinical queries;
- ✓ analizzare l'articolo trovato e leggere eventuali riferimenti bibliografici o i related articles.

Bibliografia consigliata

Costantino G, Montano N, Casazza G. When should we change our clinical practice based on the results of a clinical study? Searching for evidence: PICOS and PubMed. *Intern Emerg Med.* 2015;10(4):525-7. doi: 10.1007/s11739-015-1225-5.

PubMed User Guide https://pubmed.ncbi.nlm.nih.gov/help/

2. Primo approccio alla lettura critica

2.1 La validità di uno studio

Siete di guardia in Pronto Soccorso e arriva una donna di 40 anni con una sospetta colica renale. Pensate di farle un'ecografia alla ricerca di idronefrosi o calcoli ma, nel frattempo, chiedete una consulenza al chirurgo di guardia. Il chirurgo vi dice di fare una TC dell'addome per escludere o confermare la presenza di un calcolo e quindi decidere la migliore strategia terapeutica per la paziente. Poiché, testardi, avete il dubbio che l'ecografia da sola sia sufficiente per escludere una calcolosi renale clinicamente rilevante e non sia necessario eseguire una TC in urgenza, decidete di fare una ricerca bibliografica per rispondere alla vostra domanda (ormai siete diventati bravissimi!) e cercate su PubMed:

- P: renal colic;
- I: ultrasonography;
- C: computed tomography;
- O: sensitivity.

Trovate subito un articolo del Canadian Journal of Emergency Medicine (*CJEM 2010;12:201-6*) in cui si documenta che, in presenza di ecografia normale, è molto poco probabile si debba ricorrere ad interventi successivi anche quando una più sofisticata procedura diagnostica riveli la presenza di calcoli non rilevabili all'ecografia stessa. A questo punto, come ci comportiamo? Diamo comunque retta al consulente (… d'altronde il chirurgo è lui)? Oppure, sulla base dell'evidenza fornita da quell'articolo, decidiamo di non fare la TC e di omaggiarlo dell'articolo?

Spesso è davvero facile trovare in letteratura risposte a domande che quotidianamente ci poniamo sul lavoro, tuttavia, per non commettere errori nel leggere un articolo scientifico, dobbiamo sempre farci almeno due domande: l'articolo è "credibile"? I risultati che riporta sono rilevanti per la mia pratica clinica?

La validità generale di uno studio si può scomporre in validità esterna e validità interna. Entrambe sono indispensabili affinché i risultati dello studio si possano applicare alla pratica clinica.

La validità esterna riguarda la possibilità di estendere i risultati dello studio dalla popolazione studiata dagli autori a quella generale con cui abbiamo a che fare tutti i giorni nella pratica clinica (e in cui è compreso il nostro paziente). Per valutarla, si tratta di capire se la popolazione selezionata dallo studio (in base ai criteri di eleggibilità e di esclusione stabiliti, al setting clinico nel quale è stato

condotto, alle modalità di reclutamento seguite), sia rappresentativa anche della popolazione alla quale intendiamo applicare i risultati dello studio stesso.

Mark Zimmerman (*Am J Psychiatry 2002; 159:469–473*) ha provato ad applicare i criteri di inclusione degli studi randomizzati effettuati sulle terapie per la depressione a pazienti ambulatoriali afferenti ad un servizio di cura della depressione, rilevando che meno del 20% dei pazienti "reali" sarebbe stato arruolato nei trials descritti dagli studi. Risultati analoghi sono stati riportati nell'ambito del trattamento di neoplasie e scompenso cardiaco. È quindi evidente che i risultati ottenuti da quei trials sono applicabili a meno del 20% della popolazione reale di interesse.

Per tornare alla paziente con la colica renale, dobbiamo valutare se il setting e i soggetti dello studio siano simili al caso che stiamo trattando. Scopriamo che lo studio è stato condotto reclutando pazienti consecutivi in un ambito (il Pronto Soccorso) e con un obiettivo (valutare se l'ecografia possa essere un test di screening efficace nei pazienti che si presentano con dolori da colica renale) che riflettono bene la domanda da cui siamo partiti. Anche le caratteristiche dei pazienti (età media, sesso, gravità, presenza di co-patologie) sembrano simili a quelle dei pazienti afferenti al nostro Pronto Soccorso: possiamo quindi concludere che lo studio ha per noi una buona validità esterna.

La validità interna riguarda, invece, la correttezza metodologica dello studio, ovvero il fatto che esso sia stato disegnato, condotto ed analizzato seguendo un metodo rigoroso e appropriato per rispondere alla domanda iniziale. Essa ha quindi a che fare soprattutto con la assenza di bias nel disegno e nella conduzione dello studio. I bias possono essere dovuti a moltissimi fattori, fra i quali ad esempio la procedura di selezione dei pazienti, i metodi di raccolta delle informazioni relative ai pazienti inclusi, le procedure di definizione ed aggiudicazione degli endpoint ed altri.

Nel nostro esempio, leggendo lo studio sull'ecografia nella colica renale, dobbiamo innanzitutto valutare qual è esattamente la domanda iniziale o quesito di ricerca al quale gli autori intendevano dare una risposta con il loro studio (l'ecografia può individuare pazienti con coliche renali "a basso rischio"?), il disegno dello studio utilizzato (si tratta di uno studio di tipo retrospettivo) e se il disegno è adeguato per rispondere alla domanda.

Per capire il concetto di bias, occorre tener presente che, per ogni studio che si conduce, quello che si recluta è un campione di pazienti appartenenti alla popolazione definita dai criteri di eleggibilità e di esclusione e dalle modalità di reclutamento (meglio se si tratta di pazienti consecutivi che soddisfano i criteri stabiliti). Questo fa sì che la stima che si ottiene sia sempre soggetta ad errore casuale (random), che influisce sulla precisione delle stime, del quale si tiene conto quando si esprimono i risultati utilizzando l'intervallo di confidenza. Ad esempio, come vedremo più avanti, la precisione dei risultati forniti da uno studio può essere aumentata incrementando la dimensione del campione:

è possibile ottenere un intervallo di confidenza piccolo a piacere, al costo di aumentare le risorse dedicate allo studio.

Diversamente dall'errore casuale, gli errori sistematici, i bias, sono errori che dipendono da qualche falla nello studio, nel disegno, nell'organizzazione, nelle modalità di selezione del campione, nelle modalità di rilevazione dei dati, nei modi di analizzare i risultati. L'errore sistematico, che si traduce in bias di stima, ha a che vedere con il concetto di inaccuratezza. Mentre l'errore casuale, che ha a che fare con l'imprecisione della stima, può essere ridotto reclutando grandi campioni, l'errore sistematico può essere ridotto solo individuando ed eliminando le fonti della distorsione.

I bias si producono in vari momenti di realizzazione di uno studio, non solo durante la selezione dei pazienti, ma anche al momento della raccolta delle informazioni, e si manifestano in modi e con effetti diversi a seconda che si tratti di studi di eziologia, di prognosi, di diagnosi o di efficacia di trattamento, di studi osservazionali o sperimentali.

Ancora prima di valutare la validità di uno studio, dovremmo chiederci se ciò che è stato misurato nello studio, cioè l'endpoint considerato, è clinicamente rilevante.

2.2 Gli endpoint

Vediamo ora come si possono classificare, in base a differenti criteri, gli endpoint di uno studio.

2.2.1 Endpoint primario e secondario

L'endpoint primario è l'obiettivo principale di uno studio: consiste nella variabile che è misurata in tutti i pazienti inclusi nello studio e che ci permette di dare una risposta al quesito di partenza. Tale variabile può essere quantitativa (ad esempio la riduzione dei valori pressori dopo trattamento con farmaco antipertensivo), oppure qualitativa, in genere dicotomica (ad esempio remissione o non remissione di malattia dopo chemioterapia). In base ad esso viene calcolata la numerosità del campione, che abbiamo visto essere un aspetto fondamentale del disegno di uno studio (vedi potenza dello studio). Proprio perché in base all'endpoint primario viene costruito lo studio, i suoi risultati sono probanti, e possono perciò essere utilizzati direttamente per modificare la pratica clinica. L'endpoint primario può, quindi, essere definito come un endpoint che è in grado di fornire evidenza circa l'effetto di un intervento. Ovviamente, oltre alla significatività statistica dei risultati, andrà valutata anche la rilevanza clinica dell'endpoint primario (vedi oltre, endpoint hard vs surrogati).

Sono inoltre numerosi gli studi che definiscono più di un endpoint primario. In generale, a nostro avviso, l'endpoint primario dovrebbe essere limitato a un

singolo endpoint semplice. In tal caso lo studio risulta più solido, con meno fattori confondenti, più facilmente verificabile per la sua significatività statistica e rilevanza clinica. Ogni studio dovrebbe rispondere ad una sola domanda.

L'endpoint secondario è un endpoint che fornisce ulteriori informazioni circa l'effetto di un intervento, ma che, preso individualmente, non è sufficiente a stabilirne con sicurezza l'efficacia. Infatti, dato il notevole impiego di risorse e gli elevati costi necessari per condurre uno studio, sarebbe dispendioso usare tutte le informazioni raccolte per verificare solo l'endpoint primario. Si possono quindi porre altre domande, oltre a quella principale, e valutare la risposta ottenuta nello studio, in modo da ricavare una descrizione più completa degli effetti di un intervento. L'endpoint secondario è quindi un outcome addizionale rispetto a quello primario. Poiché il disegno dello studio clinico e, in particolare, il calcolo della numerosità campionaria non sono definiti in base all'endpoint secondario, ma solo sull'endpoint primario, i risultati dell'endpoint secondario potrebbero non essere probanti, e dovrebbero essere utilizzati più per trarne indicazioni di ricerca che suggerimenti per la modifica della pratica clinica (ruolo esplorativo).

A titolo di esempio, si possono ricordare i risultati degli studi *ELITE (Lancet 1997;349:747-752)* ed *ELITE II (Lancet 2000;355:1582-1587),* che confrontavano captopril e losartan in pazienti anziani con scompenso cardiaco. Alla fine dello studio *ELITE I,* il cui endpoint primario era la "tollerabilità renale", l'endpoint secondario "mortalità" risultò significativamente minore nei pazienti trattati con losartan. Tuttavia, in uno studio disegnato ad hoc, l'*ELITE II*, con mortalità come endpoint primario, i due trattamenti risultarono sovrapponibili.

In conclusione, i risultati di un endpoint secondario non possono influenzare la pratica clinica, finché non vengono sottoposti a verifica in studi specificamente ed adeguatamente progettati; d'altra parte, la moltiplicazione dei test di significatività sui dati di uno studio fa sì che il grado di protezione nei confronti di un errore di primo tipo si riduca.

N.B. È importante sottolineare come sia l'endpoint primario sia il secondario debbano comunque essere definiti a priori. Endpoint non pianificati potrebbero infatti risultare statisticamente significativi perché si è andati a cercare una significatività, si è cioè proceduto a quello che viene indicato come "data dredging" o "data torturing" (è ben noto che se i dati vengono torturati dallo statistico, prima o poi parlano!).

2.2.2 Endpoint hard vs surrogati

In base alla rilevanza clinica, gli endpoint (sia primario che secondario) possono essere distinti in hard e surrogati.

Gli endpoint hard sono quelli direttamente rilevanti per la vita del paziente (ad esempio mortalità, inizio della dialisi, recidiva infartuale, occorrenza di

ictus). Sono quelli che hanno più rilevanza per la pratica clinica e dovrebbero di regola essere utilizzati negli studi, in particolare come endpoint primari.

Gli endpoint surrogati sono di supposto interesse per il paziente, in quanto presumibilmente correlati a un endpoint hard in base ad un razionale fisiopatologico, e normalmente sono costituiti da un antecedente misurabile di un outcome clinico (esempio, la riduzione della proteinuria è presumibilmente correlata alla riduzione del rischio di dialisi a 5 anni, la riduzione del colesterolo alla riduzione del rischio di ictus o infarto, la soppressione di un'aritmia misurata all'holter al rischio di morte improvvisa). Vengono utilizzati perché sono più facilmente e tempestivamente misurabili e/o di incidenza maggiore rispetto agli endpoint hard.

Tuttavia, non sempre la supposta correlazione tra endpoint surrogato e hard è dimostrata. In letteratura esistono esempi drammatici di come questa presunta correlazione tra endpoint surrogato e hard sia risultata fallace. Esemplare il caso dell'utilizzo della flecainide nel peri-infarto: vari studi dimostrano come la flecainide sopprima le aritmie ventricolari nel peri-infarto, d'altra parte è noto altresì che pazienti con aritmie ventricolari nel post-infarto siano a maggior rischio di morte. Tuttavia, lo studio che ha valutato specificatamente la mortalità utilizzando questo farmaco nel peri-infarto è stato interrotto precocemente per eccesso di mortalità nel gruppo di trattamento rispetto al placebo *(N Engl J Med 1991;324(12):781-8).* Occorre quindi un'estrema cautela nel modificare la pratica clinica sulla base di studi che utilizzano degli endpoint surrogati. Non a caso, nel nostro esempio iniziale l'endpoint primario era la mortalità per ogni causa, un endpoint decisamente "hard".

2.2.3 Endpoint semplice vs composito

L'endpoint primario e quello secondario possono essere semplici o compositi (anche detti combinati).

Si parla di endpoint semplice quando l'esito di interesse è costituito da un singolo evento (ad esempio ingresso in dialisi). Si ha invece l'endpoint composito quando l'outcome dello studio è misurato dal verificarsi di uno fra più eventi possibili (ad esempio decesso o incidenza di ictus fatale o non fatale).

Gli endpoint compositi vengono spesso utilizzati negli studi perché permettono di arruolare un minor numero di pazienti e/o di proseguire lo studio per un tempo inferiore. Infatti, in questo caso non si misura l'incidenza di un singolo evento, ma la somma delle incidenze di diversi eventi, che possono verificarsi in alternativa tra di loro, consentendo dunque una numerosità campionaria inferiore. Nei risultati degli studi che utilizzano endpoint combinati dovrebbe comunque essere possibile scorporare i diversi componenti dell'endpoint in modo da riuscire a capire quali siano significativi o se la significatività è raggiunta dalla combinazione di tanti esiti, nessuno dei quali significativo e tanto meno di per sé rilevante. Citiamo ad esempio il trial *ATHENA (NEJM 2009;360:668-678)*

che confronta dronedarone e placebo nei pazienti con fibrillazione atriale (FA); l'endpoint primario di questo studio è un composito di prima ospedalizzazione per eventi cardiovascolari e decesso. Se considerato complessivamente, tale endpoint risulta significativo a vantaggio del dronedarone rispetto a placebo (31.9% vs 39.4%, HR 0.76, IC 0.69-0.84, $p<0.001$), se però andiamo a scomporre il risultato (vedi tabella 2 dello studio) ci accorgiamo che la significatività statistica riguarda solo l'ospedalizzazione (recidiva di FA), mentre la mortalità (endpoint davvero rilevante) non risulta diversa nei due gruppi di trattamento.

Nei prossimi capitoli vedremo quali sono le caratteristiche che distinguono i vari tipi di studi.

Punti chiave

- ✓ Uno studio è ad elevata validità interna quando è stato disegnato, condotto ed analizzato con un approccio metodologico rigoroso e appropriato, senza introdurre distorsioni (bias).
- ✓ Uno studio è ad elevata validità esterna quando i suoi risultati si possono estendere dallo studio alla tipologia di pazienti con cui abbiamo a che fare tutti i giorni nella pratica clinica.
- ✓ Gli endpoint o indicatori di esito rappresentano eventi che possono essere misurati in maniera oggettiva per valutare l'obiettivo dello studio.
- ✓ L'endpoint primario è la misura dell'obiettivo principale di uno studio, in generale ogni studio dovrebbe avere un solo endpoint primario.
- ✓ Per essere in grado di modificare la pratica clinica, uno studio dovrebbe avere come endpoint primario un endpoint hard (cioè clinicamente rilevante), singolo e che raggiunga la significatività statistica, stimato con sufficiente precisione.
- ✓ I risultati di endpoint secondari dovrebbero essere utilizzati per generare ipotesi per studi successivi, e non per trarre conclusioni cliniche.

Bibliografia consigliata

Christensen R, Ciani O, Manyara AM, Taylor RS. Surrogate endpoints: a key concept in clinical epidemiology. *J Clin Epidemiol.* 2024 Mar;167:111242.

Costantino G, Montano N, Casazza G. When should we change our clinical practice based on the results of a clinical study? Study endpoints. *Intern Emerg Med.* 2015 Oct;10(7):875-7.

Fleming TR, De Mets DL. Surrogate End Points in Clinical Trials: Are We Being Misled? *Ann Intern Med.* 1996;125:605-13.

Freemantle N, Calvert M, Wood J, et al. Composite outcomes in randomized trials. Greater precision but with greater uncertainty. *JAMA*. 2003;289:2554-9.

Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002 Jan 19;359(9302):248-52.

Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. *Obstet Gynecol*. 2005 May;105(5 Pt 1):1114-8.

Montori VM, Permanyer-Miralda G, Ferreira-González I, et al. Validity of composite end points in clinical trials. *BMJ*. 2005;330:594–6.

Palileo-Villanueva LM, Dans AL. Composite endpoints. *J Clin Epidemiol*. 2020 Dec;128:157-158.

3. Primo approccio alla lettura dei risultati

Un buon punto di partenza per comprendere fino in fondo, ed interpretare in maniera corretta, i risultati di uno studio scientifico è la capacità di leggere ed interpretare i dati che sono riportati nelle varie tabelle degli studi, dalle medie, alle mediane e, soprattutto, alle percentuali. Iniziamo dalla Tabella 1, quella che solitamente in un articolo scientifico descrive le caratteristiche dei pazienti arruolati.

3.1 La "Tabella 1" di un articolo scientifico: descrizione dei soggetti arruolati

Myburgh et al (*N Engl J Med 2012;367:1901-11*) riportano nella Tabella 1 del loro studio che l'età media dei 3358 pazienti trattati è pari a 63.1 ± 17 anni. Cosa significa? Il primo numero, 63, è appunto la media che ci dà un'idea generale di quanto "anziano" è il campione selezionato. Il secondo numero, la deviazione standard, ± 17 anni, sta ad indicare la variabilità delle età dei pazienti arruolati: non tutti i 3358 pazienti, ovviamente, hanno 63.1 anni, essendo quello un valore medio. Ci saranno pazienti con un'età maggiore e pazienti con un'età minore di 63.1. La deviazione standard ci dice di quanto, mediamente, le età dei pazienti arruolati si scostano dalla media. Quanto più la deviazione standard assume un valore prossimo a 0, tanto più i pazienti avranno età fra loro simili. Quanto più il valore della deviazione standard è elevato, tanto più le età dei pazienti saranno fra loro differenti. Nella stessa tabella, gli autori riportano anche che il valore dell'APACHE II score è pari a 17.0 (12.0–22.0). In questo caso 17 non sta ad indicare la media, ma, come riportato in tabella, la mediana, mentre 12.0-22.0 indica il range interquartile (IQR). Una mediana di 17 significa che, prendendo tutti i valori dell'APACHE II score dei singoli pazienti e mettendoli in ordine (crescente o decrescente), 17 è il valore che sta al centro. Vale a dire, il 50% dei soggetti ha valore <17 ed il restante 50% di soggetti un valore >17. L'interpretazione del range interquartile è molto simile: 12 e 22 rappresentano rispettivamente il 1° ed il 3° quartile. Di conseguenza, per le definizioni di quartili, sappiamo che il 25% dei pazienti arruolati ha un valore di APACHE II score < 12, un altro 25% ha valori superiori a 22 mentre il restante 50% ha valori compresi fra 12 e 22. Di conseguenza il range interquartile comprende il 50% delle osservazioni. Il range interquartile è utilizzato come misura di variabilità: quanto più è ampio l'intervallo di valori che comprende il 50% dei valori dei pazienti tanto più si ha variabilità. Se il range interquartile fosse stato 5-30, è evidente che saremmo stati in presenza di una maggiore variabilità, pazienti con valori molto diversi fra loro. Se invece fosse stato 14-15 la variabilità sarebbe

stata ridotta: infatti il 50% dei soggetti avrebbe avuto un valore di score pari a 14 o 15.

Un punto che dobbiamo tenere bene presente è che, quando descriviamo le caratteristiche dei nostri pazienti, possiamo utilizzare, per le variabili quantitative, la media e la deviazione standard solo se i dati sono simmetrici. In caso di asimmetria, media e deviazione standard forniscono una descrizione distorta. In tali casi, sarebbe meglio riportare mediana e range interquartile. Senza entrare nei dettagli, il motivo è da ricercare nel fatto che media e deviazione standard funzionano bene quando associate alla distribuzione normale (o gaussiana) dei dati. Nelle altre situazioni funzionano un po' meno bene e quindi si ritiene opportuno utilizzare mediana ed IQR. Come possiamo fare per valutare se i dati sono o non sono con distribuzione gaussiana? Innanzitutto, possiamo utilizzare un metodo grafico. Rappresentando la distribuzione di frequenza con un grafico a barre si può vedere – spannometricamente – se si ha distribuzione gaussiana o meno. I due grafici sotto riportano la distribuzione delle concentrazioni di ALT e di glucosio nel sangue di 1400 donatori di sangue. Senza dubbio, semplicemente guardando i due grafici, possiamo valutare come non gaussiana la distribuzione dei valori di ALT e gaussiana quella dei valori di glucosio.

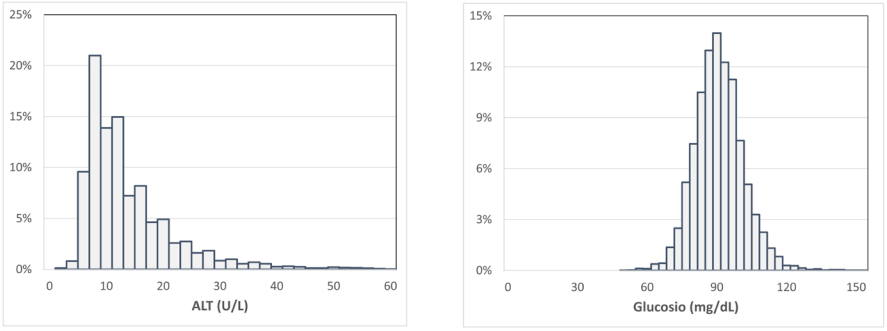

Figura 3.1. Distribuzione di frequenza dei valori di ALT e di glucosio in un campione di donatori sani

Un'altra regola che si può utilizzare consiste nel confrontare i valori di media e mediana: quando sono molto diversi fra loro, sicuramente la distribuzione sarà non gaussiana; quando invece sono simili, si potrebbe avere distribuzione gaussiana. In realtà, per una valutazione formale, esistono numerosi test statistici che si possono applicare nei vari contesti per valutare se la distibuzione dei dati è normale o approssima sufficientemente la distribuzione normale.

Abbamo visto come sono descritti i pazienti per le caratteristiche quantitative. Quando abbiamo a che fare con caratteristiche qualitative (sesso, copatologie,

familiarità etc) il dato descrittivo nella Tabella 1 di un articolo scientifico è riportato mediante conteggi e percentuali. Le percentuali generalmente non presentano problemi di interpretazione. Tuttavia, in alcune situazioni possono non essere di immediata interpretazione.

3.2 Oltre la media. Interpretare correttamente le percentuali

Nella tabella 3.1 sono riportati i risultati di uno studio condotto in Svezia (*Lancet Public Health 2021; 6: e729–38*) per valutare la prevalenza di disturbo depressivo nella popolazione generale, sul totale e separatamente per genere.

		Disordine depressivo		
		Sì	***No***	**Totale**
Genere	***Maschi***	195	2829	**3024**
	Femmine	300	2413	**2713**
	Totale	**495**	**5242**	**5737**

Tabella 3.1 Prevalenza di disturbo depressivo in un campione di 5737 individui provenienti dalla popolazione generale: frequenze assolute.

Le tabelle 3.2 e 3.3 riportano gli stessi dati, espressi però in forma percentuale (percentuali di colonna e di riga).

		Disordine depressivo		
		Sì	***No***	**Totale**
Genere	***Maschi***	6.4% (195/3024)	93.6% (2829/3024)	**100.0%**
	Femmine	11.1% (300/2713)	88.9% (2413/2713)	**100.0%**
	Totale	**8.6% (495/5737)**	**91.4% (5242/5737)**	**100.0%**

Tabella 3.2 Prevalenza di disturbo depressivo in un campione di 5737 individui provenienti dalla popolazione generale: percentuali di riga.

		Disordine depressivo		
		Sì	***No***	**Totale**
Genere	***Maschi***	39.4% (195/495)	54.0% (2829/5242)	**52.7%**
	Femmine	60.6% (300/495)	46.0% (2413/5242)	**47.3%**
	Totale	**100.0%**	**100.0%**	**100.0%**

Tabella 3.3 Prevalenza di disturbo depressivo in un campione di 5737 individui provenienti dalla popolazione generale: percentuali di colonna.

Come interpretiamo questi dati? Il punto di partenza, quando dobbiamo interpretare una percentuale, è sempre ricordarsi di considerare il denominatore, perché queste tipologie di percentuali sono di fatto frequenze relative, che ci indicano quanto è frequente il fenomeno analizzato in un gruppo standard composto da 100 persone. Così, ad esempio, 6.4%, ottenuto dal rapporto fra 195 (maschi depressi) e 3024 (totale dei maschi) ci dice quale è la percentuale di depressi fra i maschi, ovvero la prevalenza di depressione fra i maschi: in 100 maschi abbiamo 6.4 soggetti affetti da disordine depressivo. Analogamente, 11.1% (300/2713) è la prevalenza di depressione fra le femmine. Guardando la seconda tabella, vediamo invece che 60.6% (300/495) ci dice che la maggioranza dei depressi sono femmine: considerando il totale dei depressi osservati (495), il 60.6% di questi (300) sono femmine. Attenzione a trarre le conclusioni corrette quando interpretiamo queste percentuali. Come vedremo più avanti, negli studi di accuratezza diagnostica, per trarre le conclusioni corrette dovremo guardare le percentuali appropriate! Ad esempio, se volessimo parlare di rischio di depressione fra maschi e femmine, possiamo facilmente vedere come fra le nostre tabelle sia quella che riporta le percentuali di riga (ricordiamoci che sulle righe abbiamo maschi e femmine) a fornirci il risultato: su 100 maschi, 6.4 sono depressi; su 100 femmine, 11.1 sono depresse. Quindi, in base ai dati riportati, le femmine di quella popolazione sono a maggior rischio di depressione.

Riprendiamo la prima tabella e valutiamo i dati relativi alla frequenza di disordine depressivo da una differente prospettiva, proviamo a cambiare i denominatori. Abbiamo visto che le 300 femmine depresse corrispondono all'11.1% del totale delle femmine. Proviamo ora ad esprimere l'occorrenza di depressione rapportando il numero di femmine con depressione, non più al totale delle femmine (300/2713), ma al numero di femmine non affette da disordine depressivo. Otteniamo un numero, 0.124 (300/2413), che esprime lo stesso concetto visto sopra con la percentuale (frequenza relativa) in una metrica differente. Quel

numero, 0.124, ci dice sempre quanto è frequente la depressione fra le femmine, ma non è da interpretare come fosse una percentuale (perché non lo è!), ma come numero di depresse per ogni non depressa. Ciò significa che un risultato di quel rapporto pari a 0.124 ci dice che abbiamo 124 depresse ogni 1000 non depresse (ovvero 0.124 depresse per ogni non depressa). Interpretazione analoga può essere fatta per i maschi: 195/2829=0.069: fra i maschi, per 1000 non depressi abbiamo 69 depressi. Se, come abbiamo visto sopra, le percentuali sono di fatto interpretabili come rischi (probabilità), questi nuovi rapporti prendono il nome di odds, misure di frequenza che vedremo più avanti. Sostanzialmente, utilizziamo gli odds per quantificare lo stesso concetto che quantifichiamo con le probabilità (incertezza del verificarsi di un fenomeno), ma ricorrendo ad una metrica differente. Per il resto, metrica a parte, gli odds sono molto simili alle probabilità, ma possiedono proprietà matematiche che li rendono, se non indispensabili, molto utili in alcuni contesti, quali ad esempio gli studi di accuratezza diagnostica e gli studi caso-controllo.

Punti chiave

- ✓ La media non è sempre la misura migliore per sintetizzare variabili quantitative: in alcuni contesti è più appropriato utilizzare la mediana.
- ✓ Ricordarsi sempre di riportare anche una misura che quantifica la dispersione dei dati (deviazione standard o range interquartile)
- ✓ Le frequenze relative (le percentuali) sono necessarie per sintetizzare le variabili qualitative: vanno però interpretate correttamente (attenzione al denominatore!)

Bibliografia consigliata

Altman DG. Statistics in medical research. In: *Practical statistics for medical research*. Chapman and Hall: London, 1996.

Bland M. Statistica medica. Maggioli Editore. 2019.

Lang TA and Altman DG. Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines. In *Guidelines for Reporting Health Research: A User's Manual* (eds D. Moher, D.G. Altman, K.F. Schulz, I. Simera and E. Wager) 2014. https://doi.org/10.1002/9781118715598.ch25.

Larson MG. Descriptive statistics and graphical displays. *Circulation*. 2006 Jul 4;114(1):76-81.

4. I disegni degli studi

4.1 Piramide dell'evidenza

La ricerca in campo epidemiologico clinico può essere di tipo:
- osservazionale: l'esposizione al fattore di cui si intende valutare l'effetto non dipende dal ricercatore;
- sperimentale: l'esposizione al fattore di cui si intende valutare l'effetto dipende dal ricercatore.

In base alle modalità con cui selezioniamo i pazienti da includere nello studio, ed in base alle modalità di raccolta delle informazioni, possiamo distinguere gli studi in trasversali (anche detti cross-sectional) e longitudinali. Questi ultimi, a loro volta, sono distinti in retrospettivi (che vanno indietro nel tempo, esempio caso-controllo) e prospettici (che vanno in avanti nel tempo, esempio coorte e RCT).

Esistono, quindi, differenti tipi di studio che possono essere utilizzati per indagare la realtà di interesse clinico. Quando si decide di intraprendere una ricerca, ci si dovrebbe chiedere quale disegno di studio è più adatto a fornire la risposta, quello cioè in grado di produrre l'evidenza migliore. Nella figura è rappresentata la cosiddetta "piramide dell'evidenza" che mostra come, partendo dalla base, il livello o forza dell'evidenza cresce salendo verso il vertice e passando da un disegno all'altro. Fra gli altri fattori, il livello dell'evidenza dipende anche dalla qualità degli studi, vale a dire dalla capacità dei vari studi di fornire risultati non affetti da distorsioni di vario tipo. In generale, quando in uno studio si sono introdotte, involontariamente, distorsioni si dice che lo studio è affetto da bias. Come vedremo più avanti, esistono diverse tipologie di distorsioni, e gli studi osservazionali sono per loro natura tendenzialmente più esposti al rischio di bias.

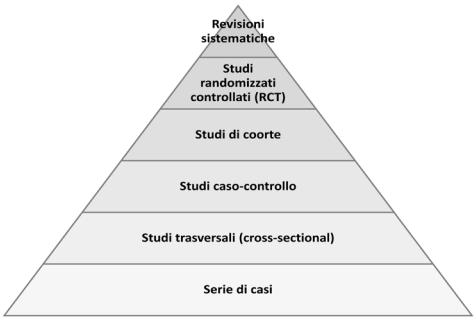

Figura 4.1 La piramide dell'evidenza

Come si può notare, la piramide dell'evidenza attribuisce elevato livello di evidenza agli studi randomizzati. Tuttavia, non sempre gli studi randomizzati forniscono la risposta migliore in termini di evidenza. Una modalità differente di rappresentazione dei diversi livelli di evidenza forniti dagli studi tiene conto della tipologia di studio (diagnosi, intervento, prognosi). Mediante il cosiddetto "trifoglio dell'evidenza", riportato in Figura 4.2, è possibile schematizzare il livello di evidenza, tenendo conto delle peculiarità delle differenti tipologie di studio. Così, ad esempio, con il trifoglio dell'evidenza è possibile mettere in luce il fatto che il disegno cross-sectional (come vedremo più avanti) fornisce evidenza di alto livello quando lo studio è diagnostico, mentre lo studio osservazionale prospettico fornisce risultati di elevata evidenza quando si vogliono studiare fattori prognostici.

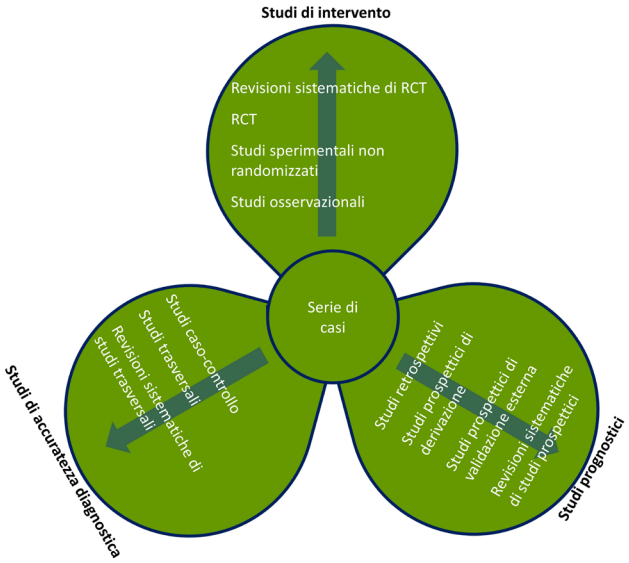

Figura 4.2 Il trifoglio dell'evidenza

4.2 Serie di casi

Un paziente di 67 anni affetto da fibrosi retroperitoneale e diabete mellito viene ricoverato nel vostro reparto per comparsa, da quattro settimane, di ittero colestatico senza segni di insufficienza epatica, per cui ha eseguito una TC addome con mezzo di contrasto e una colangio-RMN, risultate negative. Vi chiedete se possa esistere un nesso tra la diagnosi di fibrosi retroperitoneale e la successiva comparsa di diabete e colestasi. In PubMed trovate un articolo in cui, sulla base di una serie di casi, viene descritta l'associazione tra pancreatite

autoimmune (spesso complicata da comparsa di diabete) e patologie sistemiche fibrosanti, come la fibrosi retroperitoneale e una forma di colangite sclerosante responsiva alla terapia steroidea *(Best Pract Res Clin Gastroenterol. 2009;23:11-23*).

Alla base della piramide ci sono i resoconti di singoli casi o di una serie di casi. Consistono nella descrizione di un caso clinico insolito per qualche caratteristica relativa a eziologia, presentazione clinica, esito, o nella descrizione di caratteristiche cliniche osservate in una successione di pazienti accomunati da una stessa condizione. Tali pubblicazioni sono soprattutto utili a generare congetture o ipotesi che possono successivamente essere indagate con studi rigorosi. Possono anche aiutare il clinico a riconoscere associazioni non note o a trattare una patologia rara.

Spesso si potrebbe trascurare l'importanza di questo tipo di studi, ma in realtà possono essere molto utili, ad esempio, per identificare malattie nuove (i primi casi, nel 1981, di infezione da virus dell'immunodeficienza acquisita-HIV, si veda *MMWR Morb Mortal Wkly Rep. 1981;30:250-2*) oppure a scopo formativo o per fare diagnosi e pensare a trattamenti in malattie rare. Molte riviste ad esempio, riconoscendo l'importanza di questo tipo di articoli, hanno creato delle sotto riviste che pubblicano esclusivamente casi clinici.

4.3 Studi trasversali (cross-sectional studies)

4.3.1 Studi di prevalenza

Siete interessati a sapere qual è la prevalenza di microorganismi resistenti nei pazienti istituzionalizzati. Trovate un articolo in cui, per valutare tale prevalenza è stato fatto un tampone rettale di sorveglianza a tutti i pazienti ricoverati in 27 RSA del nord Italia. Dallo studio risulta che su 1947 pazienti, 991 (il 51%) era positivo per almeno un batterio con resistenza a più antibiotici (*Front Cell Infect Microbiol. 2023;13:1155320*).

4.3.1.1 Caratteristiche

Studi di questo genere, osservazionali, sono condotti reclutando un gruppo di pazienti selezionati in base a predefiniti criteri di inclusione-esclusione e procedendo, nello stesso momento, alla raccolta dell'informazione desiderata. Sono generalmente studi di prevalenza, trasversali in quanto non considerano l'evoluzione temporale del fenomeno in studio, elemento fondamentale per analizzare la relazione causale nello studio: esposizione-malattia, terapia-guarigione, trattamento-sopravvivenza. Negli studi trasversali tutta l'informazione è raccolta in un unico momento, attraverso quella che possiamo considerare una "fotografia" istantanea dei pazienti fatta al momento in cui si esegue lo studio.

4.3.1.2 Bias

I bias più diffusi nel caso degli studi di prevalenza sono relativi alla selezione della popolazione ed alla modalità di raccolta delle informazioni. Ad esempio, se si volesse valutare la prevalenza di infezioni opportunistiche sul totale delle infezioni e si eseguisse lo studio in un centro in cui è presente una casistica oncoematologica molto importante (che facilita la presenza di infezioni opportunistiche), il dato risulterebbe distorto.

4.3.2 Studi di accuratezza diagnostica

Un vostro paziente di 80 anni, ricoverato per accertamenti in merito alla comparsa di anemia, sviluppa febbre, tosse e dispnea. Sospettate una polmonite e un collega, appena tornato da un corso sull'ecografia toracica, si offre di eseguire un'ecografia al vostro paziente per verificare il sospetto diagnostico. Siete grati al vostro collega per la disponibilità, ma vi rendete conto di non sapere nulla sull'accuratezza dell'ecografia toracica nella diagnosi della polmonite e, pertanto, decidete di prendere tempo e correte al computer. Dopo una breve ricerca, trovate un articolo che sembra fare proprio al caso vostro. Nello studio, infatti, sono stati arruolati 49 soggetti con sospetta polmonite. Tutti i pazienti sono stati sottoposti ad ecografia polmonare e a radiografia del torace (*AJEM 2009;27:379–84*).

4.3.2.1 Caratteristiche

Il disegno di studio trasversale è quello ideale per gli studi di valutazione dell'accuratezza diagnostica di un test, nei quali i risultati di un nuovo test in studio (index test), che fornisce esito positivo o negativo, vengono confrontati con quelli di un test di riferimento (reference standard) che si suppone fornisca l'effettivo stato di realtà di ciascuno dei pazienti reclutati (malato/non malato), senza possibilità di errore. In questi disegni di studio, tutti i pazienti arruolati sono sottoposti contemporaneamente (da qui la trasversalità dello studio) ad index test e reference standard. In realtà, non è ovviamente possibile sottoporre un paziente contemporaneamente a due test, tuttavia l'intervallo di tempo che trascorre fra i due test deve essere di entità accettabile, tenuto conto del tipo di patologia indagata. In questo modo, è possibile stimare sensibilità e specificità del test in studio.

4.3.2.2 Bias

I bias più diffusi nel caso degli studi di accuratezza diagnostica hanno a che fare con la selezione dei pazienti e con la valutazione dell'esito dei test diagnostici. È infatti assodato che il disegno ideale per questa tipologia di studi è quello che consiste nell'arruolamento di pazienti consecutivi con il sospetto di malattia, che sono sottoposti contemporaneamente (nella pratica e breve distanza di tempo) all'index test ed al reference standard. Il venir meno della consecutività

dell'arruolamento (se non casuale) può infatti portare ad escludere inappropriatamente alcuni pazienti (ad esempio quelli più "difficili" da diagnosticare) e questo potrebbe introdurre un bias nelle stime di sensibilità e specificità, portando una sovrastima dell'accuratezza. Attenzione agli studi che selezionano due gruppi distinti di pazienti, gruppo di malati e gruppo di non malati: non sono studi trasversali, e venendo meno la consecutività, sono ad elevato rischio di bias. Per quanto riguarda, invece, la valutazione dell'index test e del reference standard, quando i due test non sono valutati in cieco l'uno rispetto all'altro, e quindi la conoscenza dell'esito di un test può influenzare l'esito dell'altro, si possono avere bias.

4.4 Studi osservazionali retrospettivi

Siete interessati a sapere quali fattori sono associati all'insorgenza di ictus in giovane età. Trovate uno studio in cui gli autori valutano se esista un'associazione tra fattori di rischio noti e ischemia cerebrale a insorgenza precoce. A tale scopo, hanno selezionato 961 pazienti tra i 25 e 49 anni con ischemia cerebrale e 1403 pazienti simili per età e sesso, ma senza ischemia cerebrale. Trovano che la presenza di fibrillazione atriale, malattie cardiovascolari, diabete mellito, colesterolo LDL, abitudine al fumo di sigaretta, ipertensione arteriosa e storia familiare di ictus sono i fattori associati alla presenza di ischemia cerebrale. (*J Am Heart Assoc. 2023 Jul 18;12:e028787).*

4.4.1 Caratteristiche

Gli studi caso-controllo sono i più noti fra gli studi retrospettivi, e sono condotti reclutando un gruppo di soggetti affetti dalla patologia di interesse (casi) e un gruppo di soggetti non affetti da quella patologia (controlli). In entrambi i gruppi si stima la frequenza di esposizione ad un dato fattore di rischio in studio. Mediante il confronto delle frequenze di esposizione nei casi e nei controlli, si cerca di stabilire se esiste un'associazione fra esposizione e malattia. Sono studi longitudinali in quanto, a differenza degli studi trasversali, i pazienti sono considerati per un appropriato intervallo di tempo (a seconda della supposta latenza fra esposizione e insorgenza clinica della malattia). Sono studi retrospettivi poiché valutano la relazione esposizione-malattia in senso inverso rispetto a quello naturale (si risale infatti dalla malattia all'esposizione, anziché discendere dalla esposizione verso l'insorgenza di malattia, come capita di fare negli studi di coorte). Questo tipo di studio fornisce una stima della forza di associazione fra malattia ed esposizione, basata sul calcolo del rapporto fra odds di esposizione nei casi e nei controlli (odds ratio, OR) come surrogato del rischio relativo, la cui stima diretta è possibile solo in uno studio prospettico o di incidenza (coorte).

4.4.2 Bias

Per quanto riguarda gli studi retrospettivi, i bias più diffusi hanno a che fare con la modalità di raccolta delle informazioni. Molte informazioni rilevanti, infatti, vengono raccolte mediante intervista dei diretti interessati a distanza di anni dal momento in cui si sono verificati i fatti di interesse. Questo può avere un effetto sulla qualità delle informazioni raccolte. Ad esempio, ammettiamo che venga sospettato, a distanza di anni, l'effetto teratogeno di un integratore utilizzato in gravidanza intervistando una popolazione di neo-mamme: è probabile che, a parità di esposizione reale, le mamme di bambini con malformazione riferiscano con maggior frequenza l'esposizione all'integratore assunto in gravidanza. Questo tipo di bias (noto anche come recall bias) potrebbe introdurre una sovrastima nella forza di associazione fra malattia ed esposizione. Un bias altrettanto importante per gli studi caso-controllo ha a che fare con la selezione dei controlli. È un bias di selezione, che si manifesta quando i controlli appartengono ad una popolazione diversa da quella che ha generato i casi. Ad esempio, se in uno studio sull'associazione fra tumore del polmone ed esposizione a fumo di sigaretta si arruolano controlli ospedalieri (perché è più semplice effettuare lo studio), si può avere un bias di selezione se la prevalenza di fumatori fra i controlli non rispecchia quella della popolazione generale da cui provengono i casi. Nel caso di una maggiore prevalenza di fumatori fra i ricoverati in un ospedale (fonte dei controlli del nostro studio) rispetto alla popolazione generale, questo bias porterebbe ad una sottostima dell'associazione fra fumo di sigaretta e tumore del polmone.

Infine, ci potrebbe essere un bias dovuto al confondimento. Questo si verifica quando un fattore, o una variabile, è associato sia all'esposizione che all'esito (casi vs controlli). Tipici confondenti sono l'età e il sesso, ma anche la gravità della malattia o le condizioni generali di un paziente al reclutamento. Ad esempio, uno studio caso-controllo ha mostrato associazione fra tumore al pancreas e consumo di caffè. Un fattore di confondimento potrebbe essere l'abitudine al fumo di sigaretta. Infatti, il fumo di sigaretta potrebbe essere associato sia al tumore al pancreas (fattore di rischio) sia al consumo di caffè (chi più fuma consuma più caffè). Se così fosse, potrebbe essere che l'effetto visto inizialmente in termini di associazione fra consumo di caffè e tumore possa essere in realtà dovuto all'effetto del fumo di sigaretta: il consumo di caffè non ha un effetto proprio, ma si trascina l'effetto del fumo di sigaretta (confondimento). Esistono tecniche statistiche per valutare l'effetto di potenziali confondenti. Nel nostro caso, basterebbe ad esempio valutare l'associazione fra caffè e tumore separatamente nei due sottogruppi costituiti da fumatori e da non fumatori. Una delle tecniche che permettono di ridurre l'effetto dei confondenti negli studi caso-controllo è il cosiddetto appaiamento (matching). Per ogni caso selezioniamo un controllo con le stesse caratteristiche in termini di potenziali confondenti (ad esempio, per ogni caso selezioniamo un controllo della stessa

fascia di età, dello stesso sesso, dello stesso status socio-economico… e così via). Tornando al nostro esempio, se l'obiettivo dello studio è valutare l'associazione fra consumo di caffè e rischio di tumore al pancreas, e se pensiamo che il fumo di sigaretta possa essere un confondente, per ogni caso fumatore selezioneremo un controllo fumatore, e per ogni caso non fumatore un controllo non fumatore, in modo da creare due gruppi (casi e controlli) bilanciati per percentuale di fumatori.

4.5 Studi osservazionali prospettici

Volete sapere se il consumo di alcool è associato alla mortalità a lungo termine nei pazienti che hanno avuto un infarto miocardico. Trovate un articolo in cui gli autori analizzano la relazione studiando una coorte di 4365 pazienti infartuati. Dividono i pazienti in quattro gruppi: non bevitori (che sono il gruppo di riferimento), bevitori lievi, moderati e forti bevitori. Nei 12 anni di follow-up, i bevitori lievi e moderati hanno un rischio di morte (lievemente) inferiore a quello dei non bevitori (*Am J Clin Nutr. 2022;115:633-642*).

4.5.1 Caratteristiche

Gli studi di coorte sono i più noti fra gli studi prospettici. Gli studi di coorte si caratterizzano per il reclutamento di un gruppo di soggetti esposti ad un dato fattore di rischio di interesse ed un gruppo di soggetti non esposti a tale fattore. I soggetti di entrambi i gruppi sono liberi da malattia al momento dell'inclusione e sono seguiti nel corso del tempo (studi longitudinali) al fine di rilevare l'insorgenza della patologia d'interesse. È, così, possibile valutare l'associazione fra esposizione e malattia rapportando l'incidenza di malattia negli esposti a quella dei non esposti. Fra quelli osservazionali, sono l'unico tipo di studio che permette una stima diretta di incidenza. Sono studi longitudinali in quanto i pazienti sono considerati per un appropriato intervallo di tempo (a seconda della supposta latenza fra esposizione e insorgenza clinica della malattia si stabilirà una lunghezza adeguata di tempo di follow-up). Sono studi prospettici in quanto indagano la relazione esposizione-malattia a partire dall'esposizione. In base alla modalità di raccolta dei dati, qualora lo studio sia effettuato utilizzando archivi ospedalieri o registri di patologia (che permettano di definire una passata esposizione e il momento di insorgenza della malattia), pur essendo studi di questo genere di fatto apparentemente retrospettivi (valutazioni fatte andando a ritroso nel tempo), si parlerà di coorte storica, e lo studio sarà comunque di tipo prospettico. In letteratura studi del genere sono a volte indicati come studi retrospettivi con dati raccolti prospetticamente. La misura di associazione che si stima negli studi di coorte è il rischio relativo (RR), definito come rapporto fra incidenze o fra rischi cumulativi in una determinata base temporale eguale per esposti e non esposti.

Possono rientrare in questa categoria anche tutti quegli studi osservazionali prospettici (esempio tipico: studi di fattori prognostici) che, pur non prevedendo l'arruolamento di una coorte di esposti e una coorte di non esposti, vanno a valutare l'associazione fra alcuni fattori e l'insorgenza di eventi futuri.

4.5.2 Bias

Per quanto riguarda gli studi prospettici, i bias più diffusi sono forse quelli che hanno a che fare con la modalità di selezione dei soggetti inclusi negli studi. Gli esposti ed i non esposti devono provenire dalla stessa popolazione, con l'unica differenza rappresentata appunto dall'essere esposti o meno al fattore di rischio di interesse.

Ci sono diversi esempi in letteratura di studi in cui si sono ottenuti risultati distorti. I primi studi che intendevano indagare il rischio di epilessia in bambini che presentavano precocemente episodi di convulsioni febbrili stimarono un rischio molto elevato di sviluppare epilessia nell'arco dell'infanzia. Questo allarme fu marcatamente ridimensionato dagli studi successivi, che riportavano un rischio molto basso. La differenza tra i primi e i secondi studi era insita nel fatto che i primi erano stati condotti in centri di eccellenza e di terzo livello, dove si recavano bambini con situazioni più complesse e quindi a maggior rischio di essere epilettici.

Un ulteriore bias, che si può manifestare negli studi prospettici che prevedono un periodo di follow-up più o meno lungo, è dovuto al fatto che si hanno pazienti che non completano lo studio, i cosiddetti persi al follow-up. Se il numero di pazienti persi è rilevante e se la perdita di questi pazienti non è casuale, ma sistematica, cioè legata a caratteristiche dei pazienti quali l'appartenenza ai diversi gruppi (esempio, esposti/non esposti) o il tipo di esito (favorevole/sfavorevole) si potrebbe essere in presenza di bias. Un modo per valutare l'effetto di questo possibile bias consiste nel fare un'analisi di sensibilità, cioè provare a fare le analisi considerando i persi al follow-up sia come se avessero riportato outcome positivo sia come se avessero riportato outcome negativo e confrontare i risultati così ottenuti. Se i risultati sono simili, l'aver perso i pazienti verosimilmente non ha introdotto bias.

Infine, anche in questi studi si può avere il bias da confondimento. Ad esempio, vogliamo condurre uno studio per valutare se una ridotta attività fisica sia associata all'incidenza di diabete di tipo 2. Selezioniamo un gruppo di soggetti che svolge attività fisica ed un gruppo di soggetti che non svolge attività fisica e valutiamo l'incidenza di diabete nei due gruppi. È noto che variabili quali peso corporeo, familiarità, età e sesso possono avere un effetto sull'outcome. Se tali variabili fossero, come è plausibile che sia, distribuite diversamente fra chi esercita o non esercita una attività fisica e non ne tenessimo conto nel disegno dello studio o nella analisi dei dati, potremmo stimare una forza di associazione fra attività fisica e insorgenza di diabete tipo 2 del tutto distorta. Inoltre, particolare

attenzione va data a distinguere quella che è una variabile di confondimento da quella che potrebbe essere una variabile esplicativa di effetto (variabile intermedia di effetto): ad esempio, il peso corporeo potrebbe essere ciò che spiega come l'attività fisica riduce, almeno in parte, l'incidenza di diabete di tipo 2. In tal caso non andrebbe trattata come variabile di confondimento, ma appunto come variabile esplicativa.

4.6 Studi sperimentali di intervento - Studi controllati randomizzati

Gli studi sperimentali sono studi che permettono la valutazione dell'efficacia di un intervento, che, in senso generale, potrebbe essere sia di prevenzione, che di diagnosi, che di cura su pazienti con definite caratteristiche al reclutamento. Si distinguono:

1. studi non controllati: studio del solo del gruppo di intervento, non includono un gruppo di controllo;
2. studi controllati: confronto fra un gruppo di intervento ed un gruppo di confronto (o di controllo) non sottoposto all'intervento in studio; l'assegnazione all'uno o all'altro gruppo può essere o meno randomizzata (vedi di seguito).

Noi vedremo in questo capitolo solo gli studi sperimentali controllati, in quanto gli studi non controllati, a causa dell'elevato rischio di bias, raramente permettono di prendere decisioni cliniche.

Siete di guardia in Pronto Soccorso e arriva un paziente di 65 anni con uno shock settico da infezione delle vie urinarie. Nonostante l'iniziale carico volemico, il paziente resta ipoteso e siete in dubbio se continuare il carico volemico o iniziare le amine. Il vostro collega "senior" dice che sicuramente le amine farebbero bene al paziente, decidete allora di documentarvi e trovate uno studio multicentrico in cui pazienti in condizione di shock settico sono stati randomizzati a ricevere amine precocemente oppure liquidi e amine solo in seconda battuta. 782 pazienti sono stati assegnati al gruppo "amine" e 781 al gruppo "liquidi". La mortalità ospedaliera entro i 90 giorni era simile nei due gruppi (14% vs 14.9%) (*N Engl J Med. 2023 Feb 9;388:499-510*).

4.6.1 Caratteristiche

Le sperimentazioni cliniche controllate, randomizzate (randomized controlled trial, RCT) e condotte in condizioni di cecità (vedi paragrafo sottostante) rappresentano il livello metodologico più elevato della ricerca clinica. Sono studi prospettici che permettono il confronto fra due (o più) gruppi di pazienti omogenei. Eventuali differenze di esito (ad esempio, guarigione, mortalità,

eventi avversi) fra i due gruppi di pazienti potranno così essere attribuite, con ragionevole certezza, esclusivamente all'effetto dell'intervento in studio. Il metodo utilizzato per garantire l'omogeneità tra i due gruppi di pazienti è la randomizzazione, vale a dire l'assegnazione dei soggetti all'uno o all'altro gruppo attraverso una procedura casuale (basata, ad esempio, sulla generazione di numeri casuali o random non conosciuta prima del reclutamento di ciascun paziente dallo sperimentatore).

L'intervento a cui sono sottoposti i pazienti del gruppo di controllo può essere un altro tipo di intervento, diverso da quello in studio (ad esempio, il farmaco tradizionale), oppure un placebo. Il placebo è un trattamento inattivo somministrato ai partecipanti del gruppo di controllo di un trial per bilanciare il possibile effetto di suggestione legato alla somministrazione di un trattamento, indipendentemente dalla sua reale efficacia, il cosiddetto "effetto placebo". La somministrazione del placebo al gruppo di controllo consente, quindi, una misura unbiased dell'efficacia del trattamento in studio.

Solo nel caso di studi mirati a valutare l'efficacia di un nuovo trattamento per patologie per cui non esistono trattamenti di riconosciuta efficacia, oppure quando il nuovo trattamento è pensato in aggiunta ad un trattamento esistente, ai pazienti del gruppo di controllo dovrebbe essere somministrato il placebo. Qualora, invece, dovessero esistere trattamenti standard di comprovata efficacia, ai pazienti del gruppo di controllo dovrebbe essere somministrato, per ragioni non solo etiche, il trattamento standard. Si è detto che l'omogeneità dei gruppi di pazienti arruolati si ottiene attraverso il procedimento della randomizzazione. Mediante la randomizzazione si avrà che il "paziente medio" (definito in termini di caratteristiche individuali note e non note: età, sesso, gravità della patologia, presenza di copatologie ed altro) del gruppo di trattamento sarà simile al paziente "medio" del gruppo di controllo. In questo modo, l'unica differenza sostanziale fra i due gruppi sarà rappresentata dal trattamento. Quindi, qualora dovessimo osservare una differenza di esito (nel nostro esempio, mortalità a 28 giorni) fra i due gruppi a confronto, potremo affermare con ragionevole certezza che il principale responsabile di questa differenza osservata è l'unico fattore che varia fra i due gruppi: il trattamento. Esistono diversi metodi, più o meno complessi, per randomizzare i pazienti. Per ridurre la possibilità di manipolazioni (volontarie o meno) nel processo di reclutamento/assegnazione occorre, comunque, che chi provvede alla generazione della sequenza dei numeri casuali sia indipendente da chi decide dell'eleggibilità e quindi del reclutamento dei singoli pazienti. Medici e pazienti dovrebbero essere tenuti all'oscuro della sequenza di assegnazione, almeno fino all'assegnazione del caso al gruppo di trattamento (allocation concealment): ogni possibilità di prevedere la sequenza di assegnazione potrebbe minare la validità dei risultati dello studio. Infatti, sapere prima a quale gruppo di trattamento il paziente verrà assegnato, potrebbe influire sulla decisione di arruolarlo e costituire un bias.

Effettuata correttamente la randomizzazione, si richiede poi che tutti i soggetti randomizzati siano analizzati nel gruppo al quale sono stati assegnati. Per vari motivi, infatti, può capitare che alcuni pazienti inclusi inizialmente nello studio, vengano esclusi prima dell'analisi dei dati. È importante distinguere fra le esclusioni che si verificano prima della randomizzazione, in base ai criteri di arruolamento definiti a priori, che assicurano comunque la validità interna dello studio, pur potendone alterare la generalizzabilità dei risultati (validità esterna), e le esclusioni che avvengono dopo randomizzazione, che alterano il confronto tra due trattamenti pregiudicando la validità interna dello studio stesso.

Un altro problema riguarda l'interruzione o il passaggio volontario dal gruppo di trattamento assegnato all'altro. L'ideale sarebbe che tutti i pazienti continuassero a sottoporsi al trattamento al quale sono stati randomizzati, poiché ogni migrazione da un gruppo all'altro di trattamento vanifica gli effetti di bilanciamento della randomizzazione. Pertanto, per ridurre il bias prodotto da tale migrazione, è stabilito di analizzare comunque i risultati anzitutto per gruppo di randomizzazione, applicando il principio dell'intention-to-treat . Con questo approccio gli esiti osservati vengono comunque conteggiati come occorsi entro il gruppo di inziale assegnazione random, senza considerare l'effettivo trattamento seguito dal singolo paziente dopo la randomizzazione.

L'esecuzione della sperimentazione in condizioni di cecità (blinding, masking) permette di limitare l'effetto di alcuni bias. Medici (investigators), pazienti (partecipants), valutatori della risposta del paziente al trattamento (assessors) devono essere tenuti all'oscuro dell'intervento assegnato al paziente, per non esserne influenzati nel proprio giudizio. La conoscenza dell'intervento ricevuto da parte dei pazienti può influenzare la risposta al trattamento e modificare la compliance. La non cecità degli investigatori che seguono il paziente potrebbe far sì che l'attitudine a favore o contro un dato intervento possa trasformarsi in un comportamento sistematicamente diverso, fino a decidere interventi ancillari o trattamenti supplementari diversi in un gruppo e nell'altro, ad incoraggiare o a scoraggiare il proseguimento dello studio, a influenzare la fiducia stessa del paziente nel trattamento che sta seguendo. La non cecità degli assessors potrebbe portare ad un'alterata definizione degli outcomes soggettivi (valutazioni cliniche, valutazione sulla base di scale, lettura di immagini). Se un assessor, che deve, ad esempio, stabilire mediante l'applicazione di una scala clinica se un paziente è guarito o meno, non è in cieco rispetto al trattamento assegnato ai pazienti, potrebbe essere portato con maggior facilità a giudicare come guarito un paziente del gruppo di trattamento rispetto ad un paziente del gruppo di controllo. Un'ulteriore cecità sempre più invocata è oggi anche quella dello statistico che elabora i dati: anche il data dredging, data torturing, data phishing possono rappresentare una pratica scorretta indotta dalla non cecità di chi analizza i dati.

Distinguiamo le seguenti tipologie di studi, in base alle caratteristiche della cecità:

- studio singolo-cieco: solo il paziente è tenuto all'oscuro dell'assegnazione;
- studio doppio-cieco: pazienti e medici che conducono la sperimentazione sono tenuti all'oscuro dell'assegnazione;
- studio triplo-cieco: pazienti, medici e tutti gli eventuali soggetti terzi chiamati a valutare la risposta al trattamento sono tenuti all'oscuro dell'assegnazione;
- studio quadruplo-cieco: cecità delle tre categorie e dello statistico che analizza i dati;
- studio… quintuplo-cieco: la scheda di assegnazione è stata smarrita e nessuno può più procedere all'unmasking del trattamento al quale è stato sottoposto il paziente!

4.6.2 Bias

Negli studi sperimentali, se ben disegnati e ben condotti, il rischio di bias può essere ridotto al minimo. I bias caratteristici di questa tipologia di studio sono quelli legati alle modalità di randomizzazione, alla mancanza di cecità ed alla gestione dei pazienti persi al follow-up (attrition bias). Abbiamo già visto nei paragrafi precedenti quali devono essere le caratteristiche di uno studio sperimentale ben fatto, e quindi esente da bias. Entriamo un po' più in dettaglio solo per il bias dovuto ai persi al follow-up. Può accadere che durante uno studio prospettico un numero rilevante di pazienti non giunga a termine dello studio, per vari motivi. Queste perdite di pazienti potrebbero introdurre un bias, soprattutto se la decisone presa dal paziente di uscire dallo studio è dovuta a motivi riconducibili al trattamento (ad esempio, ridotta compliance, presenza di eventi avversi). In situazioni del genere potremmo quindi trovarci con un certo numero di pazienti per i quali non conosciamo l'esito (endpoint), con evidenti possibili conseguenze in termini di distorsioni. Immaginiamo, ad esempio, che tutti i pazienti trattati persi al follow-up non si presentano alle visite di controllo perché deceduti: l'esclusione di questi pazienti dalle analisi introduce un evidente bias.

Per quanto riguarda il confondimento, una randomizzazione ben fatta riduce di molto il rischio di questo tipo di bias.

4.6.3 Equipoise

Uno studio clinico sperimentale, in cui si effettua il confronto fra un nuovo trattamento ed un trattamento standard, ha senso solo se, nel momento in cui lo si pianifica, si è incerti sul fatto che il nuovo trattamento sia più efficace dello standard. Infatti, se si avessero evidenze sufficientemente robuste sull'efficacia del nuovo trattamento, non sarebbe sensato condurre un RCT per effettuare il confronto: abbiamo già la risposta, non dobbiamo ricercarla con nuovi studi. È allora evidente che l'assenza di questa situazione di equipoise (vale a dire avere

evidenza di una migliore efficacia di uno dei due trattamenti) rende non etica la conduzione dello studio, in quanto esporremmo, coscientemente, la metà dei soggetti arruolati ad un trattamento che sappiamo fin da subito essere meno efficace. In conclusione, la presenza di questa incertezza (vale a dire l'assenza di evidenza) rende etico uno studio clinico sperimentale randomizzato. È importante segnalare che "evidenza" non significa la presenza di trial randomizzati controllati. Per alcune domande, anche "l'esperienza" è sufficiente. Non abbiamo necessità di fare uno studio randomizzato controllato per confermare che il fuoco brucia, bastano le evidenze empiriche!

Punti chiave

- ✓ La piramide dell'evidenza è un espediente grafico che ci permette di visualizzare differenti tipi di studio in base alla qualità dell'evidenza che possono fornire.
- ✓ Il trifoglio dell'evidenza mette in luce, a differenza della piramide, il fatto che per ogni tipologia di studio (diagnosi, prognosi, intervento) ci può essere una gerarchia qualitativa differente fra gli studi.
- ✓ Serie di casi: utili soprattutto per segnalare patologie rare o per evidenziare nuove patologie/epidemie.
- ✓ Studi trasversali (o cross-sectional studies): studi particolarmente utili per studiare la prevalenza di una malattia o per valutare l'accuratezza diagnostica di un dato test.
- ✓ Studi osservazionali retrospettivi: i pazienti sono arruolati in base al loro status di malattia (casi e controlli) e si indaga la pregressa esposizione ad un dato fattore di rischio. Sono meno complessi da organizzare e condurre rispetto agli studi prospettici, ma sono a maggior rischio di bias.
- ✓ Studi osservazionali prospettici: i pazienti arruolati sono seguiti nel tempo (periodo di follow-up) per valutare l'insorgenza di eventi di interesse. La natura prospettica di questi studi fa sì che, se ben fatti, permettano di ridurre il rischio di bias, al costo di una relativamente elevata complessità di disegno e conduzione. Alcuni esempi tipici sono gli studi epidemiologici di coorte e gli studi clinici prognostici. Permettono la stima dell'incidenza.
- ✓ Studi randomizzati e controllati: riguardano interventi o terapie o, più raramente, il confronto di strategie diagnostiche; confrontano gruppi di pazienti assegnati in modo casuale e possono essere condotti con gradi diversi di cecità.

Bibliografia consigliata

Bowers D, House A, Owens D. *Understanding clinical papers*. Ed. John Wiley and Sons Ltd, 2001.

Costantino G, Montano N, Casazza G. When should we change our clinical practice based on the results of a clinical study? The hierarchy of evidence. *Intern Emerg Med.* 2015;10(6):745-7.

Costantino G, Montano N, Casazza G. When should we change our clinical practice based on the results of a clinical study? Diagnostic accuracy studies I: the study design. *Intern Emerg Med.* 2015 Dec;10(8):1025-7.

Furuya-Kanamori L, Xu C, Hasan SS, Doi SA. Quality versus Risk-of-Bias assessment in clinical research. *J Clin Epidemiol.* 2021 Jan;129:172-175.

Greenhalgh T. How to read a paper: papers that report diagnostic or screening tests. *BMJ.* 1997;315:540-543.

Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. *Lancet.* 2002;359:145-149.

Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. *Lancet.* 2002;359:341-345.

Schulz KF, Grimes DA. Case-control studies: research in reverse. *Lancet.* 2002;359:431-434.

Kotz D, West R. Key concepts in clinical epidemiology: addressing and reporting sources of bias in randomized controlled trials. *J Clin Epidemiol.* 2022 Mar;143:197-201.

Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet.* 2002;359:515-519.

Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet.* 2002;359:614-618.

Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet.* 2002;359:696-700.

Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet.* 2002 Mar 2;359(9308):781-5.

5. Lettura dei risultati

5.1 Le misure di accuratezza diagnostica

Ritorniamo al nostro paziente di 80 anni, ricoverato per accertamenti in merito alla comparsa di anemia, con febbre, tosse e dispnea. Avevamo sospettato una polmonite e un collega ci consigliava l'esecuzione di ecografia polmonare per diagnosticare la polmonite. Analizziamo in modo più approfondito l'articolo che avevamo trovato in precedenza (*AJEM 2009;27:379–84*).

5.1.1 Sensibilità e specificità

Come abbiamo visto in precedenza, una volta noti, per ogni paziente, la positività/negatività all'index test e lo status di malattia (malato/non malato, ottenuto dal reference standard), è possibile valutare l'accuratezza diagnostica del test stimandone sensibilità e specificità, valore predittivo positivo e negativo, rapporto di verosimiglianza positivo e negativo.

Il test diagnostico ideale è quello che permette di identificare tutti i sani come quei soggetti in cui il risultato del test è negativo (tutti i negativi al test sono non malati e tutti i non malati sono negativi al test) e tutti i malati come quei soggetti in cui il test ha fornito esito positivo (tutti i malati sono positivi al test, tutti i positivi al test sono malati). Purtroppo, nel mondo reale, tutti i test presentano risultati sia "falsi positivi", non malati positivi al test, sia "falsi negativi", malati negativi al test.

Per rappresentare la relazione tra risultato del test e status del paziente si è soliti costruire una Tabella 2x2, inserendo sulle righe l'esito dell'index test (positivo, negativo) e sulle colonne lo status reale di malattia (esito reference standard: malato, non malato).

		Malati	**Non Malati**	
Index test	+	Veri positivi (VP)	Falsi positivi (FP)	Totale positivi al test
	-	Falsi negativi (FN)	Veri negativi (VN)	Totale negativi al test
		Totale malati	Totale non malati	Totale pazienti

Tabella 5.1 La tabella 2x2 di accuratezza diagnostica.

In tabella sono contenute tutte le informazioni necessarie per stimare l'accuratezza del test.

Utilizzando i dati pubblicati nell'articolo che abbiamo trovato, possiamo costruire la seguente tabella 2x2 per valutare l'accuratezza dell'ecografia polmonare (Eco) nella diagnosi di polmonite, utilizzando come reference standard la radiografia del torace (Rx).

	Polmonite		
	Sì (Rx +)	No (Rx –)	Totale
Eco +	23 VP	8 FP	31 Positivi
Eco -	1 FN	17 VN	18 Negativi
Totale	24 Malati	25 Non malati	49

Tabella 5.2 Accuratezza dell'ecografia polmonare (Eco) nella diagnosi di polmonite.

Si definisce sensibilità (Sn) di un test la percentuale, fra i malati, di coloro che risultano positivi al test:

$$Sn = \frac{VP}{VP+FN} = \frac{23}{24} = 0.958$$ (positivi al test fra i malati, ovvero i veri positivi/tutti i malati)

Un test sarà tanto più sensibile quanto più il numero dei malati con test positivo (VP) si avvicinerà al numero totale dei malati (VP+FN). La sensibilità esprime la capacità del test di identificare, come positivi, i soggetti che presentano la malattia. Un test ad elevata sensibilità risulterà positivo in quasi tutti i pazienti in cui è presente la malattia e per questo, nel caso risulti negativo, sarà molto utile per eliminare il sospetto di malattia (se i FN sono tendenzialmente pari a 0, un negativo al test non potrà che essere un VN e quindi un non malato). Un test molto sensibile alla presenza di embolia polmonare (ma non solo) è il D-dimero, la cui negatività permette di escludere una embolia polmonare acuta in atto.

Da questo deriva l'acronimo SnNOut per indicare un test a elevata sensibilità (Sn) che in caso di risultato negativo (N) permette di escludere (rule Out) la presenza di malattia.

Si definisce specificità (SP) di un test la percentuale, fra i non malati, di coloro che risultano negativi al test:

$$Sp = \frac{VN}{VN+FP} = \frac{17}{17+8} = 0.680$$ (negativi al test fra i non malati, ovvero i veri negativi/tutti i non malati)

Un test sarà tanto più specifico quanto più il numero di non malati con test negativo (VN) si avvicinerà al numero totale dei non malati (VN+FP). La specificità esprime la capacità del test di identificare, come negativi, i soggetti senza la malattia. Un test ad elevata specificità risulterà negativo in quasi tutti i non malati, e per questo nel caso risulti positivo, sarà molto utile per confermare il

sospetto di malattia (se i FP sono tendenzialmente pari a 0, un positivo al test non potrà che essere un VP e quindi un malato). Ad esempio, un test molto specifico per la presenza di carcinoma vescicale (anche se non molto sensibile) è la citologia urinaria, la cui positività permette di confermare la presenza di carcinoma vescicale. Da questo deriva l'acronimo SpPIn per indicare un test a elevata specificità (Sp) che in caso di risultato positivo (P) permette di confermare (rule In) la presenza di malattia.

Riassumendo, un test ad elevata sensibilità (con specificità media) sarà utile per il paziente in caso di esito negativo. Un test ad elevata specificità (con sensibilità media) sarà utile per il paziente in caso di esito positivo. In realtà come vedremo più avanti, l'informazione più affidabile sull'utilità ai fini decisionali di un test diagnostico è fornita dai rapporti di verosimiglianza.

Per stimare sensibilità e specificità la tabella 2X2 va letta in verticale, per colonne.

Sensibilità e specificità sono caratteristiche del test diagnostico ed il loro valore è compreso tra 0 e 1 (o tra 0% e 100% se espresse in percentuale).

5.1.2 Valori predittivi

Nella pratica clinica, le domande alle quali il medico vorrebbe poter dare una risposta non sono tanto relative alla probabilità di risultare positivo al test di un malato (o di un non malato) quanto:

- "se un paziente risulta positivo al test, quanto è probabile sia davvero malato?";
- "se un paziente è negativo al test, quanto è probabile che sia veramente non malato?".
- In altre parole, tornando al nostro paziente con sospetta polmonite, ci interessa sapere: "Se sottoponendolo ad ecografia del torace avessimo un risultato positivo, quanto sarebbe probabile che il nostro paziente sia davvero affetto da polmonite? E se l'eco fosse negativa quanto sarebbe probabile che il nostro paziente sia davvero non affetto da polmonite?". La risposta a queste domande non può ovviamente essere fornita da sensibilità e specificità, che, come abbiamo visto, considerano la probabilità di positività/negatività all'eco in malati/non malati, invece che la probabilità di malattia (n positivi/negativi all'eco).

Il valore predittivo (positivo o negativo) è la probabilità del test di darci la diagnosi corretta:

- Valore Predittivo Positivo (VPP) è la probabilità che un soggetto positivo al test sia malato:

$$VPP = \frac{VP}{VP+FP} = \frac{23}{23+8} = 0.742$$

(malati positivi al test o veri positivi/ tutti i positivi)

Valore Predittivo Negativo (VPN) è la probabilità che un soggetto negativo al test sia non malato:

$$VPN = \frac{VN}{VN+FN} = \frac{17}{17+1} = 0.944$$

(non malati negativi al test o veri negativi/ tutti i negativi)

Per calcolare i valori predittivi dalla tabella 2X2 dobbiamo leggerla in orizzontale, per righe.

Nel nostro esempio, la probabilità che un soggetto risultato positivo all'ecografia sia malato (VPP) è del 74.2%; mentre la probabilità che un soggetto risultato negativo all'ecografia sia non malato (VPN) è del 94.4%.

N.B. Mentre sensibilità e specificità, come abbiamo detto, sono caratteristiche intrinseche del test (perché dipendono esclusivamente dall'abilità del test a risultare positivo/negativo nei malati/non malati), i valori predittivi dipendono anche dalla prevalenza della malattia fra i soggetti inclusi nello studio (Malati/ totale).

Per capire il perché, facciamo due esempi numerici e proviamo a calcolare i valori di sensibilità, specificità, VPP e VPN.

	Malati		
	Sì	No	Totale
Test +	1209 (VP)	141 (FP)	1350
Test -	741 (FN)	27909 (VN)	28650
Totale	1950	28050	30000

Tabella 5.3. Valori predittivi con prevalenza pari al 6.5%.

In base a quanto riportato nella tabella 5.3 otteniamo i seguenti risultati:

- totale soggetti 30000
- prevalenza di malattia = malati/totale = 1950/30000 = 6.5%
- Sn = VP/(VP+FN) = 1209/(1209 + 741) = 62.0%
- Sp = VN/VN+ FP) = 27909/(27909 + 141) = 99.5%
- VPP = VP/(VP + FP) = 1209/(1209 + 141) = 89.6%
- VPN = VN/(VN + FN) = 27909/(27909 + 741) = 97.4%

	Malati		
	Sì	No	Totale
Test +	186 (VP)	149 (FP)	335
Test -	114 (FN)	29551 (VN)	29665
Totale	300	29700	30000

Tabella 5.4. Valori predittivi con prevalenza pari all'1%.

In base a quanto riportato nella tabella 5.4 otteniamo i seguenti risultati:
- totale soggetti 30000
- prevalenza di malattia = malati/totale = 300/30000 = 1%
- Sn = VP/(VP+ FN) = 186/(186 + 114) = 62.0%
- Sp = VN/(VN + FP) = 29551/(29551 + 149) = 99.5%
- VPP = VP/(VP + FP) = 186/(186 + 149) = 55.5%
- VPN = VN/(VN + FN) = 29551/(29551 + 114) = 99.6%

Come possiamo vedere da questi esempi, al variare della prevalenza e a parità di Sn e Sp, cambiano i valori predittivi. In particolare, all'aumentare della prevalenza, il VPP aumenta mentre il VPN diminuisce, e viceversa. La prevalenza di malattia nella popolazione da cui proviene uno specifico paziente può essere interpretata come probabilità pre-test (o probabilità a priori) di quel particolare paziente, probabilità di malattia che, in base a conoscenza ed esperienza, attribuiamo al nostro paziente prima di avere eseguito il test, di cui conosciamo sensibilità e specificità.

Data la dipendenza dei valori predittivi dalla prevalenza, ne consegue che le stime di VPP e VPN ricavate da una popolazione diversa da quella da cui proviene il nostro paziente (ad esempio quella da cui si è reclutato il campione che ha prodotto lo studio pubblicato) non sono applicabili al nostro caso.

Poiché nella nostra pratica clinica i pazienti a cui applichiamo il test sono diversi da quelli selezionati negli studi di accuratezza diagnostica, per stimare la probabilità di malattia post-test del nostro paziente occorre stabilire la sua probabilità pre-test (vedremo in seguito come farlo) e utilizzare le stime di Sn e Sp nella formula di Bayes o, meglio, utilizzare i rapporti di verosimiglianza del test (positivo e negativo).

5.1.3 Rapporti di verosimiglianza

La capacità discriminante del test positivo può anche essere espressa come la probabilità che il test risulti positivo nei malati, rispetto a quella che risulti positivo nei non malati. È evidente che un test perfetto risulterà sempre positivo nei malati e mai nei non malati, e quindi questo rapporto, detto rapporto di verosimiglianza

(RV) o, in inglese, likelihood ratio (LR) del test positivo (LR+) tenderà all'infinito. Esprimere da un punto di vista operativo il concetto appena descritto significa rapportare la Sensibilità (che tiene conto dei veri positivi) a 1- Specificità (che tiene conto dei falsi positivi). Cosa succede se il test che stiamo valutando è del tutto inutile? la probabilità che sia positivo nei malati sarà identica a quella che sia positivo nei non malati, ciò significa che il LR sarà uguale a 1. In sostanza, al netto dei denominatori (numero di malati e numero di non malati), possiamo vedere LR+ come il rapporto fra i veri positivi ed i falsi positivi: quindi, quanto più numerosi sono i VP rispetto ai FP, tanto più è probabile che un risultato positivo del test sarà un VP piuttosto che un FP e tanto grande è LR+. Quindi, in presenza di un test con esito positivo, se LR+ è elevato è (molto) più probabile che il paziente sia uno dei malati (VP), piuttosto che uno dei non malati (FP). Questo significa che per test con elevato valore di LR+ un risultato del test positivo sarà (fortemente) indicativo di presenza di malattia:

$$LR+ = \frac{\frac{VP}{MALATI}}{\frac{FP}{NON\ MALATI}} = \frac{\frac{VP}{VP + FN}}{\frac{FP}{FP + VN}} = \frac{\text{Sensibilità}}{\text{1-Specificità}}$$

Per esprimere la capacità discriminante del test negativo possiamo rapportare la probabilità che il test risulti negativo nei malati, rispetto a quella che risulti negativo nei non malati. In questo caso, è evidente che un test perfetto risulterà mai negativo nei malati e sempre nei non malati, e quindi questo rapporto, detto rapporto di verosimiglianza del test negativo (LR-) tenderà a 0. Analogamente a quanto visto sopra, esprimere da un punto di vista operativo il concetto appena descritto significa rapportare 1- Sensibilità (che tiene conto dei falsi negativi) alla Specificità (che tiene conto dei veri negativi). Cosa succede se il test che stiamo valutando è del tutto inutile? la probabilità che sia negativo nei malati sarà identica a quella che sia negativo nei non malati, ciò significa ancora una volta che il LR sarà uguale a 1. Analogamente a quanto fatto con LR+, possiamo vedere LR-, sempre al netto dei denominatori (numero di malati e numero di non malati), come il rapporto fra i falsi negativi ed i veri negativi: quindi, quanto più numerosi sono i VN rispetto ai FN, tanto meno è probabile che un risultato negativo del test sarà un FN piuttosto che un VN e tanto più piccolo sarà LR-. Quindi, in presenza di un test con esito negativo, se LR- è basso è (molto) più probabile che il paziente sia uno dei non malati (VN) piuttosto che uno dei malati (FN). Questo significa che per test con bassi valori di LR- un risultato del test negativo sarà (fortemente) indicativo di assenza di malattia:

$$LR- = \frac{\frac{FN}{MALATI}}{\frac{VN}{NON\ MALATI}} = \frac{\frac{FN}{VP+FN}}{\frac{VN}{FP+VN}} = \frac{\text{1-Sensibilità}}{\text{Specificità}}$$

Da notare che LR- (così come anche LR+) non può assumere valori negativi, il valore minimo è zero.

In pratica, i rapporti di verosimiglianza indicano quanto è più probabile che il test risulti positivo nei malati rispetto ai non malati (LR+), oppure negativo nei malati rispetto ai non malati (LR-). Un buon test sarà quindi caratterizzato da un elevato valore di LR+ (comunque >1 solitamente si dice almeno >10) e da un basso valore di LR- (comunque <1 e solitamente si dice almeno < 0.10). In realtà, i valori di LR+ e di LR- da ritenere utili nella pratica dipendono sempre dal contesto clinico in cui stiamo operando. Come vedremo più avanti, i LR si possono utilizzare per stimare le probabilità post test di malattia in un paziente che, partendo da una certa probabilità pre-test, ha avuto un test con esito positivo (LR+) o negativo (LR-).

Ritornando al nostro esempio, possiamo allora calcolare i rapporti di verosimiglianza positivo e negativo dell'ecografia polmonare per la diagnosi di polmonite:

$$LR+ = \frac{\frac{VP}{MALATI}}{\frac{FP}{NONMALATI}} = \frac{\frac{VP}{VP+FN}}{\frac{FP}{FP+VN}} = \frac{\frac{23}{24}}{\frac{8}{25}} = 2.99 \quad \text{O, più semplicemente} \quad LR+ = \frac{SE}{1-SP} = \frac{0.958}{1-0.680} = 2.99$$

$$LR- = \frac{\frac{FN}{MALATI}}{\frac{VN}{NONMALATI}} = \frac{\frac{FN}{VP+FN}}{\frac{VN}{FP+VN}} = \frac{\frac{1}{24}}{\frac{17}{25}} = 0.06 \quad \text{O, più semplicemente} \quad LR- = \frac{1-SE}{SP} = \frac{1-0.958}{0.680} = 0.06$$

È quindi 2.99 volte più probabile che un'ecografia polmonare risulti positiva in un paziente con polmonite che in uno senza polmonite. È invece 0.06 volte più probabile (quindi è meno probabile) che l'ecografia polmonare risulti negativa in un paziente con polmonite che in uno senza polmonite: detto altrimenti è 16.7 volte (1/0.06=16.7) più probabile che l'ecografia polmonare risulti negativa in un paziente senza polmonite che in uno con polmonite.

I rapporti di verosimiglianza, essendo rapporti fra due probabilità, possono variare fra 0 e +∞. Solitamente, se il test ha una qualche utilità, LR- varia fra 0 ed 1 mentre LR+ fra 1 e +∞ il positivo: a livello interpretativo si possono anche considerare come dei rischi relativi.

Riassumendo quanto detto sino ad ora, l'utilità clinica di un test diagnostico è in gran parte determinata dall'accuratezza con cui esso identifica una patologia. Tale accuratezza può essere misurata da sensibilità e specificità, dai valori predittivi positivo e negativo, dai rapporti di verosimiglianza.

I valori predittivi, come abbiamo appena visto, dipendono dalla prevalenza della malattia (o della condizione che il test va a valutare); questo fa sì che le stime dei valori predittivi di uno studio non siano quasi mai applicabili al nostro paziente. I rapporti di verosimiglianza, invece, dipendendo solo da sensibilità e

specificità, non dipendono dalla prevalenza di malattia e possono essere utilizzati per calcolare, a partire dalla probabilità pre-test, la probabilità post-test di malattia in ogni particolare paziente. In precedenza abbiamo visto che è possibile utilizzare come stima della probabilità pre-test per un singolo paziente la prevalenza di malattia nella popolazione da cui proviene quel paziente. In realtà, come vedremo più avanti, la prevalenza è solo uno dei metodi per stimare la probabilità pre-test in un singolo paziente. Esistono infatti in letteratura vari algoritmi (sottoforma di score, punteggi a cui associamo un valore di probabilità) che permettono di stimare probabilità di malattia a partire da alcune predefinite caratteristiche individuali (anagrafiche e cliniche). Riprendiamo ora il nostro esempio e riportiamo di seguito tutte le stime di accuratezza diagnostica calcolate sino ad ora:

- Sn=23/24=0.958 Sp=17/25=0.680
- VPP=23/31=0.742 VPN=17/18=0.944
- LR+=0.958/0.320=2.99 LR-=(1-0.958)/0.68=0.06

Ora vediamo come possiamo ottenere la probabilità post-test a partire dalla probabilità pre-test (stimata dalla prevalenza) utilizzando i rapporti di verosimiglianza.

Un primo metodo, abbastanza immediato, consiste nell'utilizzare uno strumento grafico: il nomogramma di Fagan (Figura 5.1). Partendo dall'asse sulla sinistra, che mostra la probabilità pre-test di malattia, si tracci la linea che congiunge la probabilità pre-test del paziente con il valore del rapporto di verosimiglianza del test, indicato sull'asse centrale. Proseguendo tale linea fino ad intersecare l'asse delle probabilità post-test del nomogramma (asse di destra) si leggerà direttamente il valore di quest'ultima probabilità per quello specifico paziente.

Nel nostro esempio, assumendo come stima della probabilità a priori la prevalenza di polmonite osservata nel campione di pazienti consecutivi selezionato al 49% (24/49), otteniamo una probabilità a posteriori del 75% circa in caso di eco positivo (linea tratteggiata) e del 5% circa in caso di eco negativo (linea continua).

In realtà, per stimare le probabilità post-test non è indispensabile utilizzare questo strumento grafico, ma si possono trovare online app molto affidabili che fanno tutti i calcoli necessari.

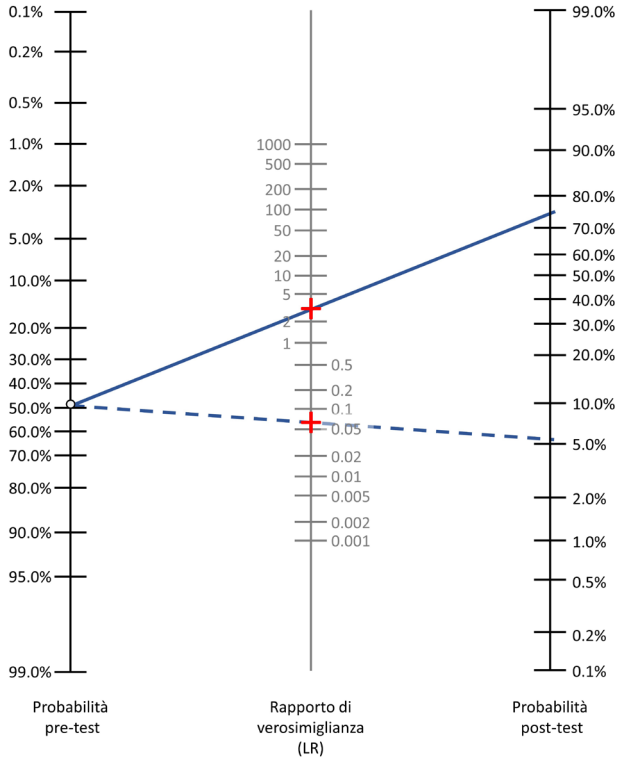

Figura 5.1. Nomogramma di Fagan.

Un ulteriore metodo, riservato ai più nerd fra voi, consiste nel ricorso alla formula che, trasformata la probabilità pre-test in odds, permette di calcolare il valore di odds post-test e quindi ritornare alla probabilità (questa volta post-test):

odds post-test= odds pre-test x LR

Si chiama odds il rapporto fra una probabilità ed il suo complemento ad 1:

$$odds = \frac{p}{(1-p)}$$

Dato un valore di odds si può ricavare la probabilità corrispondente con la formula sotto riportata:

$$p = \frac{odds}{(1+odds)}$$

Per il momento, accontentiamoci di considerare questi passaggi come se fossero una sorta di black box, fidiamoci della correttezza formale e limitiamoci ad interpretare il risultato finale. Il significato matematico degli odds sarà approfondito nei paragrafi successivi (vedi 5.4.1 e Appendice).

Proviamo a chiarire con due esempi numerici.

Esempio 1: paziente con una probabilità pre-test di polmonite del 49%

Probabilità pre-test: $p_{pre\text{-}test} = 0.49$	Odds pre-test $= \frac{p_{pre\text{-}test}}{(1-p_{pre\text{-}test})} = \frac{0.49}{0.51} = 0.96$
Test positivo: LR+=2.99	
Odds post-test = odds pre-test x LR+ = 0.96 x 2.99 = 2.87	
Probabilità post-test: $\frac{odds_{post\text{-}test}}{(1+odds_{post\text{-}test})} = \frac{2.87}{3.87} = 0.74$	
Test negativo: LR-=0.06	
Odds post-test = odds pre-test x LR- = 0.96 x 0.06 = 0.058	
Probabilità post-test: $\frac{odds_{post\text{-}test}}{(1+odds_{post\text{-}test})} = \frac{0.058}{1.058} = 0.054$	

Esempio 2: paziente con una probabilità pre-test di polmonite del 25%

Probabilità pre-test: pre-test = 0.25	Odds pre-test $\frac{p_{pre\text{-}test}}{(1-p_{pre\text{-}test})} = \frac{0.25}{0.75} = 0.33$
Test positivo: LR+=2.99	
Odds post-test = odds pre-test x LR+ = 0.33 x 2.99 = 0.99	
Probabilità post-test: $\frac{odds_{post\text{-}test}}{(1+odds_{post\text{-}test})} = \frac{0.99}{1.99} = 0.50$	
Test negativo: LR-=0.06	
Odds post-test = odds pre-test x LR- = 0.33 x 0.06 = 0.020	
Probabilità post-test: $\frac{odds_{post\text{-}test}}{(1+odds_{post\text{-}test})} = \frac{0.020}{1.020} = 0.020$	

I rapporti di verosimiglianza indicano, quindi, di quanto il risultato di un test (positivo o negativo) aumenti o diminuisca la probabilità pre-test di malattia del paziente specifico. Un rapporto di verosimiglianza uguale a 1 lascia invariata tale probabilità e quindi è assolutamente inutile utilizzare tale test; un rapporto maggiore di 1 aumenta la probabilità di malattia pre-test; un rapporto di verosimiglianza minore di 1 riduce la probabilità pre-test.

Una valutazione approssimativa dell'utilità clinica di un test può essere data dalla seguente griglia di lettura:

Utilità clinica	LR+	LR-
– Rilevante	> 10	< 0.1
– Moderata	5-10	0.1-0.2
– Modesta	2-5	0.5-0.2
– Trascurabile	1-2	0.5-1

È interessante notare come il rapporto di verosimiglianza fornisca sul test diagnostico informazioni molto più dirette, da un punto di vista clinico, di quanto facciano sensibilità e specificità prese isolatamente.

Analizzando i risultati riportati nell'Esempio 1 possiamo fare alcune interessanti considerazioni. Avrete sicuramente notato come la probabilità post-test in caso di test positivo (74%) coincide con il valore predittivo del test positivo (VPP), mentre la probabilità post-test in caso di test negativo (6%) coincide con il complemento ad 1 del valore predittivo del test negativo (1-VPN). Non è un caso: questo accade perché, nell'Esempio 1, abbiamo utilizzato come probabilità pre-test la prevalenza di polmonite dello studio. Quindi, in tutti i casi in cui la probabilità pre-test di un paziente coincide con la prevalenza di malattia nello studio che ha effettuato la valutazione, possiamo utilizzare il VPP ed il VPN forniti dallo studio quali stime delle probabilità post-test di malattia.

Infine, ritorniamo per un momento a quanto detto a proposito di test ad elevata sensibilità. Sappiamo che per escludere una diagnosi è necessario un test con una elevata sensibilità (SnNOut). Supponiamo, a titolo esemplificativo, di avere a disposizione i due test diagnostici la cui accuratezza è riportata nella Figura 5.2.

Test 1

	Malati		
	Sì	**No**	**Totale**
Test +	95	95	**190**
Test -	5	5	**10**
Totale	**100**	**100**	**200**

Sn	***0.95***
Sp	***0.05***
VPP	***0.50***
VPN	***0.50***
LR+	***1***
LR-	***1***

Test 2

	Malati		
	Sì	**No**	**Totale**
Test +	80	40	**190**
Test -	20	60	**10**
Totale	**100**	**100**	**200**

Sn	***0.80***
Sp	***0.60***
VPP	***0.70***
VPN	***0.80***
LR+	***2***
LR-	***0.3***

Figura 5.2. Due test per la diagnosi della stessa patologia. Effetto di sensibilità (Sn) e specificità (Sp) sui valori dei rapporti di verosimiglianza (LR+ e LR-).

Il primo test ha una sensibilità del 95% e una specificità del 5%, il secondo una sensibilità dell'80% e una specificità del 60%. In base alla "regola" SnNOut, saremmo portati a dire che il test da utilizzare per escludere la malattia dovrebbe essere il primo (sensibilità 95%). Calcolando il rapporto di verosimiglianza negativo, vediamo che vale 1, quindi il test è inutile per escludere la malattia, non modificando la probabilità di malattia del paziente. Per inciso, lo stesso test è inutile anche per confermare la malattia (LR+=1). Considerando il secondo test, vediamo che il rapporto di verosimiglianza negativo risulta 0.3, non eccellente, ma sicuramente migliore per escludere la malattia rispetto al primo test. Questo dimostra come non sempre ad una maggiore sensibilità corrisponda un test migliore, utile ad escludere una malattia: infatti, in questo esempio, il primo test, pur avendo una sensibilità ottimale, non dà alcuna informazione, dato che è egualmente probabile per malati e non malati risultare negativi (e anche positivi) al test, il quale quindi non contribuisce a cambiare la probabilità pre-test di malattia di qualsiasi paziente.

Abituiamoci allora a considerare l'accuratezza diagnostica di un test, soprattutto per la decisione clinica, in termini di rapporti di verosimiglianza, piuttosto che di sensibilità e specificità.

In conclusione, una volta che si disponga del valore di probabilità post-test, una domanda nasce spontanea: che cosa ce ne facciamo? Ovviamente, dipende dal livello di probabilità raggiunto e da altre considerazioni relative alla malattia e alla sua prognosi, al trattamento e al suo profilo di efficacia/sicurezza. 74% (o 50%) è un valore di probabilità sufficientemente elevato per prendere una decisione sul trattamento del paziente?

5.4% (o 2%) è un valore di probabilità sufficientemente basso per decidere di rassicurare il paziente ed escludere ogni trattamento?

Cercheremo di dare una risposta a queste domande nel capitolo relativo alle soglie decisionali.

Punti chiave – Misure di accuratezza diagnostica

- ✓ Nella pratica clinica, per poter scegliere il test più adatto a rispondere al nostro quesito diagnostico ed evitare di eseguire test scarsamente informativi se non addirittura inutili, dobbiamo conoscere le misure di accuratezza del test.
- ✓ La sensibilità di un test indica la percentuale di malati che risultano positivi al test, ovvero la capacità del test di identificare come positivi i soggetti che presentano la malattia. Un test ad elevata sensibilità che risulti negativo serve ad escludere il sospetto clinico (Test SnNOut).
- ✓ La specificità di un test indica la percentuale di non malati che risultano negativi al test, ovvero la capacità del test di identificare, come negativi, i soggetti non malati. Un test ad elevata specificità che risulti positivo serve a confermare il sospetto clinico (Test SpPIn).

- ✓ Il valore predittivo positivo di un test (VPP) indica la probabilità che un soggetto positivo al test sia malato (coincide con la probabilità di malattia post test positivo).
- ✓ Il valore predittivo negativo di un test (VPN) indica la probabilità che un soggetto negativo al test sia non malato (è pari al complemento a uno della probabilità di malattia post test negativo).
- ✓ Mentre sensibilità e specificità sono caratteristiche intrinseche del test, i valori predittivi dipendono anche dalla prevalenza della malattia, pertanto, a differenza delle stime di sensibilità e specificità, le stime di VPP e VPN fornite da uno studio particolare sono generalmente di nessuna utilità, non essendo applicabili a pazienti che provengono da popolazioni con prevalenze diverse.
- ✓ Il rapporto di verosimiglianza (LR) indica quanto è più probabile che il test risulti positivo nei malati rispetto ai non malati (LR+), oppure quanto è più probabile che il test risulti negativo nei malati rispetto ai non malati (LR-). È più informativo, da un punto di vista clinico, dei valori predittivi, poiché è caratteristico del test e non si modifica con il variare della prevalenza di malattia. È più informativo anche della sensibilità e specificità valutate isolatamente, poiché integra in un'unica misura la valutazione di capacità discriminante del test, permettendo così la valutazione di rilevanza decisionale dell'informazione che un test può fornire nel caso di un particolare paziente.
- ✓ LR+ e LR-, applicati alla probabilità pre-test di malattia di un singolo paziente, ci permettono di stimare la probabilità post-test di malattia.

5.2 Quando il test diagnostico fornisce come esito una variabile quantitativa

Un uomo di 76 anni, diabetico, iperteso e affetto da insufficienza renale cronica, arriva in Pronto Soccorso per dispnea acuta e ingravescente da circa 3 ore. All'esame obiettivo rilevati lievi edemi declivi, crepitii ai campi inferiori, saturazione di O2 88% in aria, frequenza respiratoria 30 atti/min, frequenza cardiaca ritmica a 110/minuto, pressione arteriosa 170/100 mmHg; all'ECG, tachicardia sinusale con anomalie diffuse aspecifiche della ripolarizzazione ventricolare. All'ecografia del torace, presenza di linee B su tutti i campi polmonari, come da sindrome interstiziale. Agli esami ematici, creatinina 1.8 mg/dl (valore abituale del paziente), troponina 197 ng/l (il limite per il laboratorio del vostro ospedale è 40 ng/l). Vi hanno detto che il dosaggio della troponina ha una sensibilità dell'89% e una specificità del 92% per la diagnosi di ischemia miocardica.

Come interpretate il dato della troponina, anche alla luce delle co-patologie del paziente e del quadro clinico? Il paziente è solo scompensato o ha anche un infarto miocardico? E se aveste rilevato un valore di troponina di 200 o di 400ng/l, come sarebbe cambiato il vostro giudizio? Come utilizzate le stime di sensibilità e specificità che vi sono state fornite?

Parlando di sensibilità e specificità di un test basato su di una variabile quantitativa, è indispensabile fissare il valore di cut-off in base al quale si giudica la positività del test. Cambiando cut-off, infatti, la performance del test in termini di sensibilità e specificità cambia. Quando diciamo che il test della troponina per la diagnosi di infarto ha una sensibilità dell'89% e una specificità del 92%, facciamo riferimento ad un determinato valore di cut-off.

Per semplificare, consideriamo la classica situazione in cui abbiamo un marker (un valore quantitativo misurato sui pazienti: troponina, glicemia, emoglobina, pressione arteriosa sistolica, colesterolo) i cui valori elevati siano associati a presenza di malattia (i malati hanno valori mediamente più elevati dei non malati). È buona norma giudicare la positività o meno del test avendo stabilito a priori il cut-off da usare, ma in realtà non esiste un cut-off ideale da utilizzare sempre: se vogliamo disporre di un test a elevata specificità il cut-off deve essere più alto (maggior numero di veri negativi, minor numero di falsi positivi) di quello che dobbiamo adottare per disporre di un test a elevata sensibilità (maggior numero di veri positivi, minor numero di falsi negativi). Quindi, come abbiamo già visto, agendo sul valore del cut-off siamo in grado di variare la sensibilità e la specificità di un test diagnostico. Purtroppo, però, all'aumentare della sensibilità diminuisce la specificità, e viceversa; non è possibile variare il cut-off in modo da incrementare contemporaneamente sensibilità e specificità. Possiamo facilmente vedere che all'aumentare del valore di cut-off, per i pazienti diventa più "difficile" superare l'asticella per essere classificati come positivi, e quindi:

1. abbiamo una riduzione del numero di pazienti positivi (veri e falsi)
2. abbiamo un incremento del numero di pazienti negativi (veri e falsi)

Di conseguenza, visto che aumenta il numero dei veri negativi e diminuisce il numero dei falsi positivi, avremo un incremento della specificità. Contemporaneamente, visto che diminuisce il numero dei veri positivi e aumenta il numero dei falsi negativi, avremo una riduzione della sensibilità. Analogamente, possiamo vedere che al ridursi del valore di cut-off succede l'opposto: aumenta la sensibilità e diminuisce la specificità.

In funzione dell'obiettivo che vogliamo perseguire, dobbiamo scegliere il miglior compromesso possibile fra due tipi di errore: falso positivo vs falso negativo. Vogliamo confermare oppure escludere la presenza di malattia?

5.2.1 Curva ROC

La curva ROC (Receiver Operating Characteristic) permette di rappresentare graficamente la relazione fra la sensibilità e la specificità del test al variare del cut-off. È un grafico in cui sono riportati coppie di valori, in ordinata la sensibilità (Sn) e in ascissa (1-specificità, 1-Sp), calcolati ai diversi punti di cut-off:

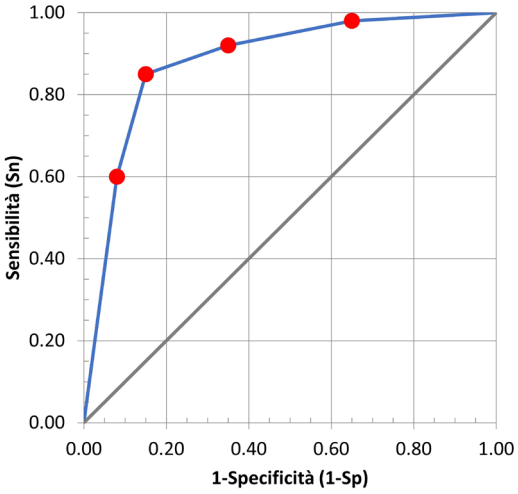

Figura 5.3. La curva ROC.

Tutte le curve ROC partono dal vertice in basso a sinistra che corrisponde ai valori Sn = 0; (1 – Sp) = 0 per finire nel vertice in alto a destra che corrisponde ai valori Sn = 1 e (1 – Sp) = 1. Ricordiamo che la sensibilità è anche indicata come proporzione di veri positivi e (1 – Sp) è indicata come proporzione di falsi positivi. Per un test diagnostico perfetto, per cui esiste un livello di cut-off che corrisponde a valori di sensibilità e specificità del 100%, la curva ROC corrispondente sarà la "spezzata" formata dai due lati (sinistro e superiore) del quadrato come riportato nella figura 5.4.

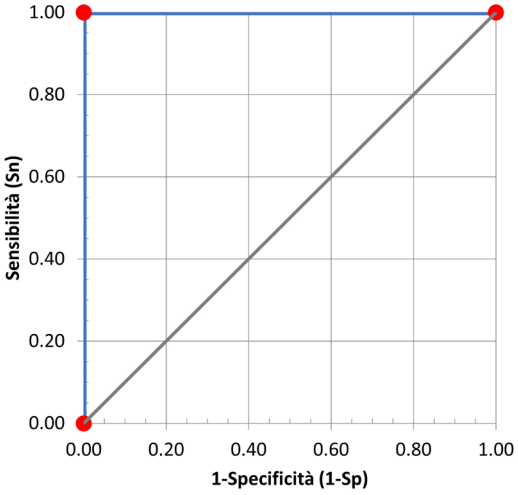

Figura 5.4. La curva ROC del test perfetto.

Per un test inutile, ovvero basto su di una variabile quantitativa identicamente distribuita in malati e non malati, vale a dire un test che non ci dà nessuna informazione aggiuntiva avendo entrambi i LR eguali a 1, la curva ROC corrispondente è la diagonale che unisce il vertice in basso a sinistra con quello in alto a destra, risultato del fatto che per ogni cut-off risulta SE = (1 – SP), come riportato in figura 5.5.

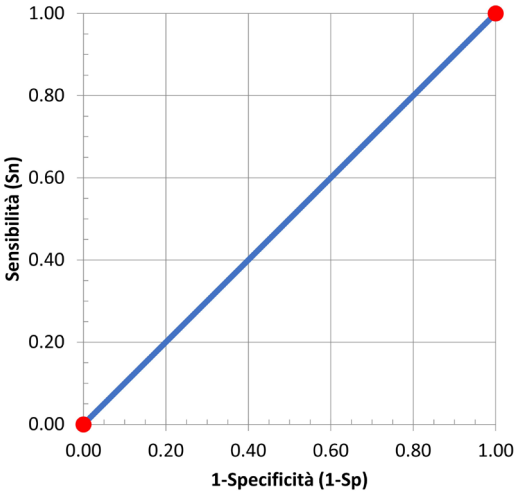

Figura 5.5. La curva ROC del test inutile.

La curva ROC corrispondente ad un test diagnostico reale è compresa fra la curva del test ideale e la curva del test inutile. Quanto più capace di discriminare malati da non malati sarà il test, tanto più la curva ROC corrispondente tenderà a passare vicino ai bordi sinistro e superiore del quadrato.

Vediamo un esempio pratico, il D-dimero misurato come agglutinazione al lattice per la diagnosi di tromboembolismo venoso. L'embolia polmonare è una patologia a presentazione clinica molto variabile, con uno spettro che comprende casi caratterizzati da elevata instabilità clinica e rapida evolutività negativa e casi in cui i sintomi sono più subdoli e di difficile interpretazione. Non diagnosticare un caso di embolia polmonare può avere conseguenze drammatiche (mortalità fino al 30%), ma gli esami considerati reference standard o dirimenti per la diagnosi sono spesso costosi, invasivi e non esenti da un elevato rischio intrinseco. Nella strategia diagnostica dell'embolia polmonare si inserisce quindi il D-dimero, un esame ad alta sensibilità, la cui negatività esclude la presenza di malattia. Il D-dimero rappresenta pertanto un test di "triage", che permette di aiutare nella decisione di procedere o no ad ulteriori accertamenti diagnostici in un paziente con sospetto di embolia polmonare.

Nella tabella seguente sono riportate le stime di accuratezza diagnostica per il D-dimero con metodo di misurazione di agglutinazione al lattice per la diagnosi

di tromboembolismo venoso in corrispondenza di tre livelli di cut-off: 250, 500 e 1000.

Cut-Off	Sensibilità	Specificità	LR+	LR-
250	0.91	0.53	1.96	0.16
500	0.88	0.59	2.14	0.21
1000	0.81	0.70	2.73	0.27

Tabella 5.5. Sensibilità, specificità e rapporti di verosimiglianza al variare del cut-off.

La curva ROC corrispondente, rappresentata in Figura 5.6, è la linea spezzata che collega i punti definite dalle coppie Sn, (1 – Sp) della tabella sopra riportata.

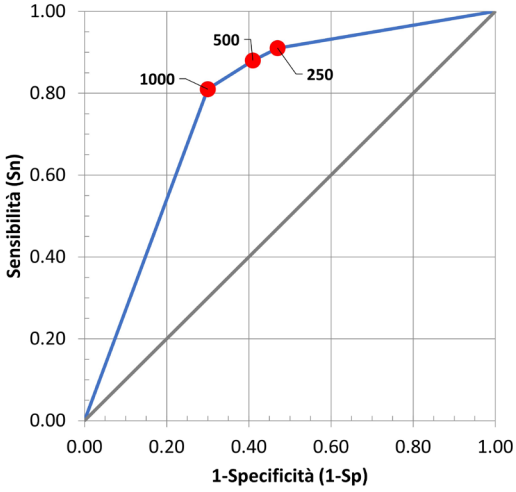

Figura 5.6. Accuratezza del D-dimero per la diagnosi di tromboembolismo venoso in corrispondenza di tre livelli di cut-off: 250, 500 e 1000.

Vediamo alcune caratteristiche "matematiche" della curva ROC che hanno importanza diagnostica. Piccoli spostamenti lungo la curva informano sulle variazioni reciproche di sensibilità e specificità per piccole variazioni del cut-off. In questo senso, è importante la pendenza locale (pendenza dei singoli tratti della spezzata) della curva: grande pendenza significa buon incremento di sensibilità con piccola perdita di specificità. Ad esempio, come si può notare dalla situazione illustrata nella Figura 5.7, quando il cut-off passa da 1000 a 500 (tratto in blu), la specificità passa da 0.60 a 0.50 (riduzione di 0.10), mentre la sensibilità passa da 0.70 a 0.97 (incremento di 0.27). Quando passiamo da 500 a 250 (tratto nero della curva ROC), la specificità passa da 0.50 a 0.40 (riduzione di 0.10), mentre la sensibilità va da 0.97 a 0.98 (incremento di 0.01). A parità di perdita di specificità, abbiamo un incremento di sensibilità molto maggiore nel

tratto blu che non nel tratto nero: il tratto blu è infatti più ripido del tratto nero (pendenza locale).

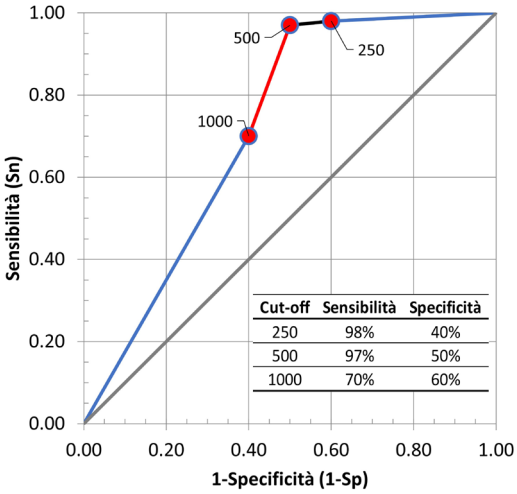

Figura 5.7. Interpretazione della pendenza di tratti di curva ROC: variazione di sensibilità e specificità al variare del cut-off.

Una caratteristica della curva spesso utilizzata (anche se talvolta in maniera poco appropriata) per confrontare la capacità diagnostica di diversi test è rappresentata dall'area sottesa (AUC, Area Under the Curve) che esprime il "potere diagnostico" del test ("diagnosticity") o capacità di discriminazione. Il test inutile, abbiamo visto, è quello che produce una curva coincidente con la diagonale principale (potere informativo del test nullo, LR per ciascun valore di cut-off pari a 1, AUC = 0.50, infatti il grafico in cui è iscritta la curva ROC è un quadrato di lato unitario, diviso a metà dalla sua diagonale). Il test ideale corrisponde ad una curva ROC con AUC = 1 (corrispondente all'area totale del quadrato). Quindi, quanto maggiore è l'AUC tanto maggiore sarà la capacità discriminante del test.

L'area sotto la curva ROC ha anche una facile interpretazione: si tratta di una probabilità, quella di classificare correttamente un malato se, estraendo a caso un paziente dalla popolazione dei non malati e un paziente dalla popolazione dei malati, si considera malato quello in cui il risultato del test è maggiore.

5.2.2 La scelta del cut-off

La curva ROC ci mostra i valori di sensibilità e specificità per i diversi valori di cut-off. Come possiamo scegliere il cut-off?

La scelta del cut-off dipende dall'utilizzo che vogliamo fare del test diagnostico, ovvero se intendiamo usarlo per confermare o per escludere il sospetto di malattia. Ad esempio, se volessimo utilizzare il D-dimero (Figura 5.8) per

escludere la malattia, dovremmo scegliere il cut-off di 250 (sensibilità alta, pochi falsi negativi, ma soprattutto LR- molto basso). Se, invece, nostro intendimento fosse di disporre di un test per confermare, se positivo, il sospetto di malattia embolia polmonare, dovremmo scegliere il cut-off di 1000, in corrispondenza del quale la specificità (o meglio il LR+) è massima (anche se comunque scarsa). In generale, quindi, ragionando sul piano dei rapporti di verosimiglianza:

– se l'intento è di utilizzare il test per rule-in, ci servirà un cut-off che fornisce un elevato valore del rapporto di verosimiglianza positivo (LR+), in modo che, in caso di positività al test, si avrà un rilevante incremento della probabilità di malattia, dalla pre-test alla post-test;
– se l'intento è di utilizzare il test per rule-out, ci servirà un cut-off che fornisce un basso valore del rapporto di verosimiglianza negativo (LR-), in modo che, in caso di negatività al test, si avrà una rilevante riduzione della probabilità di malattia, dalla pre-test alla post-test.

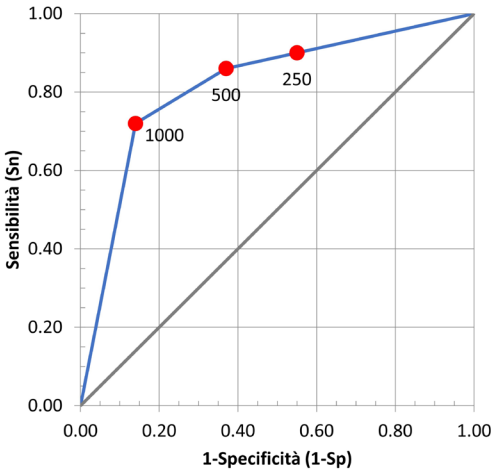

Figura 5.8. Utilizzo della curva ROC nella scelta del cut-off migliore.

Come possiamo confrontare le performance di due test? In base a quanto visto fino ad ora, un metodo grezzo è quello di utilizzare l'area sotto la curva. Questo metodo dà un indice sintetico di accuratezza diagnostica, ma non distingue fra sensibilità e specificità (necessità di rule in o di rule out della patologia sospettata). In realtà, per escludere una patologia sarà più utile un test SnNOut (rule out ovvero LR- basso) mentre per confermarla sarà da preferire un test SpPIn (rule in, ovvero LR + alto).

Torniamo all'esempio del D-dimero e confrontiamo, allora, l'accuratezza diagnostica per la diagnosi di tromboembolismo venoso di due metodi di misura diversi, ELISA ed agglutinazione al lattice (Latex).

	ELISA				Latex			
Cut Off	**Sensibilità**	**Specificità**	**LR+**	**LR-**	**Sensibilità**	**Specificità**	**LR+**	**LR-**
250	0.98	0.39	1.58	0.07	0.91	0.53	1.96	0.16
500	0.97	0.42	1.67	0.08	0.88	0.59	2.14	0.21
1000	0.93	0.58	2.23	0.13	0.81	0.70	2.73	0.27

Tabella 5.6 Accuratezza, per tre differenti cut-off, di due differenti metodi di misura del D-dimero nella diagnosi di tromboembolismo venoso.

Come possiamo decidere quale fra i due metodi utilizzare? Proviamo a rappresentare, sullo stesso grafico (Figura 5.9), le curve ROC dei due metodi.

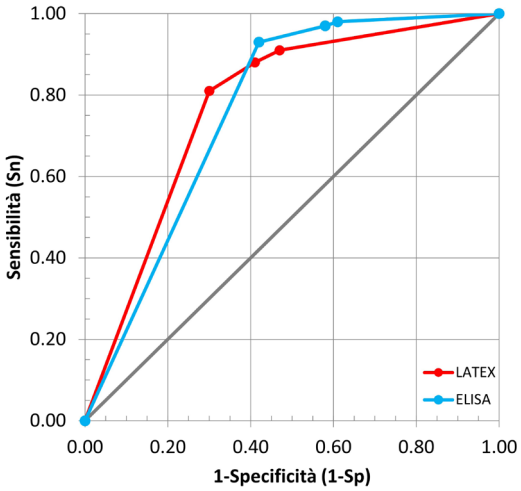

Figura 5.9. Rappresentazione grafica dei dati di accuratezza riportati in Tabella 5.6.

In generale, il test migliore sarà quello con la curva più "alta", vale a dire più vicina al punto in alto a sinistra di massima accuratezza, ovvero quella con AUC maggiore. Nel nostro caso, però, le due curve si incrociano e non si può individuare la curva migliore in assoluto. Per valori di specificità elevati (bassi valori di 1-specificità) il Latex è sicuramente migliore dell'ELISA, mentre, per valori elevati di sensibilità, ELISA risulta preferibile. Limitandosi a calcolare l'area sotto la curva, il Latex (AUC=0.7743) avrebbe un valore lievemente più alto dell'ELISA (AUC=0.7627), ma, poiché l'utilità del D-dimero è prevalentemente legata alla sua sensibilità, si nota che per valori di alta sensibilità l'ELISA è più vicino alla parte superiore del grafico rispetto al lattice ed è più utile. E, infatti, in base ai dati dell'esempio, l'ELISA con cut-off 250 è caratterizzato dal più basso valore di LR- (0.07). Quindi, in definitiva, non esiste un criterio assoluto per valutare l'accuratezza di un test diagnostico e deciderne l'utilità nella pratica

clinica: di volta in volta, avendo chiara la domanda clinica, in base all'utilizzo che vogliamo fare del test, possiamo decidere come trovare la risposta più appropriata, se sappiamo estrarre le informazioni utili dalla curva ROC.

In conclusione, dobbiamo sempre tenere presente che l'area sottesa alla curva è solo una sintesi unica dell'accuratezza globale di un test, che non tiene conto delle distinzioni fra sensibilità e specificità (e fra LR+ e LR-), utili, come abbiamo visto in precedenza, per valutare l'utilità clinica di un test. Il confronto fra due test dovrebbe sempre essere fatto valutando di quanto l'esito positivo (o negativo) è in grado di modificare la probabilità di malattia, da pre-test a post-test. Potrebbe, quindi, esserci un test con una minore AUC rispetto ad un altro, ma per il quale esiste un cut-off che fornisce migliori valori di LR (LR+ o LR-, in base all'uso che ne vogliamo fare) rispetto all'altro.

Punti chiave – Misure di accuratezza diagnostica: test variabile quantitativa

- ✓ Quando il test diagnostico è una variabile quantitativa (esempio, parametro di laboratorio) si deve stabilire una soglia di positività al test (cut-off) per poter calcolare sensibilità e specificità, come fatto con i test a risultato dicotomico.
- ✓ La curva ROC è una rappresentazione grafica delle coppie di valori di sensibilità e specificità calcolati in corrispondenza di diversi livelli di cut-off di un dato test di cui si voglia valutare l'accuratezza diagnostica.
- ✓ Nel grafico la sensibilità compare in ordinata ed il complemento ad 1 della specificità compare in ascissa. Si parla anche di "true positive ratio" e di "false positive ratio", ovvero di proporzione di veri e falsi positivi.
- ✓ È possibile esprimere la capacità discriminante stimando l'area sottesa alla curva (AUC) e usare tale misura per confrontare curve corrispondenti a test differenti, anche se abbiamo visto che questo approccio ci fornisce una visione limitata del fenomeno.
- ✓ considerando la curva ROC in dettaglio, è possibile valutare visivamente il cut-off più utile (escludere o confermare il sospetto di malattia).
- ✓ La scelta del cut-off migliore nelle diverse condizioni cliniche dipende dall'obiettivo che ci poniamo quando utilizziamo il test (per confermare, rule-in, o per escludere, rule-out, la diagnosi).
- ✓ Potremo, quindi, avere per un unico test due cut-off differenti, uno per rule-in ed uno per rule-out, scelti sulla base dei valori dei rapporti di verosimiglianza, valore elevato di LR+ (se rule-in) o valore basso di LR- (se rule-out).

5.3 La Concordanza

Siete di guardia in Pronto Soccorso. Arriva una donna di 76 anni con tosse, febbricola e dispnea presenti da circa 1 mese. Due settimane prima la paziente era già stata valutata in PS per sintomi analoghi, aveva eseguito gli esami del sangue che mostravano un minimo incremento degli indici di flogosi e del D-dimero, una radiografia del torace risultata nella norma e una tomografia computerizzata (TC) del torace con mezzo di contrasto (MDC), anch'essa negativa per trombo-embolia polmonare (TEP). La signora, con diagnosi di bronchite acuta, era stata rinviata al domicilio con terapia antibiotica e aerosol. Dopo le cure, però, la paziente vi riferisce persistenza dei disturbi; la radiografia del torace e gli esami del sangue ripetuti non sono dirimenti per la diagnosi, per cui decidete di ripetere una TC con MDC, che questa volta risulta positiva per TEP non massiva. Il radiologo di guardia, inoltre, confronta le immagini odierne con quelle eseguite 15 giorni prima e dice che, secondo lui, la TEP era già diagnosticabile nel precedente esame.

Vi sorgono allora innumerevoli domande: "Il radiologo che ha letto la prima TC è un incompetente?", "Il secondo radiologo ha valutato la prima TC col senno di poi?", "Pericolo scampato: questa volta è andata bene, la signora è ancora viva! Ma mi posso fidare dei referti delle prossime TC che chiederò nel sospetto di TEP?", "Quanto può influire la soggettività dell'operatore nel giudicare positività o negatività della TC?".

Identificando nella riproducibilità dell'osservazione il tema centrale intorno al quale i vostri quesiti ruotano, decidete di fare una ricerca bibliografica su PubMed secondo lo schema che avete imparato:

- P (patients/pathology): pulmonary embolism;
- I (intervention): CT scan;
- C (comparison): …;
- O (outcome): diagnosis, inter-observer agreement.

Nella prima schermata, trovate uno studio che sembra fare al caso vostro che parla della concordanza osservata tra 4 diversi radiologi nell'interpretazione di 46 TC eseguite per sospetta embolia polmonare. I risultati sono forniti come statistica κ che è risultata 0.82. Quando, però, gli autori hanno escluso esami positivi per TEP massiva, e non è il nostro caso, considerando TC negative o positive per TEP non massiva, la fantomatica statistica κ risulta di 0.47 (*Am J EmergMed 2009;27:1109-11).*

A questo punto, cerchiamo di capire un po' meglio di cosa si tratta.

Partiamo anzitutto da una migliore definizione del concetto di riproducibilità di un test, che ha a che fare con la coincidenza di giudizio che si osserva quando due o più operatori diversi e indipendenti, o anche lo stesso operatore in occasioni indipendenti, esaminano lo stesso quadro (interobserver agreement; intraobserver agreement).

La concordanza osservata si misura come rapporto fra numero di referti concordanti e numero totale di referti formulati. Se consideriamo due operatori che giudicano in base ad una variabile di tipo dicotomico (malattia "presente" o "assente"), la concordanza osservata sarà il rapporto tra il totale degli esami positivi o negativi per entrambi e gli esami totali eseguiti. Nella Tabella 5.7 potete trovare un esempio numerico in cui 85 radiografie del torace sono state valutate da due operatori per la diagnosi di polmonite.

		Operatore 2		
		Rx +	Rx -	Totale
Operatore 1	Rx +	61	10	71
	Rx -	3	11	14
	Totale	64	21	85

Tabella 5.7 Valutazione di 85 radiografie da parte di due operatori.

Dalla tabella si ricava una misura di concordanza osservata pari a (61+11)/85=0.847, apparentemente molto buona. Essa però è una misura grezza di accordo fra operatori e può risultare fuorviante perché non tiene conto dell'effetto del caso. Infatti, se i due operatori avessero refertato le radiografie ciascuno in base all'esito del lancio di una moneta, un certo grado di concordanza sarebbe comunque risultato presente, pur essendo del tutto attribuibile al caso. Tutte le volte che si misura l'accordo fra due operatori non si deve dimenticare questo fenomeno: ovvero che una quota di concordanza osservata è puramente attribuibile al caso e non esprime l'affidabilità di giudizio degli operatori. La statistica κ, sotto definita in formula, è stata messa a punto da Cohen, per fornire una stima di concordanza aggiustata per il caso:

$$k = \frac{\text{concordanza osservata} - \text{concordanza casuale}}{1 - \text{concordanza casuale}}$$

Come si vede, si tratta di eliminare dal numeratore la quota di concordanza osservata attribuibile al caso e di rapportare il risultato ottenuto al massimo di concordanza ottenibile, una volta che dal totale dei confronti si siano eliminati quelli concordanti per caso.

Per un maggior dettaglio nelle modalità di calcolo della concordanza casuale rimandiamo all'appendice.

Nell'esempio avevamo trovato una concordanza osservata molto buona: (61+11)/85=0.847.

Calcolando la statistica κ otteniamo una concordanza aggiustata più bassa: k = 0.535.

Oltre al fatto che non tiene conto dell'intervento del caso, la concordanza osservata come misura di riproducibilità può risultare fuorviante anche per un altro motivo. Vediamo il perché con un altro esempio.

Consideriamo la lettura di mammografie effettuate per lo screening di neoplasia della mammella: è ovvio che la maggioranza degli esami avrà esito negativo e solo una minoranza risulterà positiva, avremo cioè uno sbilanciamento simmetrico (cioè che va nella stessa direzione per i due operatori) nella distribuzione dei risultati positivi e negativi, per effetto della bassa prevalenza della malattia. Supponiamo allora di fare la seguente osservazione su 1000 mammografie.

		Operatore 2		
		Referto +	Referto –	Totale
Operatore 1	Referto +	0	10	10
	Referto –	10	980	990
	Totale	10	990	1000

Tabella 5.8 Valutazione di 1000 mammografie da parte di due operatori in un contesto di screening mammografico.

La concordanza osservata è 98%, ma la realtà è diversa: la concordanza sui referti negativi è casuale, mentre sui referti positivi la discordanza è totale. Anche in questo caso estremo $k = -0.01$ fornisce una immagine più appropriata della realtà che intendiamo valutare. Al di là del caso, non si ha concordanza fra osservatori, anzi, si può notare una tendenza sistematica a discordare nel giudizio, specie in caso di positività.

Per interpretare il valore di κ ottenuto si può fare riferimento alla seguente griglia:

Valori di κ	**Interpretazione:**
– 0.00-0.20	concordanza assente
– 0.21-0.40	concordanza modesta
– 0.41-0.60	concordanza moderata
– 0.61-0.80	concordanza buona
– 0.81-1.00	concordanza ottima

Tornando al caso clinico iniziale, alla luce di quanto abbiamo imparato sulla statistica *k* e dallo studio che abbiamo trovato, possiamo dire che quando un referto di TC torace con MDC descrive una TEP massiva, possiamo essere praticamente certi che il risultato verrà giudicato da radiologi indipendenti con una concordanza ottima, mentre quando i radiologi sono chiamati a giudicare un quadro dubbio (negativo o positivo per TEP segmentaria o sub-segmentaria), la concordanza tra gli operatori che ci dobbiamo aspettare è moderata, se non modesta, da qui nessuna sorpresa se primo e secondo radiologo discordano.

Per concludere, possiamo dire che quando ci si trova di fronte all'utilizzo di un test diagnostico, forse ancora prima che informazioni sulla sensibilità e specificità del test, può essere utile avere informazioni sulla concordanza nella lettura del test da parte di diversi operatori, soprattutto se ci interroghiamo circa

l'utilità di chiedere un secondo parere. Se poi avessimo un test di altissima sensibilità e specificità la cui interpretazione è affidabile solo se valutato da osservatori esperti di un centro specialistico di terzo livello (come spesso avviene negli studi!), è ovvio che la generalizzabilità alla pratica clinica di quel test sarà bassa.

Ci sono situazioni complesse in cui bisogna utilizzare altri metodi per valutare la concordanza. Ad esempio, se le classi diagnostiche sono maggiori di due o se la prevalenza di positività è molto bassa. Una rassegna approfondita di valutazione della concordanza si può trovare in letteratura.

Punti chiave – Concordanza

- ✓ Oltre a (e prima di) sensibilità, specificità e alle altre misure di accuratezza diagnostica fino a qui analizzate, bisogna valutare la concordanza tra gli operatori per capire se l'interpretazione di un test è riproducibile.
- ✓ Il metodo più semplice è calcolare la concordanza osservata (numero di esami concordanti/numero di esami totali) confrontando i giudizi formulati da osservatori indipendenti.
- ✓ La concordanza sopra definita è però influenzata anche dalla quota di concordanza casuale.
- ✓ Per aggiustare la stima per la quota casuale, la misura più diffusa è la statistica *k* di Cohen.

5.4 Misure di associazione per gli studi prospettici. Valutare l'efficacia di un trattamento

Un uomo di 70 anni viene ricoverato per una riacutizzazione di scompenso cardiaco; ha una cardiopatia ipocinetica post-ischemica con una frazione di eiezione del 25% e, dopo aver superato la fase acuta, ha dispnea solo per sforzi moderati. È in terapia medica massimale con ACE-inibitore, betabloccante e furosemide e ha una funzionalità renale normale; vi chiedete, quindi, se aggiungere anche un anti-aldosteronico possa dare un beneficio al paziente.

Grazie alle vostre ormai ottime capacità, recuperate facilmente lo studio *EMPHASIS-HF* pubblicato sul New England (*N Engl J Med 2011;364:11-21*).

Nello studio sono stati arruolati 2737 pazienti con FE<35% e classe NYHA II (quindi simili al nostro paziente) randomizzati a ricevere eplerenone 50 mg/die o placebo, in aggiunta alla terapia medica tradizionale. L'endpoint primario (un endpoint composito di mortalità per eventi cardiovascolari o ricovero per scompenso cardiaco) si è osservato in 249 dei 1364 soggetti del gruppo di trattamento e in 356 dei 1373 soggetti del gruppo placebo. Che cosa significano questi risultati? Dobbiamo concludere che è utile l'eplerenone?

Per rispondere a questa domanda possiamo costruire la tabella 2x2 e iniziare a fare qualche calcolo.

		Mortalità per eventi cardiovascolari o ricovero per scompenso cardiaco		
		Sì	No	Totale
Eplerenone	Sì	249	1115	**1364**
	No	356	1017	**1373**
	Totale	**605**	**2132**	**2737**

Tabella 5.9 Relazione fra assunzione di eplerenone e mortalità per eventi cardiovascolari o ricovero per scompenso cardiaco.

5.4.1 Rischio Assoluto

Qual è il rischio assoluto di mortalità o ricovero nel gruppo dei trattati?

Il rischio assoluto **è la** probabilità che un paziente, preso a caso dal gruppo considerato (trattati o non trattati), manifesti l'evento, può variare da 0 a 1 (0% e 100%) e per stimarlo dobbiamo semplicemente fare il rapporto fra il numero di soggetti trattati che hanno manifestato l'evento ed il numero totale dei soggetti trattati. Nel nostro esempio, il rischio assoluto di evento nei soggetti trattati è 249/1364 = 0.18 = 18% e nel caso dei controlli avremo 356/1373 = 0.26 = 26%. I rischi assoluti costituiscono il punto di partenza per la valutazione dell'efficacia di un trattamento o dell'effetto dell'esposizione ad un fattore di rischio. Le misure di sintesi che vedremo di seguito si basano sul confronto fra i rischi assoluti.

5.4.2 Rischio Relativo

Per un confronto fra rischio nei trattati e nei controlli si può ricorrere al rapporto fra i due rischi, calcolando il cosiddetto rischio relativo:

$$RR = \frac{\text{rischio trattati}}{\text{rischio non trattati}}$$

Il rischio relativo, quindi, essendo il rapporto fra due rischi, esprime quanto è più probabile che un evento si manifesti nei trattati rispetto ai non trattati. Il RR non è una probabilità, ma un rapporto di probabilità e, di conseguenza, può variare fra 0 ed ∞. Nel caso particolare in cui il rischio dei trattati è uguale a quello dei controlli abbiamo RR=1. Quando il rischio dei trattati è maggiore di quello dei non trattati, il RR è >1 (trattamento è fattore di rischio). Viceversa, quando il rischio dei trattati è inferiore rispetto a quello dei non trattati, il RR è <1 (trattamento è fattore protettivo).

Per quanto riguarda l'eplerenone, il RR è: 0.18/0.26=0.69: i trattati hanno un rischio, di morte o ricovero per scompenso, pari al 69% di quello dei non

trattati. Da questo si può facilmente calcolare la riduzione relativa del rischio, cioè 1-RR, che in questo caso significa che i trattati hanno un rischio del 31% (1-0.69) inferiore a quello dei non trattati, come spiegato nel paragrafo successivo.

NB: Il rischio relativo può essere calcolato solo in studi di tipo prospettico.

5.4.3 Riduzione Assoluta del Rischio

Tornando ai risultati dello studio *EMPHASIS-HF*, il rischio relativo non è l'unica misura di associazione che possiamo ricavare. In Tabella 5.9, se, invece di fare il rapporto fra il rischio di decesso o ricovero nei trattati ed il rischio di decesso o ricovero nei non trattati, ne facessimo la differenza, mettendo ora al primo posto il rischio dei controlli, otterremmo la riduzione di rischio assoluto (Absolute Risk Reduction, ARR), detta anche differenza di rischio (Risk Difference, RD):

ARR = rischio di eventi non trattati - rischio di eventi trattati

Nel nostro esempio: ARR = 0.26-0.18 = 0.08, che interpretiamo nel seguente modo. Immaginiamo di avere un gruppo di 100 pazienti che dovrebbero essere trattati con il trattamento di controllo, i deceduti sarebbero 26. Ora, immaginiamo che sia disponibile il nuovo trattamento e quindi tutti quei 100 pazienti sono sottoposti al nuovo trattamento: i deceduti sarebbero ora 18. Questo significa che con il nuovo trattamento salviamo la vita ad 8 di quei 26 pazienti che sarebbero destinati a decesso, se trattati con il controllo.

5.4.4 Riduzione Relativa del Rischio

Riprendiamo l'esempio dello studio *EMPHASIS-HF*. Consideriamo la differenza assoluta di rischio calcolata al paragrafo precedente (ARR=0.08) e valutiamo questa differenza in termini relativi rispetto al rischio del gruppo di controllo. Chiediamoci a quale quota del rischio dei controlli corrisponde questo 0.08 di riduzione. Per rispondere a questa domanda basta semplicemente fare il rapporto fra la differenza di rischio ed il rischio dei controlli, nel nostro esempio: 0.08/0.26=0.31. Proviamo ora ad interpretare il risultato ottenuto. Come abbiamo visto, il rischio di mortalità/ricovero nei controlli è pari a 0.26. Il trattamento riduce questo rischio di 0.08, riduzione che equivale, ragionando in termini relativi, al 31% del rischio iniziale dei controlli. Vale a dire, con il trattamento abbiamo abbattuto il rischio dei controlli di una quota pari al 31%. Consideriamo che ridurre il rischio dei controlli del 100% vuol dire che il trattamento elimina il rischio di mortalità/ricovero. Per tornare all'interpretazione data in precedenza al RR, si può notare come questo 31% sia esattamente lo stesso 31% di riduzione ottenuto dall'interpretazione del rischio relativo (1-RR) vista sopra.

Le misure viste sino ad ora (RR, ARR, RRR) sono utilizzabili anche per valutare l'associazione fra esposizioni ed eventi di interesse negli studi osservazionali prospettici (es. coorte).

5.4.5 Number Needed to Treat

A partire dalla riduzione di rischio assoluto (ARR), possiamo calcolare altri indici che ci permettono di esprimere le differenze fra i rischi in un modo più facilmente comprensibile anche da chi non ha un buon rapporto con la statistica. La ARR ci dice che ogni 100 pazienti trattati con il nuovo trattamento rispetto al controllo abbiamo 8 decessi in meno: dobbiamo trattare 100 pazienti per avere una riduzione di 8 decessi. Quanti pazienti dobbiamo trattare per avere 1 decesso in meno? Per rispondere a questa domanda, vediamo che non dobbiamo fare altro che partire dalla riduzione assoluta del rischio. Calcolando infatti l'inverso della riduzione di rischio assoluto (1/ARR) otteniamo il così detto Number Needed to Treat (NNT):

$$\text{NNT} = \frac{1}{\text{rischio di decesso non trattati - rischio di decesso trattati}} = \frac{1}{\text{ARR}}$$

NNT indica il numero di pazienti che devono essere trattati per ottenere, grazie al trattamento, un beneficio aggiuntivo (beneficio misurato, ad esempio, come decesso evitato).

Nel nostro esempio NNT=1/0.08=12.5.

Ciò significa che è necessario trattare con eplerenone, arrotondando sempre per eccesso, 13 pazienti per prevenire un decesso per cause cardiovascolari o un ricovero per scompenso cardiaco: in media, ogni 13 pazienti sottoposti a trattamento, 1 avrà un beneficio che, senza trattamento, non avrebbe potuto avere.

Quanto più è basso il NNT, tanto più efficace è da considerare il trattamento.

5.4.6 Number Needed to Harm

Leggendo i dati dello studio *EMPHASIS-HF* ci stiamo convincendo che l'eplerenone potrebbe essere un farmaco utile per il paziente. Tuttavia, bisogna considerare anche i possibili eventi avversi legati al trattamento stesso, e in particolare, siamo preoccupati del rischio di iperkaliemia.

Tra i 1360 pazienti trattati con eplerenone (la popolazione su cui viene calcolata la sicurezza del farmaco è quella che effettivamente ha ricevuto la terapia e, quindi, i soggetti sono meno numerosi rispetto a quelli della Tabella 5.9), 109 hanno sviluppato iperkaliemia durante il trattamento. Nella popolazione di controllo tale effetto collaterale si è verificato in 50 dei 1369 pazienti che hanno ricevuto il placebo. I dati sono riportati in Tabella 5.10.

		Iperkaliemia		
		Sì	**No**	**Totale**
Eplerenone	**Sì**	109	1251	**1360**
	No	50	1319	**1369**
	Totale	**159**	**2570**	**2729**

Tabella 5.10 Relazione fra assunzione di eplerenone e iperkaliemia.

Analogamente a quanto abbiamo visto relativamente all'efficacia, anche per quanto riguarda gli eventi avversi, possiamo calcolare RA, RR e ARR:

- RA trattati=109/1360=0.08=8%
- RA non trattati=50/1369=0.03=3%
- RR=0.08/0.03=2.66
- ARR=0.08-0.03=0.05=5%

Essendo il trattamento fattore di rischio per l'insorgenza di iperkaliemia, possiamo ottenere il così detto Number Needed to Harm (NNH). Il NNH ha lo stesso significato del NNT, riferendosi però agli eventi avversi, ovvero indica ogni quanti trattati si verifica in media un evento avverso:

$$\text{NNH} = \frac{1}{\text{ARR}}$$

Nel nostro esempio: NNH=1/0.05=20, ovvero ogni 20 pazienti trattati ci aspettiamo un evento avverso aggiuntivo.

Quanto più è elevato il valore di NNH, tanto più sono rari gli eventi avversi dovuti al trattamento considerato.

Per decidere se applicare un trattamento ai propri pazienti sulla base di NNT e NNH bisogna: confrontare la rilevanza degli eventi considerati come beneficio e "harm"; confrontare il valore assoluto di NNT rispetto a NNH. Nel nostro esempio, tenendo conto che gli eventi avversi osservati nello studio *EMPHASIS-HF* sono in generale clinicamente poco rilevanti (iperkaliemia, lieve insufficienza renale, ginecomastia, ipotensione arteriosa), che l'outcome primario è molto rilevante (mortalità e ospedalizzazione), e che NNH è quasi il doppio di NNT, ci potremmo sentire abbastanza sicuri nel prescrivere l'eplerenone al nostro paziente con scompenso cardiaco. **È** da segnalare che, di regola, gli studi randomizzati non sono gli studi migliori per valutare gli eventi avversi dei farmaci, che sono rilevati in maniera più appropriata dagli studi osservazionali di registro.

Sebbene NNT e NNH e ARR forniscano un'informazione più rilevante dal punto di vista clinico (ad esempio, quanti pazienti devo trattare per salvare una vita), essi risentono maggiormente della possibile non generalizzabilità dei risultati. In modo simile a VPP e VPN dipendono dalla prevalenza di eventi nello

studio originale. Se la prevalenza di eventi nella nostra popolazione è molto diversa da quella dello studio, non potremo trasferire queste misure di efficacia ai nostri pazienti.

Viceversa, il rischio relativo, pur essendo generalizzabile perché non dipende dalla prevalenza di eventi, potrebbe dare un'informazione fuorviante, portando alla sovrastima dell'efficacia di un trattamento. Ad esempio, un trattamento con RR di 0.7, che implica una riduzione del 30% del rischio di partenza, avrà un impatto diverso se somministrato ad un paziente con rischio base del 40%, il cui rischio verrebbe ridotto al 28%, NNT=8.3, rispetto ad un paziente con rischio base dell'1%, il cui rischio sarebbe ridotto allo 0.7%, NNT=334. Il significato clinico del RR dipende anche dalla prevalenza di malattia: spesso un trattamento con un RR elevato, se applicato ad una popolazione con bassa prevalenza di malattia, si traduce in un'efficacia modesta.

5.5 Misure di associazione negli studi retrospettivi

5.5.1 Odds ratio

Avete ricoverato un paziente con scompenso cardiaco e avete fatto una profilassi antitrombotica durante l'ospedalizzazione, dovete decidere se proseguirla o meno dopo la dimissione. Trovate uno studio caso-controllo in cui sono stati arruolati come casi soggetti non ricoverati affetti da trombosi venosa profonda (TVP) e come controlli pazienti non affetti da TVP. Negli uni e negli altri sono stati rilevati fattori che si presume siano associati all'incidenza di TVP *(Thrombosis Research 2010. 126;367–372)*. Dallo studio si ricava che 32 dei 774 casi con TVP e 130 dei 7740 controlli erano anche affetti da scompenso cardiaco (dati adattati per semplicità).

In questo caso, non è possibile calcolare l'incidenza di TVP nei due gruppi, perché i soggetti sono stati selezionati "a tavolino" in base alla presenza o meno di TVP e quindi la prevalenza di TVP nel campione selezionato non rispetta la prevalenza della popolazione. Mentre nel RCT visto in precedenza avevamo definito da disegno due sottogruppi (gruppo dei trattati vs gruppo dei controlli), da confrontare in base all'occorrenza di un evento di interesse (rischio di decesso/ospedalizzazione), nei disegni caso-controllo la situazione si ribalta. Visto che i due sottogruppi che definiamo da disegno sono proprio i casi ed i controlli (soggetti con TVP vs soggetti senza TVP), pot Tabella5_11giusta remo allora solamente confrontare questi due sottogruppi in base alla frequenza di caratteristiche di nostro interesse, ad esempio l'esposizione ad un fattore di rischio (presenza di scompenso cardiaco). Potremo quindi vedere come il 4.1% (32/774) dei casi in confronto con l'1.7% (130/7740) dei controlli erano scompensati. Questo confronto, fra il "rischio" di essere scompensato fra i casi e fra i controlli, ci permette di concludere che presumibilmente c'è associazione

fra scompenso cardiaco e TVP. Non possiamo però calcolare il rischio relativo di TVP, dato dal rapporto fra il rischio di TVP negli scompensati rapportato al rischio di TVP fra i non scompensati, che sarebbe la misura statistica più facilmente interpretabile da parte nostra. L'unica cosa che possiamo fare è "inventare", ricorrendo ad artifici matematici (vedi dettagli in appendice), una misura di associazione che sia una stima del rischio relativo e che sia interpretabile esattamente come un rischio relativo. Per tale motivo, si utilizza come misura di associazione, in sostituzione del rischio relativo, l'Odds Ratio. Cerchiamo di chiarire con l'esempio della TVP. Anche in questo caso possiamo organizzare i dati in una tabella 2x2 (Tabella 5.11).

		TVP		
		Sì	**No**	**Totale**
Scompenso cardiaco	**Sì**	32	130	**162**
	No	742	7610	**8352**
	Totale	**774**	**7740**	**8514**

Tabella 5.11 Scompenso cardiaco e trombosi venosa profonda (TVP) in 8514 pazienti.

Come dicevamo in precedenza, in uno studio così disegnato non ha senso calcolare il rischio di TVP fra i soggetti con e senza scompenso cardiaco, poiché lo studio è stato condotto reclutando i pazienti proprio in base alla presenza o assenza di TVP, e non in base alla presenza o assenza di scompenso cardiaco. Per questo, il 19.8% (32/162, percentuale di soggetti con TVP fra i 162 con scompenso) non è interpretabile come rischio di TVP, essendo influenzato dalla decisione di reclutare un numero di controlli pari a 10 volte quello dei casi. Se infatti gli autori si fossero limitati, esattamente nella stessa situazione, a reclutare un eguale numero di casi e controlli i dati sarebbero quelli riportati nella Tabella 5.12.

		TVP		
		Sì	**No**	**Totale**
Scompenso cardiaco	**Sì**	**A** 32	**B** 13	**45**
	No	**C** 742	**D** 761	**1503**
	Totale	**774**	**774**	**1548**

Tabella 5.12 Scompenso cardiaco e trombosi venosa profonda (TVP) in 1548 pazienti. Campione composto da egual numeri di casi e di controlli.

Da questa tabella risulterebbe un rischio di TVP fra gli scompensati pari al 71% (32/45), rispetto al 19.5% della tabella precedente. Negli studi caso-controllo non è quindi possibile calcolare il rischio relativo. Quello che però risulta

invariato nelle due tabelle è la misura di associazione chiamata odds ratio (OR), che può essere calcolata facendo il rapporto dei prodotti incrociati di ciascuna tabella:

$$OR = \frac{A * D}{B * C}$$

Considerando i dati in Tabella 5.11: OR = 32 x 7610 / 130 x 742 = 2.52;
Considerando i dati in Tabella 5.12: OR= 32 x 761 / 13 x 742=2.52.

L'OR è interpretabile esattamente come un rischio relativo: valori prossimi ad 1 sono indicativi di assenza di associazione fra esposizione ed evento, mentre valori minori (o maggiori) di 1 indicano che siamo in presenza di un fattore protettivo (di rischio).

Tornando al nostro esempio, 2.52 è un valore di OR che segnala l'esistenza di un'associazione positiva fra le due condizioni considerate (TVP e scompenso): grossolanamente, possiamo concludere che il rischio di TVP nei pazienti affetti da scompenso cardiaco è più del doppio (2.52 volte) rispetto a quelli senza scompenso cardiaco.

In appendice potete trovare un approfondimento matematico riguardante l'odds ratio.

L'odds ratio, come il RR di cui è un surrogato, può assumere valori compresi tra zero e $+\infty$:

- **OR = 1** assenza di associazione tra esposizione e malattia;
- **OR < 1** associazione negativa (l'esposizione può proteggere dalla malattia);
- **OR > 1** associazione positiva (l'esposizione può causare la malattia).

Se utilizzato come stima del rischio relativo in studi in cui la prevalenza di eventi è alta, l'odds ratio risulta maggiore del RR e quindi interpretato come RR in realtà ne rappresenta una sovrastima. Per chiarire questo concetto ricorriamo, nelle tabelle 5.13 e 5.14, ad esempi con numeri scelti "a tavolino".

Prevalenza 50%		Malattia		
		Sì (casi)	No (controlli)	Totale
Esposizione	Sì (esposti)	70	30	**110**
	No (non esposti)	30	60	**90**
	Totale	**100**	**100**	**200**

Tabella 5.13 Associazione fra esposizione e malattia: scenario con prevalenza del 50%.

$$\text{RR} = \frac{\frac{70}{110}}{\frac{30}{90}} = 2 \quad \text{OR} = \frac{70 \times 60}{30 \times 30} = 3.5$$

Prevalenza 0.75%		Malattia		
		Sì (casi)	No (controlli)	Totale
Esposizione	Sì (esposti)	10	990	**1000**
	No (non esposti)	5	995	**1000**
	Totale	**15**	**1985**	**2000**

Tabella 5.14 Associazione fra esposizione e malattia: scenario con prevalenza dello 0.75%.

$$\mathrm{RR} = \frac{\frac{10}{1000}}{\frac{5}{1000}} = 2 \quad \mathrm{OR} = \frac{10 \times 995}{990 \times 5} = 2$$

L'OR, così come il RR, rappresenta sempre una stima campionaria di quella che intendiamo sia una misura di forza di associazione fra esposizione e malattia, e quindi è necessario ricorrere sempre al calcolo dell'intervallo di confidenza se si vuole fornire un'informazione utile per la pratica clinica.

L'OR è una stima del RR che tende a sovrastimare l'effetto (sia protettivo che di rischio). Abbiamo detto che l'OR è da utilizzare esclusivamente negli studi retrospettivi, perché nei prospettici possiamo calcolare direttamente il RR. Tuttavia, in molti studi prospettici si ricorre comunque ad analisi mediante OR. Questo accade perché in molti contesti di studi prospettici si ricorre al metodo di analisi statistica di regressione logistica, che consente di aggiustare le stime di associazione per le variabili di confondimento o di studiare l'effetto di variabili di interazione, in modo molto efficiente.

Punti chiave – Studi prospettici e retrospettivi: misure di associazione

- ✓ Riduzione assoluta del rischio (ARR) è la differenza tra il rischio assoluto (RA) nei trattati (esposti) e nei controlli (non esposti).
- ✓ Dalla differenza di rischio assoluto si può calcolare il number needed to treat (NNT), che esprime il numero di pazienti da trattare per ottenere un beneficio (evitare un decesso, un evento cardiovascolare, un ricovero) e permette di avere una stima più immediata della rilevanza clinica dei risultati dello studio.
- ✓ Il rischio relativo (RR) è il rapporto tra il rischio assoluto (RA) (di decesso, di eventi) nel gruppo dei trattati e il rischio nel gruppo di controllo; esprime quanto è più probabile o meno probabile che un evento si manifesti nei trattati rispetto ai controlli.
- ✓ Il RR può essere calcolato in tutti gli studi prospettici (coorte, RCT) in cui

l'incidenza di malattia o il rischio cumulativo in un determinato intervallo di tempo è direttamente stimabile.

- ✓ La riduzione relativa del rischio assoluto è una ulteriore modalità di calcolo dell'effetto del trattamento (esposizione), che consiste nel valutare quanto la riduzione assoluta pesa sul rischio assoluto del gruppo di controllo (non esposti) e lo si interpreta, quindi, come una variazione percentuale.
- ✓ Quando, come accade per gli studi caso-controllo, a causa del disegno di studio non possiamo calcolare i rischi assoluti (ovvero l'incidenza), non possiamo nemmeno calcolare il RR e dobbiamo ricorrere ad una sua stima, l'odds ratio (OR).
- ✓ L'odds ratio permette di fornire una stima del rischio relativo in alcune condizioni particolari; è una buona approssimazione del RR quando l'incidenza di eventi è rara.
- ✓ RR e OR hanno il vantaggio di poter essere applicati anche a una popolazione con incidenza di eventi diversa da quella dello studio originale, ma possono portare ad una sovrastima dell'effetto del trattamento se la prevalenza di eventi è bassa.
- ✓ La riduzione di rischio assoluto e il NNT sono clinicamente più utili per decidere per il singolo paziente, ma è necessario che la prevalenza di eventi nello studio originale sia simile a quella della popolazione a cui voglio applicarne i risultati.

Bibliografia consigliata

Accuratezza diagnostica

Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. *BMJ*. 1994;308:1552.

Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ*. 1994;309:102.

Altman DG, Bland JM. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots. *BMJ*. 1994;309:188.

Costantino G, Montano N, Casazza G. When should we change our clinical practice based on the results of a clinical study? Diagnostic accuracy studies II: the diagnostic accuracy. *Intern Emerg Med*. 2016 Aug;11(5):755-7.

Deeks JJ, Altman DJ. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329;168-169.

Jaeschke R, Guyatt G, Sackett DL, et al. Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test A. Are the Results of the Study Valid? *JAMA*. 1994;271(5):389–391.

Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, III: how to use an article about

a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA*. 1994;271:703-707.

Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J, for the Evidence-Based Medicine Working Group. Users' Guides to the Medical Literature: XV. How to Use an Article About Disease Probability for Differential Diagnosis. *JAMA*. 1999;281(13):1214–1219.

Stein PD et al. d-Dimer for the exclusion of acute venous thrombosis and pulmonary embolism. A systematic review. *Ann Int Med*. 2004;140:589-602.

The task force for the diagnosis and management of the acute pulmonary embolism of the european society of cardiology. Guidelines on the diagnosis and management of acute pulmonary embolism. *Eur Heart Jour*. 2008;29:2276-2315.

Concordanza

Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993 May;46(5):423-9.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977 Vol.33,pp.159-174.

Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol*. 1988;41(10):949-58.

Misure di associazione

Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. *BMJ*. 1998;317:1318.

Bland JM, Altman DJ. Statistics Notes: The odds ratio. *BMJ*. 2000;320;1468.

Higgins JPT, Li T, Deeks JJ (editors). Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.

Holmberg MJ, Andersen LW. Estimating Risk Ratios and Risk Differences: Alternatives to Odds Ratios. *JAMA*. 2020;324(11):1098–1099.

Storm CL, Garvan CW. Proportions, Odds, and Risk. *Radiology*. 2004;230:12-19.

Viera AJ. Odds Ratios and Risk Ratios: What's the Difference and Why Does It Matter? *Southern Medical Journal*. 2008;101(7):730-734.

6. Quesiti, Prove, verifiche di ipotesi e p-values

Nel vostro reparto viene ricoverato per il terzo episodio di scompenso cardiaco nell'arco di due mesi un paziente di 82 anni affetto da ipertensione, cardiopatia ischemica, BPCO e insufficienza renale cronica moderata. L'ecocardiogramma mostra il peggioramento di una nota stenosi aortica, attualmente di grado severo. Il paziente è ovviamente ad alto rischio e poco candidabile per un intervento chirurgico, vi chiedete quindi se una sostituzione valvolare transcatetere potrebbe migliorare la sua prognosi.

Vi mettete subito all'opera su PubMed, trovate uno studio del 2010 sull'argomento (*NEJM 2010;363:1597-1697*). In questo trial vengono randomizzati a trattamento tradizionale (terapia medica o valvuloplastica con palloncino) vs impianto di valvola transcatetere per via femorale (*TAVI*) 358 pazienti con stenosi aortica severa, giudicati non sottoponibili ad intervento chirurgico. L'endpoint primario dello studio è la mortalità per ogni causa ad 1 anno, e risulta inferiore nel gruppo trattato con TAVI rispetto al gruppo di controllo (30.7% vs 49.7%).

Ogni studio dovrebbe nascere dall'esigenza di rispondere ad una domanda, che nel nostro caso potrebbe essere: "La TAVI migliora la prognosi dei pazienti con stenosi aortica severa non candidabili all'intervento chirurgico?". Obiettivo dello studio è dunque cercare di rispondere a questa domanda, attraverso la formulazione e la successiva verifica di un'ipotesi.

L'evento che nello studio viene concretamente misurato in maniera oggettiva per verificare l'ipotesi è il cosiddetto "endpoint", o indicatore di esito (vedi capitolo 2). Nel nostro caso, ciò che nello studio si è scelto di misurare e confrontare tra i due gruppi di trattamento è la mortalità ad un anno per tutte le cause.

6.1 La verifica delle ipotesi in medicina

6.1.1 Ipotesi nulla e ipotesi alternativa

Quando conduciamo uno studio scientifico, sia per valutare l'associazione fra esposizione e malattia, che per valutare l'efficacia di un nuovo trattamento, formuliamo un'ipotesi e la sottoponiamo a valutazione. La ricerca scientifica parte infatti da una domanda, alla quale si deve cercare di dare una risposta. Per trovare la risposta si deve trasformare quella domanda di partenza in un'ipotesi, la quale a sua volta deve essere valutata, o meglio, verificata: se ne deve valutare la veridicità confrontandosi con la realtà (dati). L'approccio scientifico che si segue è quello ipotetico deduttivo, che consiste appunto nel formulare (prima di raccogliere i dati!) un'ipotesi scientifica da confrontare con le osservazioni.

L'ipotesi che andiamo valutare è detta ipotesi nulla, ed è solitamente indicata con H0. Mentre verrebbe spontaneo pensare che l'ipotesi formulata sia da confermare (accettare) mediante le osservazioni (dati), in realtà, seguendo l'approccio falsificazionista (una teoria è scientifica se può essere contraddetta, falsificata, da esperimenti scientifici, dai dati) si deve procedere, in un certo senso, al contrario, in maniera controintuitiva. Poiché, in base a questi approcci, l'unica conclusione "certa" alla quale potremo giungere sarà che l'ipotesi formulata è falsa, non sarà possibile accettare, ma solo "dimostrare" falsa l'ipotesi di partenza[1]. È quindi evidente che formulare l'ipotesi affermando che "il trattamento sperimentale è più efficace del controllo" ai nostri fini non può funzionare, visto che la potremo esclusivamente dimostrare falsa, non la potremo mai accettare. Formulare invece l'ipotesi affermando che "il trattamento sperimentale ha la stessa efficacia del controllo" comporta la conseguenza che rifiutare la nostra ipotesi, sulla base delle osservazioni del nostro campione, significa concludere che trattamento e controllo hanno differente efficacia. Impostare il ragionamento seguendo questa logica significa, quindi, che H0 deve essere formulata in termini di assenza di un effetto (esempio, pari efficacia dei due trattamenti a confronto).

Quindi, tornando agli esempi in ambito clinico, l'ipotesi che formuliamo (ipotesi nulla, H_0), è che l'esposizione o l'intervento che stiamo valutando non siano associati all'outcome di interesse, ovvero che i due trattamenti a confronto abbiano pari efficacia, oppure, in contesto eziologico, che il rischio di malattia negli esposti e nei non esposti sia uguale (tradotto in misure di associazione: RR=1, OR=1, HR=1, riduzione del rischio=0). L'ipotesi nulla è valutata rispetto ad un'altra ipotesi, che in statistica è detta ipotesi alternativa (H_1), che afferma: i due trattamenti hanno differente efficacia (alternativa a due code) oppure, ad esempio, uno dei due trattamenti è più efficace dell'altro (alternativa a una coda)[2]. Obiettivo degli studi è quindi riuscire a dimostrare, con un certo livello di sicurezza di non sbagliare, che H_0 è falsa.

Tornando all'esempio precedente, in questo caso l'ipotesi nulla è che la mortalità per ogni causa ad un anno nel gruppo sottoposto a TAVI sia uguale alla mortalità nel gruppo di controllo, e obiettivo dello studio è osservare se la eventuale differenza di mortalità rilevata a un anno fra gruppo TAVI e

1 K.R. Popper Logica della scoperta scientifica. Il carattere autocorrettivo della scienza. Einaudi, 2010

2 L'approccio classico all'inferenza statistica prevede che l'ipotesi alternativa possa essere formulata in modi differenti, in base all'obiettivo dello studio. Parlando degli studi classici in ambito clinico, ad esempio, studi di valutazione dell'efficacia di un farmaco A rispetto al controllo B, mentre H0 ricordiamo deve essere formulata in termini di assenza di efficacia, abbiamo tre differenti possibilità per formulare H1: Efficacia A >Efficacia B; Efficacia A ¹ Efficacia B; Efficacia A > Efficacia B.

gruppo di controllo sia significativamente diversa (sperabilmente minore, ma non necessariamente).

6.1.2 Errore di primo e secondo tipo e p-value

Sappiamo bene che reclutando un campione di 358 pazienti, non l'intera popolazione, è sempre possibile osservare una mortalità differente fra i due gruppi per puro effetto del caso, non per un'effettiva differenza di efficacia fra i due trattamenti. È quindi necessario stabilire a partire da quale differenza osservata potremo sentirci ragionevolmente sicuri nel concludere con un rifiuto dell'ipotesi nulla. I risultati dello studio mostrano che la mortalità per ogni causa ad un anno nel gruppo trattato con TAVI è risultata inferiore rispetto alla mortalità nel gruppo di controllo (30.7% vs 49.7%).

Dai dati pubblicati possiamo costruire la tabella 2x2 e calcolare poi le misure di associazione.

		Decesso ad 1 anno		
		Sì	**No**	**Totale**
TAVI	**Sì**	55	124	179
	No	89	90	179
	Totale	**144**	**214**	**358**

Tabella 6.1 Decesso a un anno in 358 pazienti con stenosi aortica severa randomizzati a impianto di valvola transcatetere per via femorale (*TAVI*) o trattamento tradizionale (terapia medica o valvuloplastica con palloncino).

Rischio decesso TAVI=55/179=30.7%

Rischio decesso controlli=89/179=49.7%

RR=(55/179)/(89/179)=0.62

Questo significa che la TAVI riduce davvero il rischio di decesso nella popolazione dei malati di nostro interesse? Oppure la differenza di mortalità che abbiamo osservato è da attribuire ad altri fattori, casuali, diversi dal trattamento?

Per poter rispondere a questa domanda dobbiamo valutare:

– che lo studio sia stato condotto in modo metodologicamente corretto, senza bias;
– che un RR di 0.62 sia sufficientemente diverso da 1 (RR dell'ipotesi nulla): "sufficientemente diverso" significa che la probabilità che questa differenza sia dovuta al caso sia sufficientemente bassa (di solito inferiore al 5%);
– che il valore vero di RR abbia comunque una rilevanza pratica.

In questo capitolo ci occuperemo del terzo punto.

Se un effetto del caso portasse ad una conclusione erronea che il trattamento è efficace, avremmo un errore che è definito di I tipo; se viceversa, portasse ad una conclusione erronea che non c'è effetto, avremmo un errore di II tipo.

- **Errore di I tipo**: rifiutare erroneamente l'ipotesi nulla, considerando più efficace la TAVI quando è invece di pari efficacia rispetto al trattamento "standard". Tale errore è assimilabile ad un falso positivo. Generalmente si indica con α la probabilità di commettere tale errore.
- **Errore di II tipo**: non rifiutare erroneamente l'ipotesi nulla, ovvero concludere in maniera erronea che non abbiamo elementi per dimostrare che TAVI e trattamento standard hanno differente efficacia. Tale errore è assimilabile ad un falso negativo. Generalmente si indica con β la probabilità di commettere tale errore.

La statistica ci aiuta a valutare l'effetto del caso attraverso il famigerato p-value. Il p-value non è la probabilità che i risultati di uno studio siano "veri", bensì la probabilità che, posto che l'ipotesi nulla fosse vera (e lo studio non fosse influenzato da bias o confondenti), la differenza osservata tra i risultati (o una differenza ancora più estrema di quella osservata) sia dovuta al caso. Nel nostro esempio, risulta RR = 0.62 (minor mortalità nel gruppo TAVI) con $p<0.001$, che significa che, se la TAVI non fosse efficace nel ridurre la mortalità, noi osserveremmo un RR di 0.62 o inferiore in media meno di 1 volta ogni 1000 studi effettuati nelle stesse condizioni ($p<0.001$). Il p-value non dà informazioni sulla probabilità che HR sia effettivamente 0.62. E il p-value, purtroppo, non dà nemmeno informazioni definitive sulla verità (falsità) dell'ipotesi nulla.

Generalmente si rifiuta H_0 di fronte a risultati la cui probabilità di essere osservati per effetto del caso sia inferiore al 5% (sì, è proprio lui, il mitico $p<0.05$, l'unica cosa che, forse, vi ricordate del corso di statistica), cioè quando, decidendo di rifiutare l'ipotesi nulla, si corre un rischio di errore di I tipo (risultato dello studio falsamente positivo) inferiore al 5%. La motivazione per cui solitamente si utilizza la soglia del 5% per la significatività ha origini storiche[3]: fu per primo RA Fisher, considerato da molti il padre della statistica moderna, ad introdurre una soglia di questa entità, ritenendosi soddisfatto di sbagliare (errore I tipo) non più di 1 volta su 20. E, da allora, la maggior parte dei ricercatori utilizza il 5% quale soglia per la significatività statistica. In realtà, ci sono studi in cui è opportuno ricorrere a valori soglia più bassi, ad esempio 1% e anche minori. E, addirittura, c'è chi ha proposto di utilizzare, in ambiti particolari, il 10%[4].

L'errore di II tipo riguarda, invece, la probabilità beta (fissata solitamente al 20% o al 10%) che la non significatività dei risultati ottenuti sia dovuta al caso. Se, ad esempio, volessimo valutare se gli svedesi hanno più facilmente i capelli biondi degli italiani e prendessimo casualmente 5 soggetti svedesi e 5 italiani, nonostante sappiamo che in effetti gli svedesi hanno una prevalenza maggiore di capelli biondi, la differenza tra i risultati osservati non sarebbe statisticamente

3 Fisher RA. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 1926; 33:503-513.

4 Costantino, G. & Casazza, G. 2014. Randomization, ethics and clinical research. Internal and emergency medicine, 9, 797-798

significativa, ma motivo di ciò risiede solo nel fatto che abbiamo arruolato troppi pochi soggetti.

Il complemento a 1 di beta (1-β) rappresenta la potenza dello studio (che quindi è generalmente fissata all'80% o al 90%). La potenza di uno studio indica la probabilità di vedere una differenza giudicata significativa tra due trattamenti quando questa esiste. A parità di altre condizioni, essa è tanto maggiore quanto è più grande il campione e quanto più grande è la differenza vera. In termini pratici, dire che uno studio ha una potenza dell'80% equivale a dire che, se osserviamo una certa differenza predefinita (la minima rilevante dal punto di vista clinico) tra due gruppi (ad esempio terapia meglio del placebo), abbiamo l'80% di probabilità di riconoscerla come statisticamente significativa in quello studio.

Il valore di β (rischio di studio falsamente negativo) è, come abbiamo detto, solitamente fissato al 20% perché in genere si preferisce essere "conservativi" ed evitare di commercializzare un nuovo farmaco non efficace, anche a discapito del fatto che si possano giudicare inefficaci farmaci in realtà efficaci. Avrete notato la similitudine fra i concetti appena espressi (rischio di errore di I e II tipo) con sensibilità e specificità trattati nel capitolo sull'accuratezza diagnostica. Se volessimo riassumere, la potenza dello studio può essere vista come la sensibilità, mentre il complemento dell'errore di I tipo (vale a dire non rifiutare un'ipotesi vera) può essere assimilata alla specificità.

Ovviamente, si può anche decidere di abbassare sia α che β, al costo di dover selezionare un campione più numeroso; infatti α e β entrano in gioco per il calcolo della numerosità del campione di un certo studio per dimostrare un dato endpoint. Gli studi che non dichiarano a priori i valori prescelti di α e β sono deboli.

6.2 Intervalli di confidenza

Dal momento che gli studi ci permettono di ottenere una stima campionaria (cioè un risultato che è solo espressione del campione osservato), è utile esprimere i risultati fornendo un intervallo, ovvero un range, che permetta di "esportare" il risultato ottenuto alla popolazione generale: l'intervallo di confidenza (confidence interval, CI). Il calcolo dell'intervallo di confidenza, la cui ampiezza definisce l'imprecisione di stima per un determinato grado di certezza di essere nel vero (95% in genere), dipende fondamentalmente dal numero di soggetti inclusi nello studio: quando aumenta la dimensione del campione, l'ampiezza si riduce e aumenta la precisione della nostra stima.

Ritornando al nostro esempio, il rischio di decesso nei pazienti sottoposti a TAVI, RR = 0.62, è stima di un valore vero della popolazione che, al 95%, dobbiamo considerare incluso fra i due estremi dell'intervallo di confidenza 0.47 e 0.81. Ciò significa che è plausibile che il vero valore di RR nella popolazione da cui provengono i pazienti inclusi nello studio sia come minimo 0.47 come

massimo 0.81. Non è impossibile, ma riteniamo poco plausibile, sulla base dei dati a nostra disposizione, che il vero valore di RR sia maggiore di 0.81 o minore di 0.47. L'interpretazione dell'IC al 95% potrebbe essere fatta nel seguente modo: gli autori dello studio concludono che il vero valore di RR per TAVI vs controllo è compreso fra 0.47 e 0.81. La probabilità che la loro conclusione sia vera (livello di confidenza) è del 95%. Se fra i due valori estremi dal punto di vista clinico pratico giudichiamo non ci sia differenza rilevante, possiamo concludere che la qualità della stima ottenuta è buona, se invece giudicassimo di scarso interesse pratico una riduzione del 19% e di notevole interesse una riduzione del 53%, dovremmo considerare lo studio non conclusivo dal punto di vista pratico.

Si può quindi intuire come p-value e IC siano due modi diversi di rappresentare l'informazione fornita dal campione. Ricordiamo che l'ipotesi nulla dello studio era che non ci fosse differenza di mortalità fra TAVI e controllo (RR=1). Dai risultati abbiamo visto che RR=1 non è un valore compreso nell'IC al 95%, in base ai dati ottenuti dallo studio è quindi poco plausibile che il vero valore di RR possa essere pari ad 1. Possiamo allora, con ragionevole certezza, rifiutare l'ipotesi nulla e affermare con la stessa ragionevolezza che possa essere vera una qualsiasi ipotesi alternativa (per il valore di RR) fra quelle contenute nell'intervallo di confidenza al 95%.

Vediamo di rendere più chiaro questo concetto con un esempio grafico riportato in Figura 6.1.

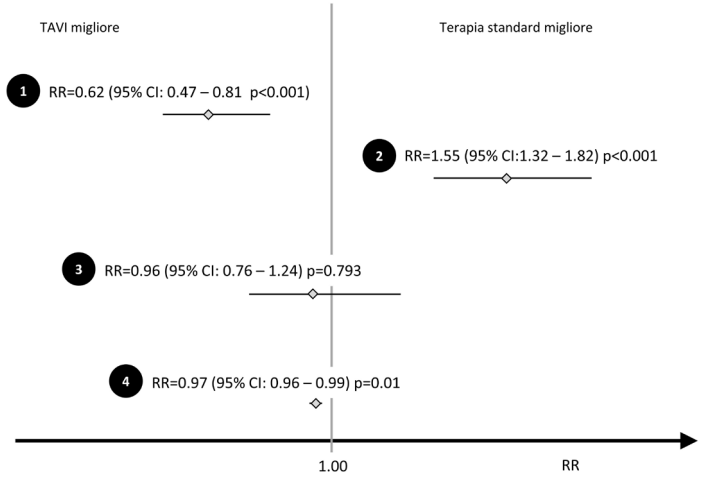

Figura 6.1 Rappresentazione grafica di 4 scenari di ipotetici studi che confrontano il rischio di decesso ad 1 anno in pazienti sottoposti a TAVI vs pazienti sottoposti a terapia standard. Il caso n°1 è tratto da *NEJM 2010;363:1597-1697*. I casi n° 2-3-4 sono stati costruiti con dati di fantasia.

Il caso n°1 rappresenta i risultati dello studio che abbiamo analizzato, in cui RR=0.62 e il limite superiore dell'IC al 95% è minore di 1: possiamo quindi affermare che la TAVI è più efficace della terapia standard, dato che il p-value minore di 0.05 indica che l'osservazione fatta è da giudicarsi statisticamente significativa. Guardando l'IC non possiamo stabilire quale sia l'esatto valore vero di *p*, ma sappiamo che è minore di 0.05.

Nel caso n°2 RR = 1.55, il limite inferiore dell'IC al 95% è maggiore di 1: possiamo quindi affermare che la TAVI è significativamente meno efficace della terapia standard, con p-value minore di 0.05. Anche in questo caso, guardando l'IC non possiamo sapere il valore esatto di *p*, ma sappiamo comunque che è minore di 0.05.

Nel caso n°3, RR=0.97: in questo caso l'IC al 95% comprende il valore 1. Semplicemente guardando il grafico, possiamo concludere che abbiamo osservato nel gruppo TAVI una riduzione di mortalità non sufficiente a farci escludere l'ipotesi nulla, ovvero non statisticamente significativa. Analogamente ai due casi precedenti, non possiamo sapere il valore esatto di *p*, ma sappiamo comunque che è sicuramente maggiore di 0.05.

Come dicevamo in precedenza, l'ampiezza dell'IC ci dà un'idea della precisione di una stima. Guardiamo il caso n°3 ed il caso n°4: possiamo notare che in entrambi i casi la stima di RR è 0.97. Tuttavia, mentre nel n°3 l'IC comprende RR=1, nel caso n°4 il limite superiore dell'IC è minore di 1, e quindi il RR osservato risulta statisticamente significativo. Ciò può essere dovuto al fatto che nel caso n°3 è stato arruolato un minor numero di pazienti, e di conseguenza la stima dello stesso valore (RR=0.97) è meno precisa. Inoltre, sempre nel caso n°3, il valore vero di RR potrebbe essere sia 0.76 (quindi una riduzione di mortalità del 24% nel gruppo sottoposto a TAVI) sia 1.24 (un aumento di mortalità del 24%).

Infine, possiamo osservare come un risultato statisticamente significativo non sia obbligatoriamente anche rilevante dal punto di vista clinico: nel caso n°4, la riduzione della mortalità è compresa tra l'1 e il 4% (RR=0.97, IC 0.96-0.99). Tuttavia, una terapia che riduce solo del 3% la mortalità (e ha comunque delle possibili complicanze) non appare utile in questo contesto. La significatività statistica è stata raggiunta solo grazie all'elevata numerosità campionaria. Quindi significatività statistica e rilevanza clinica di un risultato sono due concetti da tenere separati.

Da tutto ciò che abbiamo detto emerge che la non significatività di uno studio effettuato con ipotesi nulla di uguaglianza non permette di asserire che i due trattamenti siano di pari efficacia, bensì unicamente che non si è osservata una differenza campionaria di efficacia statisticamente significativa ovvero tale da permetterci di rifiutare l'ipotesi nulla. Per utilizzare una frase cara anche a David Sackett: «Assenza di evidenza non significa evidenza dell'assenza»[5] .

5 Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995 Aug 19;311(7003):485

In appendice trovate le formule per calcolare gli intervalli di confidenza al 95% delle più comuni misure di epidemiologia clinica.

6.3 Intervalli di confidenza o p-values?

Abbiamo quindi visto come intervalli di confidenza e p-values forniscono informazioni in parte sovrapponibili sulla plausibilità di una ipotesi, alla luce dei dati raccolti con il nostro campione. Partendo da due punti di vista differenti, sia i p-values, sia gli intervalli di confidenza ci permettono di capire se l'ipotesi nulla è da rifiutare o meno. Tuttavia, vi sarà a questo punto chiaro che i p-values forniscono un'immagine parziale dei risultati ottenuti con il nostro studio: si limitano a dirci se l'ipotesi è da rifiutare o meno, ma, ad esempio, nel caso di un risultato non statisticamente significativo, non ci permettono di capire se la mancanza di significatività è verosimilmente da attribuire ad una limitata efficacia, piuttosto che ad una bassa potenza dello studio (cioè studio troppo piccolo).

Gli intervalli di confidenza, invece, oltre a portarsi dietro un'informazione analoga a quella fornita dai p-values relativa al rifiutare o meno l'ipotesi nulla, ci permettono una lettura più completa dei risultati di uno studio. Gli intervalli di confidenza ci consentono di effettuare valutazioni articolate sulla rilevanza clinica dei risultati, tenendo in considerazione gli estremi dell'intervallo, e valutando le possibili ricadute cliniche. Ad esempio, nello studio ANDROMEDA-SHOCK (*JAMA 2019;32:654-664*) gli autori hanno confrontato il dosaggio dei lattati rispetto al refill capillare nella gestione del paziente con sepsi. I risultati dell'endpoint primario (mortalità) non sono significativi (HR[6]=0.75; p<0.06). Se guardiamo però l'intervallo di confidenza al 95% dell'HR, questo va da 0.55 a 1.02. Clinicamente potremmo interpretare questi risultati come se, nella migliore delle ipotesi, la mortalità potrebbe essere dimezzata (HR=0.55 è interpretabile come riduzione di mortalità del 45%), oppure nella peggiore delle ipotesi, aumentata di circa il 2%. Vista la mortalità dei pazienti con sepsi, potremmo decidere di utilizzare la strategia gestionale suggerita nonostante il risultato dello studio non sia significativo, potremmo cioè essere disposti ad accettare un rischio di un piccolo incremento di mortalità soppesandolo con un beneficio di una possibile grande riduzione della stessa.

In conclusione, il suggerimento è di guardare sempre gli intervalli di confidenza e ragionare sui valori dei due estremi, cercando di capire quali sarebbero

6 Gli autori di questo studio hanno utilizzato un particolare approccio statistico che fornisce come risultato una misura di confronto, l'hazard ratio (HR), che tratteremo nel capitolo 7. Ai fini della comprensione dell'esempio riportato, è sufficiente che sappiate che HR è una misura di confronto che si interpreta esattamente come un rischio relativo (RR). In questo caso, considerando HR=0.75 significa che il dosaggio dei lattati in confronto con il refill capillare riduce il rischio di mortalità del 25%.

le conseguenze cliniche nel caso fosse vero il valore del limite superiore e quali le conseguenze nel caso fosse vero il valore del limite inferiore, effettuando una sorta di analisi decisionale in scenari estremi.

6.4 Numerosità del campione

Abbiamo visto come dalla numerosità del campione dipende la precisione di una stima, e quindi la capacità di uno studio di dimostrare una differenza di efficacia tra due trattamenti.

Come si fa a decidere quanti soggetti arruolare?

Uno studio dovrebbe essere della dimensione giusta per permettere al ricercatore di raggiungere l'obiettivo che si è prefissato. Vale a dire, dovrebbe avere un numero di pazienti idoneo per essere in grado di individuare come statisticamente significativa una differenza clinicamente rilevante, ad esempio di mortalità, fra il gruppo TAVI ed il gruppo di controllo, quando questa differenza è realmente presente.

Generalmente si pensa che, quanto più è elevata la numerosità del campione, tanto migliore è lo studio. In realtà, questa affermazione non sempre si rivela corretta. Al contrario, si dovrebbe preferire una numerosità più piccola possibile, compatibilmente con il contesto e con l'obiettivo dello studio. Infatti, per avere campioni di dimensioni elevate potrebbe essere necessario protrarre lo studio per parecchi anni, per cercare di arruolare più pazienti, impedendo l'applicazione degli eventuali risultati positivi dello studio alla popolazione oppure sottoponendo i pazienti arruolati a un trattamento non utile e potenzialmente nocivo e costoso, se il trattamento non fosse efficace.

Da un punto di vista statistico, avere tanti pazienti significa avere tante informazioni sul fenomeno che si vuole studiare. Avere un campione la cui numerosità coincide con quella della popolazione reale che si intende studiare ci permetterebbe di eliminare del tutto la variabilità casuale dovuta al campionamento. Ma nella realtà la variabilità c'è e dobbiamo solo cercare di ridurla quel tanto che basta per raggiungere i nostri obiettivi.

Dato che generalmente le popolazioni da cui provengono i pazienti inclusi nei nostri studi sono di dimensioni molto elevate (alcune volte non conosciamo nemmeno la dimensione della nostra popolazione di riferimento), dobbiamo per forza di cose tenere conto della variabilità dovuta a campionamento. Questa variabilità campionaria, oltre a dipendere dalla variabilità intrinseca del fenomeno che stiamo studiando, è in stretta relazione con la dimensione campionaria: diminuisce quando aumenta la dimensione campionaria. Disegnare uno studio di dimensione ottimale ci permette di avere stime (di OR, RR, HR) abbastanza precise (o meglio, stime con una precisione sufficiente) e ci permette, inoltre, di saggiare la nostra ipotesi in maniera efficiente.

La determinazione della dimensione campionaria è fatta tenendo conto degli elementi che definiscono il sistema di ipotesi (nulla ed alternativa) come definite in precedenza. Facciamo un passo indietro. Nei paragrafi precedenti abbiamo definito le ipotesi e gli errori di primo e di secondo tipo. Ipotesi statistiche che dovrebbero sempre essere formulate prima di iniziare uno studio, in quanto sono la trasposizione numerica del quesito clinico che ci spinge ad effettuare lo studio.

Gli autori dello studio sulla TAVI affermano: «We estimated that with a sample of 350 patients, the study would have at least 85% power to show the superiority of TAVI over standard treatment with respect to the primary endpoint, assuming that 1-year mortality would be 37.5% in the standard-treatment group and 25% in the TAVI group». Vediamo, elemento per elemento, quali sono i fattori in gioco.

Innanzitutto, si deve avere un'idea della misura dell'endpoint nel gruppo di controllo (gli autori assumono mortalità del 37.5%). In secondo luogo, si deve definire quale differenza minima si è interessati a individuare fra i due gruppi nello studio (differenza minima clinicamente rilevante: 12.5% nel caso TAVI, ottenuto dalla differenza 37.5%-25%): vale a dire, vorremmo essere in grado di ottenere un risultato statisticamente significativo in tutti i casi in cui tale differenza fosse del 12.5% o maggiore.

Per capire l'effetto della differenza minima clinicamente rilevante sulla dimensione del campione, partendo da una mortalità del 37.5% nel gruppo di controllo, avremmo bisogno di moltissimi pazienti (qualche migliaio per gruppo), per riuscire ad individuare una riduzione di mortalità del 2.5% (37.5% gruppo di controllo vs 35% gruppo TAVI). Al contrario, avremmo bisogno di molti meno pazienti (qualche decina per gruppo) per riuscire ad individuare una riduzione di mortalità del 30% (da 37.5% a 7.5%).

Schematizzando, quindi, la dimensione campionaria dipende da:

- misura dell'endpoint nel gruppo di controllo (solitamente desunta da dati di letteratura o da studi preliminari);
- differenza minima fra nuovo trattamento e controllo, clinicamente rilevante, che vogliamo essere in grado di individuare con lo studio (decisa dal ricercatore: variazione minima dell'endpoint che rende "appetibile" il nuovo trattamento);
- potenza dello studio: aumentando la potenza desiderata (come abbiamo visto, valori tipici 80%-90%), aumenta il numero di pazienti che dobbiamo arruolare;
- rischio di errore di I tipo: generalmente è fissato al 5%; volendo effettuare uno studio con un rischio di errore più basso, il numero di pazienti richiesto aumenta.

Combinando questi quattro elementi mediante le formule appropriate per il tipo di endpoint prescelto e per le misure di rischio di interesse (OR, RR, HR),

si ottiene il numero minimo di pazienti che dobbiamo arruolare. Molte volte si pianifica l'arruolamento di un numero maggiore di pazienti, rispetto al minimo necessario, al fine di tenere conto dell'eventuale perdita (drop-out), casuale, di pazienti durante lo svolgimento dello studio.

Il calcolo della numerosità campionaria ci permette, quindi, di definire la dimensione minima di cui abbiamo bisogno per riuscire a verificare il sistema di ipotesi che abbiamo pianificato. Uno studio sottodimensionato (che prevede l'arruolamento di un numero di pazienti inferiore a quello ottimale) non dovrebbe neppure iniziare, in quanto inutile spreco di risorse. Ovviamente, studi descrittivi e studi esplorativi (che non prevedono verifica di ipotesi) potrebbero essere condotti anche senza definizione della dimensione campionaria secondo le modalità esposte in precedenza.

Tra i parametri che abbiamo presentato, l'unico valore che dovrebbe essere definito in maniera oggettiva (basato su dati di letteratura) è l'incidenza dell'endpoint nel gruppo di controllo. Questo potrebbe a posteriori rivelarsi diverso da quanto previsto (ad esempio, nello studio TAVI la mortalità campionaria osservata nel gruppo di controllo è stata del 49.7%, molto alta per pensare che sia una stima della mortalità ipotizzata del 37.5%) e influenzare la significatività e la generalizzabilità dei risultati.

La differenza minima clinicamente rilevante è l'elemento più delicato nella definizione del sample size. Indica la variazione (esempio, riduzione di mortalità, riduzione incidenza di infarto, incremento di guarigioni) che riteniamo importante da un punto di vista clinico: al termine dello studio saremo disposti a utilizzare nella pratica clinica il nuovo trattamento solo se avrà dimostrato una differenza di endpoint rispetto al controllo almeno pari (o superiore) a quel valore da noi ritenuto rilevante. Per chiarire il concetto, tornando all'esempio TAVI, gli autori hanno definito come clinicamente rilevante una riduzione assoluta di mortalità di 12.5% ed hanno dimensionato lo studio per essere in grado di individuare come statisticamente significativa una riduzione pari appunto ad almeno 12.5%. Ciò significa che per noi riduzioni di mortalità più basse di valore (ad esempio riduzione del 7%) non sono ritenute rilevanti dal punto di vista clinico. Oppure, la scelta è stata determinata per "risparmiare" sulla numerosità campionaria, rischiando la non significatività di risultati anche potenzialmente rilevanti.

6.5 Analisi per sottogruppi

Spesso negli studi il confronto tra braccio di intervento e braccio di controllo viene eseguito non solo sull'intera casistica, ma anche entro alcune sottopopolazioni di pazienti con caratteristiche particolari (ad esempio, diabetici o non diabetici, classi di età, causa ischemica o non ischemica dello scompenso) per indagare un possibile diverso effetto di un intervento in categorie specifiche

di pazienti, in base a presupposti fisiopatologici più o meno ben fondati. Analogamente a quanto detto per gli endpoint secondari, i risultati nei diversi sottogruppi, anche se statisticamente significativi, non sono probanti, cioè non possono essere automaticamente utilizzati per prendere decisioni cliniche. Anche in questo caso, l'analisi è da considerarsi esplorativa: prima di essere trasferibile nella pratica clinica, dovrà essere verificata in uno studio successivo disegnato ad hoc per saggiare quella particolare ipotesi.

Inoltre, come per gli endpoint, i sottogruppi devono essere definiti a priori nel protocollo di studio sulla base di una domanda clinica rilevante. Un'eventuale definizione a posteriori, con lo scopo di trovare almeno un sottogruppo nel quale il risultato risulti significativo, sarebbe totalmente scorretta: più aumentiamo il numero di sottogruppi, più facilmente troveremo risultati significativi per puro caso. Ad esempio, nello studio ISIS2, che analizzava l'utilità dell'Aspirina nei pazienti con infarto del miocardio, effettuando un'analisi per sottogruppi, risultava che l'Aspirina era utile in tutti i pazienti tranne in quelli nati sotto il segno dei Gemelli! (*Lancet 1988;2(8607):439-69*). La definizione a priori dei sottogruppi in modo trasparente, e secondo un chiaro razionale, è attualmente assicurata dal fatto che molte tra le riviste scientifiche più importanti richiedono che il protocollo di un trial sia registrato prima di poter cominciare lo studio (vedi ad esempio http://clinicaltrials.gov/). Non sempre, però, anche in questo modo, si riesce a verificare che tutte le analisi per sottogruppi siano state dichiarate in anticipo.

Infine, tutte le volte che in uno studio si fanno molte analisi statistiche (siano analisi di sottogruppo, oppure analisi su numerosi endpoint secondari) si dovrebbe correggere il valore dell'errore di primo tipo utilizzato per stabilire la significatività (il famoso $p<0.05$) riducendolo in funzione del numero di test statistici effettuati. Ad esempio, l'approccio semplice suggerito da Bonferroni consiste nello stabilire, per uno studio che preveda 10 test differenti sui dati, il livello di significatività statistica sia stabilito a $p<0.005$ (cioè, 0.05/10).

Punti chiave

- ✓ Quando vogliamo effettuare uno studio clinico (ad esempio,
- ✓ di efficacia di un farmaco) dobbiamo formulare due ipotesi, un'ipotesi nulla (H0: il farmaco ha la stessa efficacia del farmaco di confronto) e un'ipotesi alternativa (H1). Il nostro obiettivo è riuscire confutare l'ipotesi nulla.
- ✓ α e β indicano le probabilità di commettere errore rispettivamente di I tipo (concludere che il nuovo trattamento è più efficace del confronto quando i due trattamenti sono di pari efficacia) e di II tipo (concludere che il nuovo trattamento non è più efficace del vecchio, quando in realtà lo è).
- ✓ Il complemento a 1 di β (1- β) è la potenza di uno studio, probabilità di osser-

vare un risultato statisticamente significativo quando la supposta differenza esiste.

- ✓ Il p-value indica la probabilità che uno studio risulti falsamente positivo; di solito si accetta che tale probabilità possa essere al massimo il 5% (p-value <0.05).
- ✓ L'ampiezza dell'intervallo di confidenza di una stima indica l'imprecisione della stima: intervallo stretto stima precisa, intervallo ampio stima imprecisa. Se l'intervallo di confidenza comprende il valore di indifferenza tra i trattamenti, non possiamo dichiarare falsa l'ipotesi nulla (ma nemmeno dichiarare vera l'ipotesi alternativa).
- ✓ Sia intervallo di confidenza, sia il p-value ci permettono di capire se un'ipotesi è da rifiutare. Ricordiamoci però che l'intervallo di confidenza ci dà informazione molto più completa rispetto al solo p-value.
- ✓ La numerosità del campione di uno studio viene calcolata in base a:
  - incidenza dell'endpoint nel gruppo di controllo;
  - differenza minima clinicamente rilevante (la rilevanza è decisa da noi) che vogliamo essere in grado di individuare;
  - potenza dello studio (generalmente fissata all'80% o 90%);
  - rischio di errore di primo tipo (generalmente fissato al 5%).
- ✓ La numerosità campionaria deve essere calcolata a priori, in fase di definizione del protocollo.

Bibliografia consigliata

Davidoff F. Standing statistics right side up. *Ann Intern Med.* 1999;130:1019.

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed).* 1986;292(6522):746-50.

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995.

Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med.* 1994;121:200.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 2016 Apr;31(4):337-50.

Grimes DA, Schulz KF. Surrogate End Points in Clinical Research: Hazardous to Your Health. *Obstet Gynecol.* 2005;105:1114–8.

Grimes DA, Schulz KF Sample size calculations in randomised trials: mandatory and mystical. *Lancet.* 2005; 365:1348–53.

Hawkins AT, Samuels LR. Use of Confidence Intervals in Interpreting Nonstatistically Significant Results. *JAMA*. 2021;326(20):2068-2069.

7. Introduzione ai metodi di analisi statistica

Abbiamo visto in termini molto generali il procedimento da seguire per verificare ipotesi scientifiche. Ma nella pratica come facciamo a stabilire se un'ipotesi (di assenza di effetto) sia da rifiutare o meno? Si deve ricorrere a vari test statistici, che sono in sostanza procedure particolari che cercano di stabilire, in termini probabilistici, quanto sia verosimile osservare campioni come quelli che abbiamo selezionato, quando una certa ipotesi nulla è vera.

7.1 Quando l'endpoint è un evento

Ad esempio, se nella realtà il trattamento non ha effetto (RR=1) e nel nostro campione abbiamo osservato un certo effetto (ad esempio, leggendo quanto riportato nell'articolo TAVI: decesso ad 1 anno, 30.7% nel gruppo TAVI vs 49.7% nel gruppo terapia standard, RR=0.62, p<0.001) possiamo rifiutare l'ipotesi che RR=1? In questa situazione, per valutare la significatività statistica delle differenze osservate, potremmo utilizzare il test *z*, oppure il test *chi-quadrato*, oppure ancora il *test esatto di Fisher*. Infatti, nei metodi, gli autori dello studio riportano che per confrontare variabili categoriche è stato utilizzato il test esatto di Fisher. Se volessimo invece confrontare le sensibilità (o le specificità) di due differenti test diagnostici a cui abbiamo sottoposto gli stessi pazienti (disegno intraindividuale), per vedere quale dei due test è migliore, dovremmo utilizzare il *test di McNemar*. Quindi, per ogni situazione abbiamo uno o più test appropriati dal punto di vista statistico. Nella figura 1 è riportata una sintesi dei test statistici più appropriati nelle diverse situazioni.

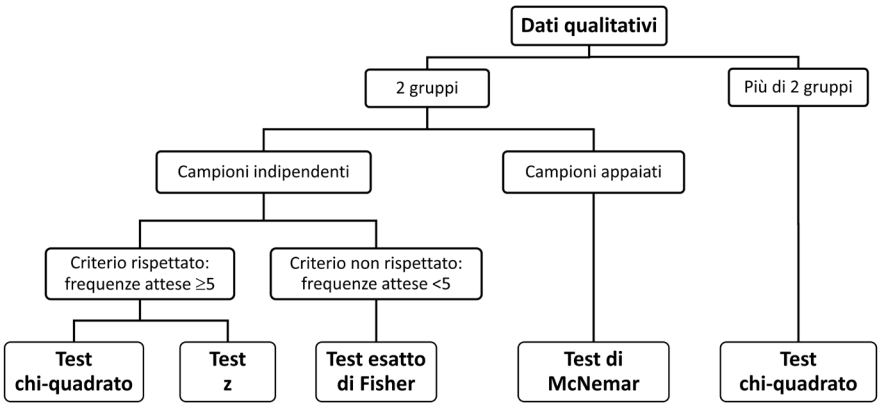

Figura 7.1 Algoritmo di analisi per dati qualitativi.

7.2 Quando l'endpoint è una variabile quantitativa

In uno studio si è valutato l'effetto dell'ossicodone rispetto alla morfina per l'analgesia post-cesareo. Sono state arruolate 77 donne, 38 randomizzate al gruppo ossicodone, 39 al gruppo morfina e si è valutato il dolore a riposo dopo 6 ore, mediante scala VAS (da 0 a 10). È risultato che le donne trattate con ossicodone hanno mostrato un minor dolore (3.80±1.52) rispetto alle donne trattate con morfina (4.96±1.49), e questa differenza è risultata statisticamente significativa (p=0.002). Quale test statistico potrebbe essere stato utilizzato dagli autori dello studio per valutare la significatività della differenza di dolore medio fra i due gruppi? Essendo il dolore misurato su scala quantitativa il test appropriato è il test *t di Student* per dati indipendenti, oppure in alternativa l'equivalente *test non parametrico di Wilcoxon* (somma dei ranghi). La scelta fra i due test va fatta valutando, come già accennato quando abbiamo parlato delle analisi descrittive riportate nella Tabella 1 di uno studio scientifico, la distribuzione normale o meno dei dati. Se i valori di VAS hanno distribuzione approssimativamente normale, possiamo utilizzare il test t di Student, in caso contrario dovremo ricorrere al test non parametrico. Approccio che è stato seguito in uno studio in cui si è valutato l'effetto del desametasone, rispetto al placebo (in aggiunta a terapia antibiotica), sulla durata della degenza in pazienti ricoverati per polmonite. La durata mediana di degenza è stata pari a 6.5 giorni (IQR, 5.0–9.0) per il gruppo desametasone contro i 7.5 giorni (IQR, 5.3–11.5) per il gruppo placebo. Questa differenza è risultata statisticamente significativa… p=0.048. Avendo solitamente la durata di degenza una distribuzione asimmetrica, gli autori hanno giustamente scelto di riportare i dati mediani, invece che medi, ed hanno utilizzato per il confronto fra i due gruppi il *test non parametrico di Mann-Whitney*, in sostituzione del test t di Student. Lasciando per un momento da parte la scelta del test statistico, avrete sicuramente notato un p-value di valore molto vicino alla soglia 0.05 ed un effetto di riduzione di 1 giornata. Questo significa che il trattamento riduce, rispetto al placebo, la durata della degenza in misura statisticamente significativa. Questa riduzione è anche clinicamente rilevante? Il test statistico non ci fornisce risposta a questa domanda, la rilevanza clinica di un risultato deve essere valutata pensando al contesto clinico nel quale andremo a utilizzare i risultati dello studio e valutando le conseguenze nella pratica.

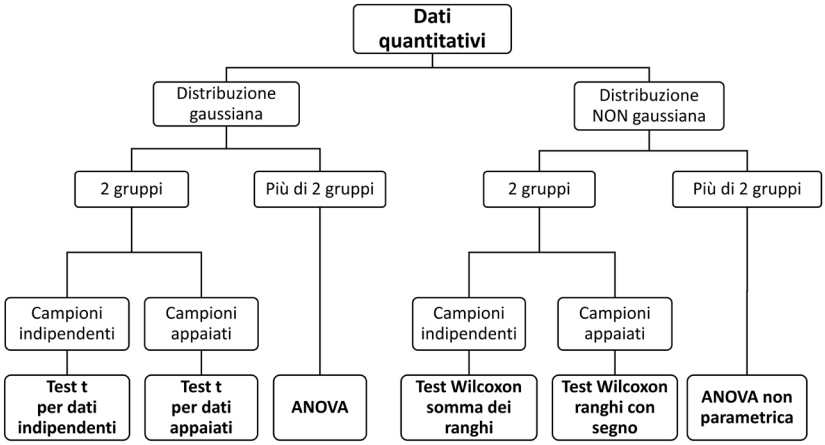

Figura 7.2 Algoritmo di analisi per dati quantitativi.

7.3 Modelli di analisi più popolari

Infine, in moltissime situazioni, l'approccio migliore per l'analisi dei dati richiede l'uso di modelli statistici per effettuare le cosiddette analisi di regressione, vale a dire analisi statistiche in cui si valuta l'associazione fra uno o più fattori indipendenti (o cause, in alcuni contesti indicati anche come predittori) ed una variabile cosiddetta dipendente (o effetto). In base al tipo di variabile dipendente che abbiamo a disposizione, si possono avere il modello di regressione lineare (variabile continua), regressione logistica (variabile dicotomica, evento, ad esempio decesso) o regressione di Cox (variabile dicotomica ma si analizza tempo al verificarsi dell'evento). Una ulteriore distinzione ci permette di parlare di regressione univariata (una sola variabile indipendente) o regressione multivariata (più di una variabile indipendente). I modelli di regressione multivariata sono molto utili quando si vuole tenere conto dell'effetto di eventuali confondenti.

7.3.1 Regressione logistica

La regressione logistica è condotta mediante un particolare modello statistico in cui valutiamo la relazione esistente fra una o più cause (variabili indipendenti: esposizione a fattori di rischio, caratteristiche cliniche/demografiche individuali, trattamenti, etc) e un effetto (variabile dipendente: evento di interesse, quale il decesso, il ricovero, l'infarto, lo stroke, etc). In sostanza, ricorrendo a particolari modelli matematici riusciamo a stimare l'effetto che i fattori considerati hanno sull'odds (quindi "rischio") di evento. Cioè, con questo approccio siamo

in grado, ricorrendo a formule, di stimare (predire) il rischio di evento individuale (variabile dipendente, soggetto i-esimo) partendo da alcune caratteristiche individuali (predittori o confondenti, variabili indipendenti). Consideriamo, ad esempio, di voler valutare l'associazione fra trattamento e decesso. Riuscendo a stimare il rischio di decesso nei trattati e nei non trattati, siamo anche in grado di ottenere una stima del rischio relativo. In realtà, visto che, come accennato sopra, stimiamo gli odds di decesso nei trattati e nei non trattati, l'associazione fra trattamento e decesso potrà essere valutata mediante stima di un odds ratio invece che di un rischio relativo. Ma la sostanza in termini di interpretazione non cambia. Non entreremo nel dettaglio di come sia possibile, a partire da una formula, che è molto simile all'equazione di una retta, ottenere la stima di un odds ratio. Ci focalizzeremo sull'interpretazione dei risultati. Per chi non si fida… rimandiamo ad alcuni tutorial reperibili in letteratura e riportati nei suggerimenti bibliografici al termine del capitolo.

Zhou et al (*Lancet 2020; 395: 1054–62*) hanno condotto uno studio osservazionale per valutare l'associazione fra alcuni fattori individuali (parametri clinici e di laboratorio) ed il rischio di decesso ospedaliero in pazienti ricoverati per Covid-19. Dalle analisi preliminari, l'età anziana, la presenza di malattie cardiovascolari, valori elevati di SOFA score erano alcune fra le variabili che sembravano differenziare i deceduti dai non deceduti. Per valutare l'associazione di tutti questi (ed altri) fattori individuali con la mortalità ospedaliera, gli autori hanno condotto un'analisi di regressione logistica, che in questa situazione è forse la modalità di analisi più appropriata. Il decesso ospedaliero è stato considerato come variabile dipendente, mentre l'età, SOFA score, sesso, esposizione a fumo di sigaretta, malattia cardiovascolare ed altre comorbidità, parametri di laboratorio (fra cui D-dimero, ferritina, etc) sono stati considerati come variabili indipendenti. L'obiettivo era andare a vedere come la presenza/assenza di una caratteristica (per le variabili categoriche) o il valore di una variabile (per le variabili quantitative) incidevano sul rischio di decesso. Fra i risultati, gli autori riportano che l'età è molto associata al rischio di decesso (odds ratio per ogni anno di incremento d'età 1.14, 95% CI 1.09–1.18; $p<0.0001$). Ciò significa che per ogni anno di incremento d'età si ha un incremento del rischio (dell'odds) di decesso del 14%. Anche la presenza di malattia cardiovascolare è associata alla mortalità (OR=21.4, 95% CI 4.6–98.8; $p<0.0001$), così come ad esempio il D-dimero di valore elevato (valori > 1 μg/mL vs <=0.5 μg/mL, OR=20.0, 95% CI 6.5–61.6; $p<0.0001$) e valori elevati di SOFA score (OR=6.1, 95% CI 3.5–10.9; $p<0.001$). Tutti questi risultati sono stati ottenuti mediante analisi di regressione logistica, i cui risultati possono essere (e generalmente lo sono) espressi mediante odds ratio, che quantificano la forza dell'associazione fra le due variabili (predittore ed evento). In particolare, i risultati appena riportati sono stati ottenuti da una regressione univariata. Questo significa che età, presenza di malattia

cardiovascolare e valore di D-dimero sono stati valutati, per l'effetto che possono avere sul rischio di decesso, separatamente gli uni dagli altri. Un dubbio che può sorgere è: ma se i soggetti che hanno malattia cardiovascolare sono anche quelli più anziani, quell'OR di 21.4 non potrebbe essere (in parte) dovuto anche al fatto che i soggetti con malattia cardiovascolare sono anche quelli più anziani? Detta in altri termini, l'età potrebbe avere introdotto un confondimento? Un modo per effettuare queste valutazioni consiste nell'effettuare una regressione multivariata (i puristi la chiamerebbero regressione multipla). Vale a dire, considerare in un opportuno modello statistico congiuntamente tutte le variabili indipendenti, in modo che si possano ottenere stime di effetto aggiustate per i potenziali confondenti. Infatti, gli autori dello studio riportano anche i risultati del modello multivariato, da cui si ottengono le seguenti stime: età, OR=1.10, 95% CI 1.03–1.17, p=0.0043; D-dimero, per valori > 1 μg/mL vs <=0.5 μg/mL, OR=18 ·42, 95% CI 2.64–128.6, p=0.0033; valori elevati di SOFA score, (OR=5.65, 95% CI 2.61–12.23, p<0.001; presenza di malattia cardiovascolare, OR=2.14, 95% CI 0.26–17.79, p=0.48). Come si può notare, ora la presenza di malattia cardiovascolare non è più un fattore significativamente associato al rischio di decesso, e l'OR è passato da 21.4 a "solo" 2.14 (e non statisticamente significativo). Cosa può essere successo? Presumibilmente, come avevamo accennato in precedenza, in quel 21.4 non c'è il solo effetto della malattia cardiovascolare, ma potrebbe essere compreso l'effetto dell'età, o del D-dimero o del SOFA score. Al di là della significatività statistica, l'interpretazione di quell'OR ottenuto dal modello multivariato è la seguente: la presenza della copatologia malattia cardiovascolare, a parità di tutte le altre condizioni considerate (vale a dire: stessa età, stesso valore di D-dimero e stesso valore di SOFA score) causa mediamente un raddoppio (2.14) del rischio di decesso. Quindi, questo OR=2.14 quantifica l'effetto indipendente sulla mortalità della malattia cardiovascolare, mentre nel valore di OR=21.4 era presumibilmente compreso l'effetto dell'età o di altri fattori.

In conclusione, possiamo dire che è solo con l'analisi multivariata che si riesce ad ottenere una stima dell'effetto indipendente di ciascun fattore considerato. Come abbiamo visto in precedenza con l'esempio della malattia cardiovascolare e dell'età, può infatti capitare che un fattore veicoli un effetto non proprio. Quindi l'ideale sarebbe sempre fare analisi con modelli multivariati includendo il maggior numero possibile di variabili. Attenzione però che, dall'altro lato, è rischioso includere troppe variabili. C'è infatti il rischio del cosiddetto "overfitting", vale a dire l'inclusione di un numero eccessivo di variabili che fa sì che si trovino associazioni spurie, e non reali: con troppe variabili si rappresenta al meglio la situazione osservata, ma non si ha una buona generalizzazione ad altri contesti analoghi.

Quindi, è buona norma individuare a priori le variabili da considerare, sulla base di un substrato fisiopatologico, sulle quali condurre analisi univariate,

selezionare quelle significative su cui costruire il modello multivariato, per arrivare alla fine ad identificare le variabili davvero rilevanti, ovvero quelle significative anche in multivariata. Questo approccio è necessario anche in tutti i modelli di regressione che vedremo in seguito.

7.3.2 Regressione di Cox

Molte volte negli studi clinici siamo interessati a valutare anche quando si verificano gli eventi di interesse ed a confrontare i tempi al verificarsi degli eventi nei gruppi a confronto. In questi contesti si ricorre al modello di regressione di Cox, che ci permette di valutare l'associazione fra variabili indipendenti e variabile dipendente tenendo anche conto del tempo a cui si verificano gli eventi (la cosiddetta "time to event analysis").

In sostanza, anche in questo caso, ricorrendo a particolari modelli matematici, riusciamo a stimare l'effetto che i fattori considerati hanno sul rischio istantaneo (istante per istante) di evento, il cosiddetto *hazard rate*, che può variare nel tempo. Cioè, con questo approccio siamo in grado di calcolare il rischio istantaneo di evento individuale partendo sempre da alcune caratteristiche individuali (predittori o confondenti, variabili indipendenti). I rischi istantanei di decesso (l'hazard rate) per i trattati e per i non trattati si possono combinare in una sorta di rischio relativo, mediante il calcolo del cosiddetto *hazard ratio*, rapporto fra due hazard rate (appunto trattati vs non trattati), interpretabile esattamente come se fosse un rischio relativo. A rigore, l'hazard ratio illustra quanto dopo il gruppo dei trattati avrà l'evento.

Sitbon et al (*JACC 2002;40:780-8*) hanno condotto uno studio il cui obiettivo era l'individuazione dei fattori associati alla sopravvivenza a lungo termine dei pazienti con ipertensione polmonare primitiva trattatati con epoprostenolo. Sono stati arruolati 178 pazienti con diagnosi di ipertensione polmonare primitiva, in classe *NYHA* III o IV, trattati con epoprostenolo. Al momento della diagnosi, per ciascun paziente sono state raccolte informazioni demografiche e anamnestiche, e sono inoltre state effettuate misure di emodinamica. La stima del contributo dato dai fattori prognostici considerati è stata effettuata ricorrendo al modello di regressione di Cox, che applica l'analisi della regressione allo studio della sopravvivenza. Tale metodo risulta più potente di quello di Kaplan-Meier: infatti mentre quest'ultimo si basa, per il confronto, esclusivamente sulla stratificazione dei pazienti in gruppi omogenei per le caratteristiche considerate, il che riduce rapidamente il numero dei pazienti disponibili per stimare la curva di sopravvivenza gruppo-specifica, il modello di Cox, basandosi sull'analisi della regressione, utilizza sempre tutti i dati per le stime di effetto specifico di ciascuna co-variata considerata, naturalmente fatto salvo il rispetto degli assunti di validità (*proportional hazard*) L'analisi porta così a stimare il contributo specifico di ciascuna variabile, aggiustato per ogni

altra, al rischio totale. Nella tabella 7.1 sono riportate le stime univariate di effetto riportate nell'articolo di Sitbon et al.

Variabili	Hazard Ratio (95% CI)	*p-value*
Età > 44 anni	1.17 (0.71-1.94)	0.535
Genere (femminile/maschile)	0.95 (0.53-1.71)	0.877
Storia di scompenso destro (sì/no)	2.19 (1.31-3.64)	0.003
Storia di sincope (sì/no)	0.75 (0.44-1.25)	0.226
TPR (total pulmonary resistance) ≥ 35.4 U/m2	0.65 (0.39-1.09)	0.102

Tabella 7.1 Risultati riportati in uno studio *(JACC 2002;40:780-8).*

Come si legge questa tabella?

L'Hazard Ratio (HR) è una misura di associazione che valuta la relazione esistente fra alcuni fattori ed il rischio di evento (in questo caso, decesso). È una misura simile al rischio relativo (ma non è un rischio relativo!) ed è data dal rapporto fra due rischi istantanei (hazard), in ogni istante, definiti in modo tale da tenere conto di come i decessi si distribuiscono nel tempo. Da un punto di vista interpretativo, l'HR è simile a RR. ipotizzando che l'endpoint che stiamo considerando sia un evento negativo (decesso: un valore pari ad 1 indica che il fattore non ha effetto sul rischio di evento, mentre HR<1 indica che il fattore considerato è un fattore protettivo ovvero che la sua presenza migliora la prognosi e infine HR>1 indica che il fattore considerato è un fattore di rischio: la presenza peggiora la prognosi). Semplificando a fini interpretativi, possiamo dire che l'HR, tenendo conto della differenza di rischio istante per istante, in sostanza ci dice anche quanto prima (o dopo) si verificano gli eventi in un gruppo rispetto all'altro.

Tornando ai risultati riportati nella tabella, vediamo che un solo fattore comporta un incremento di rischio significativo ed altri una riduzione non statisticamente significativa. La storia di scompenso cardiaco destro è associata significativamente alla mortalità: pazienti che hanno avuto scompenso cardiaco destro sono a maggior rischio di decesso (HR=2.19), che raggiunge la significatività statistica (p=0.003) rispetto ai pazienti che non hanno avuto scompenso cardiaco destro. Questo significa che, in ogni istante, i pazienti che hanno avuto scompenso cardiaco sono "mediamente" a rischio più che doppio (2.19) di decesso rispetto ai non scompensati. Vista da un'altra prospettiva, dato che si tiene conto del tempo all'evento, questo significa anche che mediamente i decessi si verificano prima nel gruppo dei soggetti non scompensati rispetto agli scompensati. Ovviamente questo 2.19 è un valore "medio" valutato in tutto il periodo di osservazione. Invece avere un valore di TPR ≥ 35.4 U/m2 porta ad una riduzione del rischio di decesso del 35%

(HR=0.65), ma in questo caso non c'è significatività statistica del risultato (p=0.102).

Anche con il metodo di Cox è possibile effettuare un'analisi multivariata che consente di considerare tutte le variabili di interesse simultaneamente, così da individuare quelle associate all'outcome al netto dell'effetto delle altre co-variate presenti nel modello e da escludere quelle il cui effetto è da attribuire a confondimento. Vediamo ora un esempio di analisi di Cox univariata e multivariata. Fang et al (*J Am Coll Cardiol 2011;58:395–401*) hanno condotto uno studio per valutare i fattori associati al rischio di sanguinamento maggiore in soggetti con fibrillazione atriale trattati con warfarin. È stata considerata una serie di predittori, e dall'analisi univariata è emersa un'associazione statisticamente significativa fra alcune variabili ed il sanguinamento maggiore. Ad esempio, sono risultate significativamente associate a sanguinamento: anemia (HR=4.2, p<0.001), età>75 anni (HR=2.5, p<0.001), emorragie precedenti (HR=2.2, p<0.001), diagnosi di tumore (HR=1.7, p<0.001), ipertensione (HR=1.5, p<0.001), precedente sanguinamento gastrointestinale (HR=2.1, p<0.001) e presenza di cirrosi (HR=2.6, p=0.03). Successivamente è stata condotta un'analisi multivariata, da cui è risultato che fra le variabili significative in univariata solo anemia (HR=3.28), età>75 anni (HR=2.03), emorragie precedenti (HR=1.68) e ipertensione (HR=1.31), oltre a malattia renale severa (HR=2.63), erano risultate significativamente associate al rischio di sanguinamento. Quindi, diagnosi di tumore, cirrosi e precedente sanguinamento gastrointestinale, che erano tutte significative in univariata, perdono la loro significatività all'analisi multivariata. Presumibilmente, questo accade a causa di un'associazione fra queste e alcune degli altri predittori considerati.

In conclusione, abbiano visto che nel valutare l'associazione fra predittori ed eventi di interesse non si può prescindere dall'effettuare analisi multivariate, per cercare di eliminare l'effetto di potenziali fattori di confondimento e per essere ragionevolmente certi che gli effetti misurati siano davvero riconducibili ai predittori.

Nella figura 7.3 è riportato un algoritmo che mostra quale tipo di modello utilizzare nei vari contesti, in funzione del tipo di endpoint dello studio.

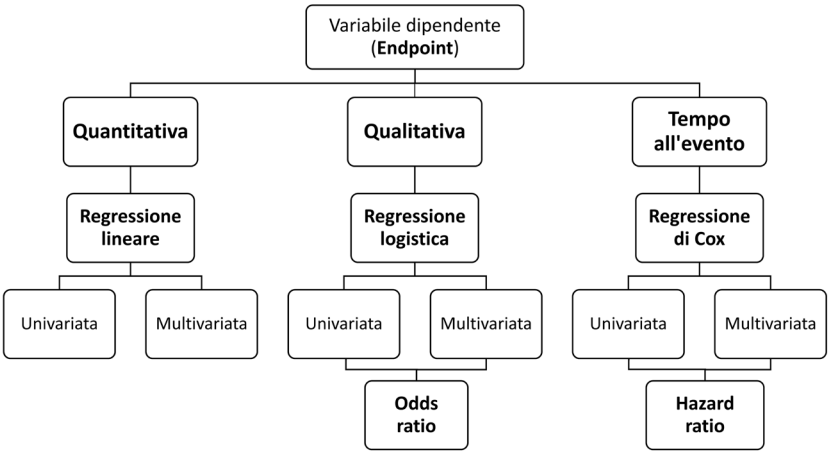

Figura 7.3 Algoritmo di analisi per dati mediante i modelli illustrati nel testo.

Punti chiave

- ✓ Per ogni tipologia di endpoint esistono uno o più appropriati test statistici per effettuare confronti.
- ✓ I modelli di regressione più popolari sono il modello logistico, per endpoint dicotomici, ed il modello di Cox, per il tempo all'evento.
- ✓ Il modello logistico permette di ottenere stime di odds ratio (OR), mentre il modello di Cox permette di ottenere stime di hazard ratio (HR).
- ✓ I modelli di regressione permettono di effettuare confronti aggiustando per potenziali confondenti (analisi multivariata).

Bibliografia consigliata

Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall, London. 1991.

Altman DG, Bland JM. Time to event (survival) data. *BMJ*. 1998;317(7156):468-9.

Bland M. *Statistica medica*. Maggioli Editore. 2019.

Schober P, Vetter TR. Logistic Regression in Medical Research. *Anesth Analg*. 2021;132(2):365-366.

Schober P, Vetter TR. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. *Anesth Analg*. 2018;127(3):792-798.

Tolles J, Lewis RJ. Time-to-Event Analysis. *JAMA*. 2016;315(10):1046–1047.

Tolles J, Meurer WJ. Logistic Regression Relating Patient Characteristics to Outcomes. In: Livingston EH, Lewis RJ. eds. *JAMA Guide to Statistics and Methods*. McGraw-Hill Education; 2019.

8. Studi di Prognosi, revisioni sistematiche e altri studi particolari

8.1 Gli Studi di prognosi

In alcune situazioni, può essere di interesse valutare quale potrebbe essere l'evoluzione di una patologia in un orizzonte temporale di breve o lungo periodo.

8.1.1 Disegno di studio

Una paziente di 35 anni è ricoverata nel vostro reparto per dispnea ingravescente comparsa da qualche mese. Dopo un iter completo, la diagnosi è di ipertensione polmonare primitiva. La paziente è comprensibilmente preoccupata e vi chiede qual è la sua prognosi. Vi ricordate di avere letto tempo fa uno studio (un po' datato!) in cui si valutava la storia naturale dell'ipertensione polmonare primitiva mediante l'osservazione nel tempo di 130 pazienti. L'outcome di interesse primario era la mortalità. Osservando i pazienti per almeno 10 anni, gli autori sono stati in grado di descrivere l'andamento nel tempo della mortalità dei pazienti con ipertensione polmonare primitiva. Il risultato (*Fuster V. Circulation. 1984;70(4):580-7*) è rappresentato mediante la curva di sopravvivenza (curva di Kaplan-Meier), in cui è rappresentata la probabilità di sopravvivenza al variare del tempo, dove tempo=0 indica l'ingresso del paziente dello studio (ad esempio, la diagnosi).

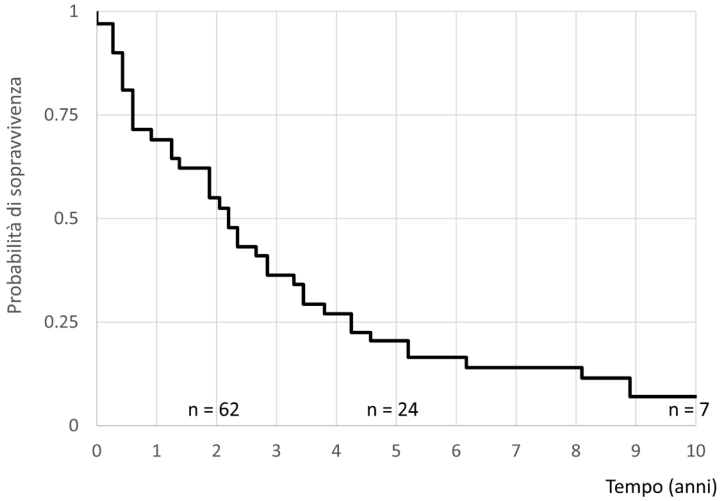

Figura 8.1 Curva di Kaplan-Meier (dati estratti da *Fuster V. Circulation. 1984;70(4):580-7).*

In generale, senza conoscere altre caratteristiche di quella paziente, i risultati di quello studio ci dicono che la nostra paziente avrà una probabilità di sopravvivenza a 2 anni di poco più del 50% ed a 5 anni del 25% circa.

"Fare una prognosi" significa utilizzare alcune caratteristiche del nostro paziente per cercare di predire, in maniera il più possibile accurata, l'evoluzione della malattia. Gli studi di prognosi confrontano frequenza e/o tempo di occorrenza di un particolare esito (favorevole o sfavorevole) di interesse in gruppi di pazienti con diverse condizioni di partenza. Raramente lo studio di prognosi è semplicemente inteso a descrivere la "storia naturale" della malattia (studio descrittivo), più spesso lo scopo è stimare il potere segnaletico di caratteristiche, fattori, condizioni particolari dei pazienti (indicatori prognostici) presenti all'esordio della malattia in base ai quali prevederne l'andamento (studi analitici). Gli indicatori prognostici sono variabili che devono essere facilmente rilevabili nei singoli pazienti (in maniera riproducibile) di tipo demografico (ad esempio sesso, età), clinico (ad esempio stadio di malattia, gravità di malattia, complicanze da malattia, comorbidità), anamnestico (ad esempio abitudine al fumo, consumo di alcool, attività fisica, abitudini dietetiche). Quindi, uno studio prognostico è uno studio osservazionale prospettico, il cui obiettivo principale è cercare di predire il rischio di esito di interesse nel singolo paziente, mediante l'informazione fornita dalle variabili individuali ritenute importanti (indicatori prognostici). I risultati di uno studio prognostico possono essere espressi semplicemente in termini di individuazione dei fattori di rischio principali, oppure mediante la costruzione di uno score (punteggio) prognostico che cerca di quantificare numericamente il rischio di evento, oppure, ancora, mediante la derivazione di una regola decisionale (clinical prediction rule) che classifica il paziente in categorie di rischio (ad esempio, basso-medio-alto) sulla base della presenza di alcune caratteristiche individuali. Alla base di questi tre approcci ci sono comunque analisi della relazione fra potenziali predittori (fattori prognostici) ed evento di interesse effettuate mediante particolari modelli statistici, fra cui i più utilizzati in questo ambito sono il modello logistico (se l'endpoint è valutato ad un tempo fisso: esempio, evento a 10 giorni) ed il modello di Cox (se si valuta il cosiddetto tempo all'evento, cioè se e quando si verifica l'evento).

Per quanto riguarda il disegno di questi studi (che, ricordiamo, sono osservazionali), alcuni fra i punti principali da tenere in considerazione, per evitare l'introduzione di bias, sono:

- i criteri di selezione pazienti (definizione dettagliata criteri inclusione/esclusione)
- la durata del follow-up (tempo all'evento: breve oppure lungo termine, in base all'interesse clinico)
- la scelta dell'outcome (definizione evento e modalità di rilevazione)
- la scelta dei predittori (definizione; modalità di rilevazione; devono essere presenti a t0, momento di inizio del follow-up)

8.1.2 Bias

I bias più diffusi nel caso degli studi di prognosi sono relativi alla selezione dei pazienti, nel senso che una non corretta definizione dei criteri di inclusione ed esclusione potrebbe far sì che alcuni sottogruppi di pazienti con alcune caratteristiche siano sistematicamente sovra o sottorappresentati. Inoltre, dal momento che c'è di mezzo un periodo di follow-up, si potrebbe essere portati ad escludere quei pazienti "difficili" da seguire, oppure si potrebbe avere una elevata quota di pazienti che si ritirano dallo studio (persi al follow-up). Infine, soprattutto quando il numero di potenziali fattori prognostici da misurare è elevato, la presenza di dati mancanti (missing) anche per uno solo dei predittori previsti può portare all'esclusione del paziente. Quando l'esclusione del paziente dallo studio, per le varie motivazioni dette, è in qualche modo associata ad alcune caratteristiche individuali (ad esempio, non riusciamo a raccogliere i dati per i pazienti più gravi, che con più probabilità avranno l'evento), si potrebbe avere un bias. Inoltre, un ulteriore bias potrebbe essere dovuto alla scarsa riproducibilità della misura dei fattori prognostici di interesse (ad esempio, il terzo tono cardiaco potrebbe essere molto importante per la diagnosi e prognosi dei pazienti con scompenso cardiaco, ma la concordanza tra diversi medici nel riconoscimento di questa anomalia è bassissima).

8.1.3 Variabili e clinical prediction tools

Siete di guardia di notte in Pronto Soccorso e giunge con il 118 un paziente di 72 anni per caduta in bagno, complicata da trauma cranico. Dopo aver escluso fratture e aver suturato una profonda lesione lacero-contusa frontale, il collega chirurgo ve lo affida per gli accertamenti del caso. Il paziente non ricorda la dinamica della caduta, tuttavia, la moglie, accorsa nel bagno per il rumore, vi riferisce di aver trovato il paziente svenuto a terra, ma di essere riuscita a svegliarlo dopo pochi secondi. Apprendete, inoltre, che il paziente è affetto da cardiopatia ischemica in esiti di infarto miocardico per cui assume regolarmente ASA, statina, beta-bloccante e ACE-inibitore, ed è inoltre in terapia con inibitori di pompa protonica per gastrite cronica. La pressione in clino- e ortostatismo risulta normale e l'elettrocardiogramma risulta normale. Sareste orientati a interpretare l'evento come episodio sincopale neuro-mediato post-minzionale, tuttavia, mentre siete in attesa dell'esito degli esami ematici, vi chiedete se il paziente possa essere dimesso con tranquillità al termine dell'osservazione per il trauma cranico o non sia piuttosto a rischio di un qualche evento avverso a breve termine e pertanto meritevole di ricovero ospedaliero. Vi ricordate, allora, di avere letto, qualche tempo fa, uno studio italiano (STePS, vi pare, *J Am Coll Cardiol. 2008;51:276-83*), in cui veniva valutata la prognosi a breve e lungo termine dei pazienti giunti in Pronto Soccorso per sincope e venivano individuati i fattori prognostici associati alla morte o altri eventi avversi. In particolare, gli

autori hanno effettuato un'analisi per individuare i fattori più importanti nel predire un evento maggiore a 10 giorni.

Sono stati presi in considerazione 14 fattori, e si è trovato, mediante analisi univariata, che ECG anomalo, anamnesi positiva per trauma, assenza di prodromi, sesso maschile, età maggiore di 65 anni, storia di BPCO, di cardiopatia strutturale, di scompenso cardiaco, sono fattori di rischio per eventi maggiori entro 10 giorni dalla sincope, in quanto per tutti si rileva un OR significativamente maggiore di 1. Non è chiaro, però, se ciascun fattore abbia un suo effetto specifico o se qualcuno di essi abbia solo un effetto apparente, essendo in realtà associato ad un altro fattore fra quelli considerati (confondimento). Ad esempio, si potrebbe pensare che ECG anomalo e scompenso cardiaco siano associati fra loro così che, quando uno è noto, l'altro in realtà non aggiunge alcuna informazione utile alla prognosi. Come abbiamo visto, una modalità di analisi che ci permette di valutare l'apporto di ciascuna variabile, indipendentemente dalle altre, è per l'appunto la regressione logistica multipla, un metodo di analisi multivariata. Nello studio STePS, l'analisi multivariata ha dimostrato che solo ECG anomalo, trauma, assenza di sintomi precedenti la sincope e sesso maschile sono indicatori indipendenti di prognosi sfavorevole (danno ciascuno un contributo all'aumento il rischio di un evento maggiore entro 10 giorni dall'episodio di sincope considerato), mentre le altre variabili (fra cui età maggiore di 65 anni, presenza di cardiopatia strutturale, BPCO e scompenso cardio-circolatorio) si sono dimostrate non aggiungere informazioni prognostiche utili. In tabella sono riportati, per le sole variabili significative all'analisi multivariata, i valori di OR con gli intervalli di confidenza al 95%. Il fattore che risulta il più forte predittore di evento a 10 giorni, ad esempio, è un ECG anomalo: pazienti con anomalie all'ECG hanno un rischio di evento a 10 giorni quasi 7 volte maggiore di quello di pazienti senza anomalie ECG (OR=6.9, 95% IC 3.1-15.1).

Predittore	**Odds Ratio (95% CI)**	***p-value***
Anomalie ECG alla presentazione	6.9 (3.1 – 15.1)	<0.001*
Trauma	2.9 (1.4 – 5.9)	0.004*
Assenza di prodromi	2.4 (1.2 – 4.8)	0.016*
Sesso maschile	2.2 (1.0 – 4.5)	0.037*

Tabella 8.1. Predittori di evento maggiore a 10 giorni in pazienti giunti in Pronto Soccorso per sincope. Dati tratti da *J Am Coll Cardiol. 2008;51:276-83.*

Questo è un esempio di studio finalizzato all'individuazione dei principali fattori di rischio ed alla stima della forza dell'associazione fra fattori di rischio ed evento.

Sempre nell'ambito della sincope, alcuni autori hanno costruito una regola per predire eventi maggiori entro 30 giorni dall'accesso in PS per evento sincopale (studio ROSE, *J Am Coll Cardiol. 2010;55:713-21*). Mediante analisi di regressione logistica multipla sono stati individuati come fattori predittivi di esito sfavorevole BNP>=300pg/ml (OR=7.3), sangue occulto nelle feci (OR=13.2), emoglobina <=90g/l (OR=6.7), saturazione O2<=94% (OR=3.0), ECG con onda q patologica (OR=2.8), che hanno aggiunto a dolore toracico e bradicardia per ottenere uno score di rischio semplicemente definito come segue: è da considerare a rischio di evento maggiore a 30 giorni il paziente che presenti almeno 1 dei 7 fattori sopra elencati. Quindi, sempre partendo da una analisi logistica e da stime di OR, sono stati individuati i principali fattori di rischio per evento a 30 giorni. In aggiunta, gli autori hanno deciso di combinare questi fattori mediante un algoritmo, in base al quale è considerato a rischio ogni soggetto che presenta almeno uno dei 7 fattori riportati. Ai fini pratici, nella clinica è utile questa regola decisionale? Gli autori riportano che questa regola (rule) ha una sensibilità del 92.5% ed una specificità del 73.8% (VPP=22.4%, VPN=99.2%) nell'individuare un evento maggiore a 10 giorni.

Rimanendo ancora nell'ambito della sincope, altri autori (studio OESIL, *Eur Heart J. 2003;24(9):811-9*) hanno studiato i fattori prognostici per la mortalità per ogni causa a 12 mesi in pazienti presentatisi in PS per sincope. Il disegno e la conduzione dello studio sono stati in generale simili a quelli degli esempi precedenti. Considerato che l'endpoint è stato valutato a medio/lungo termine (12 mesi), gli autori hanno scelto di analizzare i dati mediante modello di Cox. Hanno così effettuato un'analisi univariata e multivariata, individuando in multivariata l'età>65 anni (HR=1.42), anamnesi di malattia cardiovascolare (HR=1.34), sincope senza prodromi (HR=1.13) e ECG anomalo (HR=1.29) quali fattori di rischio per la mortalità a 12 mesi. Infine, gli autori hanno concluso le analisi combinando questi fattori in uno score, vale a dire calcolando per ogni singolo paziente un punteggio dato dal numero di fattori prognostici (fra i quattro individuati) presenti. Questo score varia fra 0 (nessun fattore presente) e 4 (paziente con tutti i fattori) ed analogamente a quanto visto con l'esempio della rule (ROSE) può essere utilizzato per valutare il rischio di evento. In base ai dati riportati nello studio citato, pazienti con uno score pari a 4 sembrerebbero avere un rischio di decesso a 12 mesi elevato (stimato pari a 57% circa), mentre soggetti con valore di score pari a 0 o 1 sembrerebbero avere un rischio di decesso molto basso (0% e 0.8%, rispettivamente). Gli autori, infine, riportano che questo score sembrerebbe essere utile nella pratica clinica, perché caratterizzato da un valore di area sotto la curva ROC (AUC), pari a 0.90 circa.

Nei tre esempi appena visti abbiamo considerato i risultati ottenuti dagli autori dei tre studi sui loro pazienti. Sicuramente, in base ai dati riportati, sono stati individuati importanti fattori prognostici. Usereste nella pratica clinica i risultati di uno qualunque dei tre studi per valutare il rischio di evento in uno dei vostri pazienti? Una regola generale che dobbiamo tenere ben presente è cha la validità generale di un qualunque studio del genere deve essere sempre valutata anche in contesti simili (per caratteristiche generali dei pazienti), ma diversi (per pazienti) da quello in cui il modello prognostico è stato creato. Potrebbe infatti essere che alcuni fattori prognostici siano ottimi stando allo studio che li ha individuati (studio di derivazione del modello prognostico), ma magari siano un po' meno buoni quando utilizzati nella pratica clinica. Sono allora necessari i cosiddetti studi di validazione, vale a dire studi condotti per cercare di replicare i risultati ottenuti dallo studio di derivazione. Gli studi di validazione possono essere condotti nello stesso contesto (stesso ospedale) dello studio di derivazione, ma su pazienti differenti (validazione interna). Oppure possono essere condotti arruolando pazienti differenti in contesti differenti da quello di derivazione (validazione esterna). Solo se anche gli studi di validazione esterna dimostrano che lo score o la rule hanno una buona capacità di predire eventi, possiamo pensare di farne uso nella pratica clinica. In assenza di studi di validazione esterna, se ne sconsiglia fortemente l'uso. In realtà, come vedremo in seguito, per utilizzare con sicurezza questi score, bisognerebbe anche confrontarsi con il giudizio clinico e fare un trial randomizzato e controllato per vedere se l'implementazione dello score nella clinica dia un beneficio. Come vi sarete già resi conto, gli studi di derivazione e validazione di score e rule sono, per molti aspetti, simili a quelli diagnostici.

Punti chiave – Studi prognosi

- ✓ Gli studi predittivi (di prognosi) costituiscono la base per l'informazione per il paziente e per le decisioni mediche.
- ✓ Quando leggiamo uno studio prognostico prestiamo attenzione al disegno adottato dagli autori: selezione pazienti, individuazione e misura dei predittori, misura dell'endpoint.
- ✓ Le analisi statistiche sono principalmente condotte mediante modelli di regressione logistica (generalmente quando l'evento è a breve termine; risultati espressi mediante odds ratio) o regressione di Cox (generalmente quando l'evento è a lungo termine e siamo interessati al tempo all'evento; risultati espressi mediante hazard ratio).
- ✓ Le stime prodotte dagli studi sull'utilità dei fattori prognostici (solitamente espresse tramite OR o HR) devono essere accurate e riproducibili e possono essere sintetizzate e combinate in score o prediction rules.

- ✓ Per l'applicazione nella pratica clinica, è necessario che i modelli predittivi siano stati validati esternamente e possibilmente confrontati con il giudizio clinico.
- ✓ Facciamo comunque un utilizzo critico (no fiducia cieca): la stima del rischio di evento futuro ottenuta dai modelli prognostici è solo una fra le informazioni disponibili per la gestione del paziente.

8.2 Studi osservazionali di intervento

In alcune situazioni in cui si vuole valutare l'efficacia di un intervento rispetto ad un controllo, per vari motivi, potrebbe non essere possibile condurre studi sperimentali e si deve quindi ricorrere a studi osservazionali. Alla luce di tutto quanto abbiamo visto sino ad ora, sappiamo benissimo quali sono i principali punti di forza di uno studio sperimentale randomizzato e quali sono i principali punti di debolezza di uno studio osservazionale prospettico. In particolare, gli studi osservazionali possono essere soggetti al bias di confondimento, che, non potendo essere neutralizzato dalla randomizzazione, deve essere gestito con tecniche di analisi statistica.

8.2.1 Il propensity score

Geleris et al (*N Engl J Med 2020;382:2411-8*) hanno condotto uno studio osservazionale per valutare l'efficacia dell'idrossiclorochina nei pazienti affetti da Covid-19. Hanno arruolato 1376 pazienti consecutivi affetti da Covid-19, ricoverati in un ospedale di New York dal 07/03/2020 al 08/04/2020. Di questi, 811 erano stati trattati con idrossiclorochina, mentre i restanti 565 avevano ricevuto altri trattamenti. L'endpoint primario scelto dagli autori è stato l'evento composito decesso o intubazione. Al termine del periodo di osservazione, si sono avuti 262 eventi (32.3%) nel gruppo idrossiclorochina e 84 eventi (14.9%) nel gruppo che ha ricevuto altri trattamenti. La conclusione, tenendo conto del tempo all'evento, è che i soggetti trattati con idrossiclorochina sono a maggior rischio rispetto agli altri: HR=2.37 (95% CI: 1.84-3.02). Quindi la conclusone è che l'idrossiclorochina è un fattore di rischio, in quanto raddoppia il rischio di decesso/intubazione. Però… questo è uno studio osservazionale, la scelta di trattare o meno con idrossiclorochina è stata presa per varie motivazioni dai medici che avevano in cura i soggetti, e non mediante randomizzazione dai ricercatori che hanno condotto lo studio. Questo fattore può avere introdotto un bias (confondimento): la scelta di trattare o meno con idrossiclorochina potrebbe essere stata presa dai medici valutando in qualche modo la gravità del paziente, e decidendo di trattare i più gravi. Se così fosse, la gravità del paziente sarebbe associata al trattamento (tratto i più gravi). Possiamo ovviamente ipotizzare anche un'associazione fra gravità e evento (i più gravi saranno a maggior

rischio di decesso/intubazione). Ecco allora che in queste condizioni la gravità del paziente rappresenterebbe un perfetto confondente nell'analisi della relazione fra idrossiclorochina e decesso/intubazione. Per cercare di controllare questo effetto di confondimento, gli autori hanno condotto un'analisi, indicata come "propensity score matched analysis", in cui hanno tentato di controllare il possibile confondimento dato dalla selezione dei pazienti da parte dei medici curanti. In base a questa analisi, per ogni paziente arruolato è stato calcolato un punteggio (score) che in qualche modo quantifica la "propensione" individuale di ricevere idrossiclorochina, considerando fattori demografici o clinici, risultati di test di laboratorio, presenza di co-trattamenti. Il propensity score cerca di quantificare la probabilità che il paziente riceva il trattamento in base alle proprie caratteristiche (per esempio, in base all'età, le co-patologie, la gravità clinica). Dopo aver calcolato il propensity score, si possono utilizzare diverse tecniche per ridurre l'effetto del confondimento sull'outcome di interesse. In questo caso, gli autori hanno condotto l'analisi appaiando (matched) i soggetti nei due gruppi di trattamento sulla base del valore del propensity score. Nel dettaglio, se ho un paziente nel gruppo idrossiclorochina per il quale abbiamo calcolato uno score (che, ricordiamo, è la propensione a ricevere idrossiclorochina sulla base delle sole caratteristiche individuali elencate in precedenza) pari ad esempio a 0.275, allora devo cercare fra i non trattati un paziente con lo stesso valore di score (o un valore molto simile) da includere nell'analisi matched. Il senso di questa analisi sta nel fatto che se abbiamo due pazienti con caratteristiche simili (in termini di fattori demografici, fattori clinici, risultati di test di laboratorio, presenza di co-trattamenti), uno nel gruppo idrossiclorochina ed uno nel gruppo che ha preso altri trattamenti, possiamo assumere che il fatto che uno dei due sia stato trattato e l'altro non trattato possa essere dovuto al caso. In un certo senso, l'analisi matched con propensity score "mima" la randomizzazione in studi osservazionali. Quindi, eseguendo la stessa operazione di appaiamento per tutti i pazienti del gruppo di idrossiclorochina, avremo due gruppi a confronto molto simili per tutte le caratteristiche che abbiamo considerato, e quindi eventuali differenze di incidenza di eventi possono essere attribuite all'effetto del trattamento. In un certo senso, l'analisi con propensity score è molto simile, concettualmente, all'analisi con modelli di regressione multipla. La differenza principale, dal punto di vista statistico, è che la si può ritenere un po' più efficiente della regressione. Per quanto riguarda l'esecuzione pratica, esistono differenti approcci, basati su assunti differenti che possono essere più o meno appropriati nei differenti contesti. Oltre all'analisi matched descritta sopra, che può avere il difetto di non fare uso di tutti i pazienti (quelli non appaiati/non appaiabili sono esclusi), esistono altri approcci. Uno fra questi consiste nel calcolare la differenza di rischio di evento fra i due gruppi pesando ogni paziente in base all'inverso del proprio propensity score: i pazienti trattati con idrossiclorochina che avevano (in base allo score) un'elevata probabilità di

essere trattati (elevato valore dello score) peseranno meno, nelle analisi, dei pazienti che avevano una bassa probabilità di essere trattati con idrossiclorochina. Facendo un ragionamento analogo per i pazienti del gruppo di controllo, riusciamo così ad attribuire un peso a ciascun paziente nell'analisi, tenendo conto del trattamento ricevuto e della probabilità di ricevere quel trattamento sulla base delle caratteristiche individuali. Così facendo, si cerca di ridurre l'effetto del confondimento.

Tornando al nostro esempio, fra altre analisi, gli autori dello studio hanno pubblicato i risultati dell'analisi fatta con il matching, includendo tutti gli 811 pazienti del gruppo idrossiclorochina e solo 274 dell'altro gruppo di trattamento (questi ultimi sono i soli pazienti nel gruppo non-idrossiclorochina che hanno valore di propensity score appaiabile a quello di almeno un paziente del gruppo idrossiclorochina) che mostra HR=0.98 (95% CI: 0.73-1.31). In definitiva, questa analisi mostra che non c'è differenza di efficacia fra i due gruppi, visto che HR è praticamente pari ad 1. Possiamo quindi supporre che quell'HR di 2.37 che si ottiene dall'analisi grezza sia influenzato dall'effetto di potenziali confondenti (di fattori demografici, fattori clinici, risultati di test di laboratorio, presenza di co-trattamenti). Ad un risultato simile si sarebbe giunti utilizzando un modello di regressione multipla, aggiustando per gli stessi confondenti. La differenza principale fra analisi con propensity score ed analisi di regressione multipla sta nel fatto che con la prima analisi tutti i confondenti vengono sintetizzati in un unico valore numerico (il propensity score), mentre con le analisi di regressione ogni confondente è una variabile a sé, con tutti i pro e i contro del caso.

Nonostante questo metodo permetta di ridurre i bias in uno studio osservazionale di intervento, non si possono predere decisioni cliniche sul singolo paziente in base ai risultati di uno studio osservazionale di questo tipo, anche se le analisi sono state condotte utilizzando il propensity score.

Punti chiave – Studi osservazionali di intervento

- ✓ Nel caso si voglia valutare l'efficacia di un intervento mediante studio osservazionale, si è ad elevato rischio di bias da confondimento.
- ✓ Per cercare di controllare l'effetto del confondimento si possono condurre le analisi statistiche mediante utilizzo del propensity score.
- ✓ Dobbiamo tenere presente che, comunque, gli studi che analizzano i dati mediante propensity score non possono sostituire gli studi sperimentali randomizzati, che, quando fattibili, forniscono, se ben condotti, la migliore evidenza.

8.3 Gli studi di non inferiorità

Nel vostro ambulatorio di medico di medicina generale arriva un paziente di 77 anni, affetto da ipertensione arteriosa, diabete mellito e fibrillazione atriale permanente in terapia anticoagulante orale. Ha deciso di trasferirsi con la moglie in un piccolo paese di montagna ed è preoccupato perché l'ospedale più vicino si trova ad oltre 60 Km di distanza e, avendo uno scarso controllo dell'INR, di solito deve eseguire controlli ogni 15 giorni. Vi chiede pertanto se non sia il caso di sospendere la terapia anticoagulante. Spiegate al paziente che il suo rischio cardio-embolico è piuttosto elevato e che, al suo posto, cerchereste di continuare il warfarin, però capite perfettamente che la gestione della terapia potrebbe diventare un problema importante.

Vi ricordate di aver partecipato ad un congresso in cui si parlava di un farmaco anticoagulante orale, il dabigatran, che non necessita di controlli della coagulazione e decidete di capirci di più. Cercate su PubMed (aiutandovi come al solito con l'acronimo "PICO") "atrial fibrillation" AND "warfarin" AND "dabigatran" AND "stroke", mettete come limite "Randomized Controlled Trial" e trovate rapidamente il trial *RE-LY* pubblicato sul New England nel 2009 *(N Engl J Med 2009;361:1139-51)*.

Si tratta di un trial di non inferiorità in cui sono stati randomizzati a terapia con dabigatran (100 o 150 mg x 2/die) o warfarin più di 18000 pazienti con fibrillazione atriale a rischio di ictus. L'endpoint primario era l'incidenza di ictus o embolismo sistemico. Nei metodi leggete che lo studio è stato disegnato per valutare la non-inferiorità del dabigatran rispetto al warfarin nella prevenzione di ictus o embolia sistemica. Per soddisfare l'ipotesi di non inferiorità, il limite superiore dell'intervallo di confidenza del rischio relativo (di ictus o embolia sistemica del dabigatran rispetto al warfarin) doveva essere minore di 1.46. Una volta stabilita la non inferiorità di entrambe le dosi di dabigatran, tutte le analisi successive sono state condotte per testarne la superiorità.

I risultati dello studio mostrano che l'endpoint primario si è verificato nell'1.53% all'anno nei pazienti in terapia con dabigatran 110 mg, nell'1.11% all'anno in quelli in terapia con dabigatran 150 mg e nell'1.69% all'anno in quelli in terapia con warfarin. In base alle conclusioni degli autori, entrambe le dosi di dabigatran sono risultate non inferiori al warfarin.

Come interpretiamo questi risultati? Che cosa significa che il dabigatran è risultato "non inferiore" al warfarin?

8.3.1 Logica e disegno di studio

Abbiamo imparato che il "gold standard metodologico" della ricerca clinica sono gli studi randomizzati-controllati. Se vogliamo testare un nuovo trattamento, la cosa migliore è condurre uno studio in cui si confronta il nuovo farmaco contro placebo. Tuttavia, quando vogliamo testare un trattamento

alternativo ad una terapia già esistente, non sarebbe etico condurre uno studio in cui si somministra un placebo a uno dei due gruppi di malati. In questo caso, si disegnerà uno studio che confronti il nuovo trattamento con quello standard. Il nuovo trattamento potrà essere paragonato a quello in uso tramite uno studio di superiorità in cui si cercherà di dimostrare che il nuovo farmaco è più efficace. Nel caso in cui non si pensa che il nuovo farmaco sia più efficace del vecchio, ma che abbia un profilo di sicurezza, o un costo o una comodità per il paziente migliori, si potrebbe disegnare uno studio di non inferiorità in cui ci accontentiamo di dimostrare che il nuovo farmaco è anche meno efficace del vecchio, ma entro un limite considerato accettabile.

Uno studio di non inferiorità confronta un nuovo trattamento (T) con un trattamento di comprovata efficacia (C) con l'obiettivo di dimostrare che la nuova terapia è non meno efficace della prima di un certo margine definito a priori, detto appunto "margine di non inferiorità" (indicato con Δ). Di solito si utilizza un disegno di non inferiorità quando non siamo interessati a (e potremmo non essere in grado di) dimostrare la superiorità di T rispetto a C, ma quando il primo trattamento offre dei vantaggi, ad esempio è meno costoso o più semplice da somministrare, oppure ha meno effetti collaterali. Si utilizza un disegno di non inferiorità quando siamo interessati a vedere se T è di efficacia simile a C: vale a dire, saremmo propensi a sostituire C con T qualora riuscissimo a dimostrare la superiorità di T, ma anche una lieve inferiorità (definita con il margine di non inferiorità) di T rispetto a C. Nell'esempio del trial *RE-LY*, sareste contenti di utilizzare il dabigatran nel vostro paziente se si dimostrasse non inferiore al warfarin (quindi anche peggiore, ma entro un certo limite), perché è sicuramente più maneggevole e gestibile.

Vediamo ora più nel dettaglio qualche aspetto metodologico degli studi di non inferiorità.

8.3.2 Ipotesi nulla e ipotesi alternativa

Abbiamo già visto come, negli studi di superiorità, per riuscire a dimostrare che una terapia è più efficace di un'altra o del placebo, bisogna ottenere un risultato che consenta di giudicare falsa l'ipotesi nulla di partenza (cioè che non vi sia differenza di efficacia tra i due trattamenti che si stanno confrontando). Negli studi di non inferiorità ipotesi nulla e ipotesi alternativa sono esattamente ribaltate. Nella Tabella 8.2 sono riportate le caratteristiche degli studi di non inferiorità, in confronto con quelli di superiorità.

	Trial di superiorità	Trial di non inferiorità
Ipotesi nulla (H_0)	T e C hanno pari efficacia	T è inferiore a C
Ipotesi alternativa (H_1)	T e C hanno efficacia diversa	T è migliore o uguale a C
Errore di primo tipo	concludere erroneamente che T è diverso da C	concludere erroneamente che T è migliore o uguale a C
Errore di secondo tipo	concludere erroneamente che T e C hanno pari efficacia	concludere erroneamente che T è inferiore a C

Tabella 8.2. Definizione di H0, H1, errore di I e II tipo per gli studi di superiorità e di non inferiorità. T nuovo farmaco; C trattamento di controllo.

8.3.3 Il margine di non inferiorità (Δ)

L'elemento critico di questo tipo di trial risiede nello stabilire un margine di non inferiorità appropriato. Ci sono due tipi di approccio:

- stabilire "a tavolino" il margine di non inferiorità sulla base di quella che può essere considerata una differenza di efficacia poco rilevante clinicamente (ovviamente è un criterio soggettivo!);
- scegliere il margine di non inferiorità facendo riferimento ai risultati di studi precedenti in cui era stato confrontato il farmaco attivo con il placebo, stabilendo che il margine di non inferiorità debba essere inferiore (ad esempio la metà) del minore effetto che ci aspetteremmo di trovare confrontando il farmaco con il placebo. Ad esempio, se la riduzione assoluta di mortalità con un trattamento C rispetto al placebo è del 10%, il nuovo trattamento T che vogliamo testare con uno studio di non inferiorità dovrebbe essere inferiore al trattamento di controllo C di non più del 5%.

Nello studio *RE-LY* il limite superiore dell'IC del rischio relativo dell'outcome primario del dabigatran rispetto al warfarin deve essere minore di 1.46. Questo margine di non inferiorità è stato definito partendo dai risultati di una metanalisi di studi che confrontavano warfarin e placebo nella prevenzione dell'ictus in pazienti con fibrillazione atriale, dimezzando il limite superiore dell'intervallo di confidenza al 95% del rischio relativo della terapia di controllo rispetto al warfarin.

8.3.4 Numerosità campionaria

La numerosità del campione dipende dal livello di confidenza, rischio di errore di tipo II e D. In generale, potrebbe essere sufficiente arruolare un minor numero di pazienti rispetto ai trial di superiorità, perché è possibile scegliere "a piacere" il margine Δ. Modificando (anche "a tavolino") il margine di non inferiorità, si modifica la numerosità campionaria: quanto più è ampio Δ, tanto

minore è il numero di pazienti che bisognerà arruolare, e viceversa. Per questo bisogna leggere con particolare attenzione questo tipo di trials.

Nello studio *RE-LY* viene calcolata una numerosità campionaria di 15000 pazienti per avere una potenza dell'84% per dimostrare la non inferiorità di entrambe le dosi di dabigatran.

8.3.5 Analisi ed interpretazione dei risultati

L'analisi per intention-to-treat è universalmente riconosciuta come il metodo più conservativo per analizzare i dati negli studi di superiorità. Sfortunatamente non si può dire altrettanto per quanto riguarda i trial di non inferiorità. Infatti, ammettiamo di somministrare un nuovo farmaco che vogliamo testare (T) e confrontarlo con un farmaco efficace (C). Se nel gruppo trattato con T avessimo tanti pazienti che sospendono il farmaco (ad esempio perché ha molti effetti collaterali) o che cambiano il braccio di trattamento nel corso del trial iniziando il farmaco alternativo C, e questi pazienti venissero comunque inclusi nell'analisi come appartenenti al gruppo di trattamento T, potremmo concludere erroneamente che i farmaci sono equivalenti quando invece T è inferiore.

Dall'altro lato, un'analisi "per-protocol", escludendo i dati di coloro che violano il protocollo, potrebbe portare a risultati falsati in entrambe le direzioni.

Pertanto, gli studi di non inferiorità sono spesso analizzati con entrambi gli approcci e, solo se entrambi i risultati sono a favore della non inferiorità, il trial è considerato positivo.

I risultati di uno studio di non inferiorità vengono interpretati guardando esclusivamente l'intervallo di confidenza della stima dell'effetto (nell'esempio è il RR). La decisione finale (rifiuto o non rifiuto dell'ipotesi che il trattamento nuovo è inferiore allo standard) dipende da dove si colloca l'intervallo di confidenza dell'effetto del nuovo trattamento rispetto al margine di non inferiorità e rispetto all'uguaglianza dei due trattamenti. Cerchiamo di capire meglio con un esempio grafico (Figura 8.2) che mostra tutti i possibili risultati di un trial di non inferiorità. La linea verticale rossa tratteggiata indica il margine di non inferiorità; la linea verticale continua rappresenta l'uguaglianza tra i due trattamenti; le linee orizzontali contrassegnate dalle lettere sono gli intervalli di confidenza.

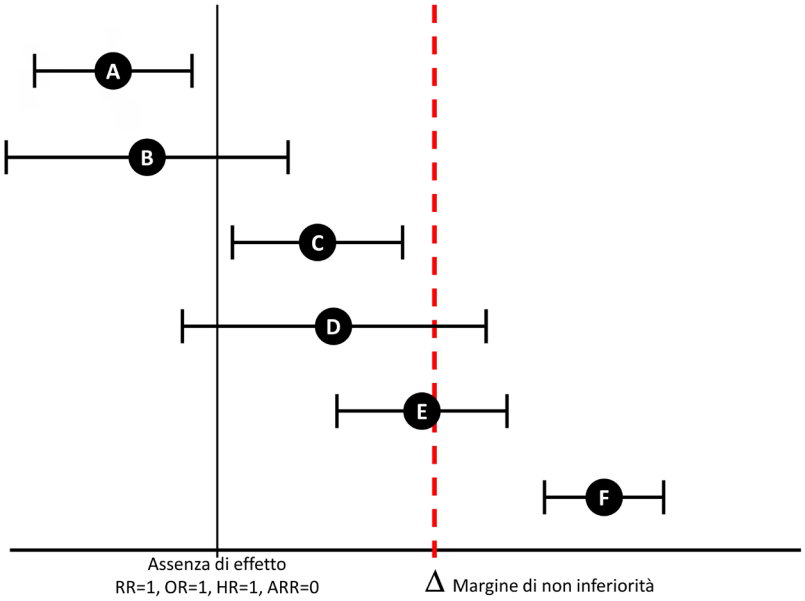

Figura 8.2 Illustrazione dei possibili scenari di risultati di studi di non inferiorità (modificata da JAMA 2006; 295:1152-1160).

I casi A, B e C rappresentano risultati di studi in cui viene dimostrata la non inferiorità del nuovo trattamento perché l'intervallo di confidenza (IC) della misura di efficacia è completamente al di sotto di D. Vediamo poi come in A l'IC è completamente al di sotto della soglia di assenza di effetto, per cui il nuovo trattamento sembrerebbe addirittura superiore al trattamento standard. Invece, nel caso C l'IC è al di sotto del margine di non inferiorità ma non comprende quello di assenza di differenza di efficacia, significa che il nuovo trattamento è non inferiore in base alla definizione dello studio, ma è comunque peggiore. Nei casi D ed E non possiamo dimostrare la non inferiorità perché l'IC va oltre D; infine, nel caso F l'IC è completamente oltre il margine di non inferiorità e quindi il nuovo trattamento è inferiore.

Da questi esempi possiamo capire come sia fondamentale interpretare i risultati di uno studio di non inferiorità con cautela e sempre alla luce sia dell'IC che del margine di non inferiorità. Per illustrare meglio il concetto (Figura 8.3), prendiamo ad esempio il caso D della figura precedente, in cui non possiamo affermare la non inferiorità del nuovo trattamento (scenario A, margine D1). Se nello stesso studio avessimo stabilito un margine di non inferiorità più ampio (scenario B, margine D2), l'IC sarebbe rimasto completamente al di sotto del margine di non inferiorità e il nuovo trattamento sarebbe risultato non inferiore. Infine, se, pur mantenendo il primo margine di inferiorità (D1), avessimo aumentato la numerosità del campione (scenario C), avremmo avuto in IC più

ristretto e anche in questo caso (a differenza che nello scenario A) avremmo potuto dichiarare la non inferiorità del nuovo trattamento.

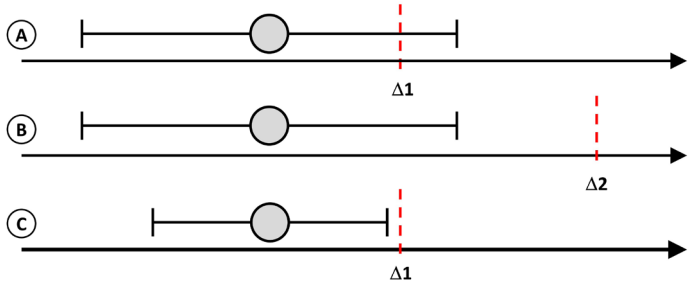

Figura 8.3 Illustrazione dei possibili scenari di risultati di studi di non inferiorità: effetto del margine e dell'ampiezza dell'intervallo di confidenza (modificata da *JAMA 2006; 295:1152-1160*).

Schematizzando i risultati del trial RE-LY, in figura 8.4 vediamo come il dabigatran risulti non inferiore al warfarin nella prevenzione degli eventi cardio-embolici.

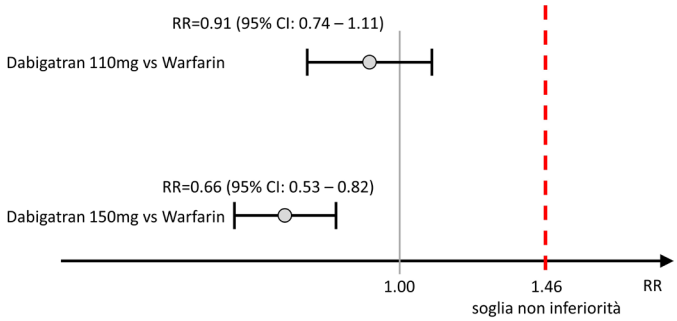

Figura 8.4 Risultati dello studio *RE-LY (N Engl J Med 2009;361:1139-51)*: rischio relativo di ictus o embolia sistemica (endpoint primario dello studio). Il valore 1.46 rappresenta il margine di non inferiorità.

Questo esempio numerico ci permette di chiarire quanto dicevamo prima a proposito del fatto che, in genere, gli studi di non inferiorità permettono di arruolare un minor numero di soggetti rispetto ai trial di superiorità. Prendiamo ad esempio il confronto tra dabigatran 110 mg e warfarin: il RR è 0.91 con un IC al 95% tra 0.74 e 1.11. Se questo fosse stato il risultato di uno studio di superiorità, il trial sarebbe risultato non conclusivo: per dimostrare la superiorità, a parità di tutte le altre condizioni, avremmo dovuto arruolare un numero molto maggiore di soggetti, in modo da avere RR=0.91 con IC al 95% più ristretto, ad esempio 0.86-0.99.

Una volta dimostrata la non inferiorità, è possibile cercare di valutare se il nuovo trattamento sia anche superiore utilizzando dei test definiti a priori e un'analisi per intention-to-treat. Nello studio *RE-LY*, ad esempio, l'analisi di superiorità mostra che il dabigatran alla dose di 150 mg è superiore al warfarin.

Viceversa, in un trial di superiorità non si può fare a posteriori un'analisi di non inferiorità: se uno studio non riesce a dimostrare la superiorità di un nuovo farmaco rispetto ad un farmaco esistente, non possiamo in nessun modo dire che il farmaco sia uguale o peggiore del trattamento standard. Invece, se noi a priori accettassimo che possa essere "non troppo peggiore", potremmo disegnare ad hoc uno studio di non inferiorità.

8.3.6 Per una lettura critica

Abbiamo analizzato le varie caratteristiche metodologiche degli studi di non inferiorità dalle quali sembra emergere un ruolo importante di questo tipo di disegno nel dimostrare un possibile ruolo di farmaci, forse non altrettanto efficaci, ma con altri vantaggi rispetto ai trattamenti esistenti. Tuttavia, per una loro interpretazione attenta e rigorosa, bisogna fare attenzione ad alcuni aspetti.

Innanzitutto, è importante che venga definito in maniera chiara il razionale: vogliamo studiare un nuovo farmaco che ha dei potenziali vantaggi (ad esempio, minor costo o effetti collaterali, maggior facilità di utilizzo) rispetto ad un trattamento di comprovata efficacia. Se i risultati del trial fossero positivi, il nuovo trattamento potrebbe essere migliore, uguale o "non troppo peggiore" del trattamento standard. Ci va bene che possa anche essere "non troppo peggiore", proprio perché ha degli altri benefici che non fanno parte dell'outcome che stiamo valutando, altrimenti non ci sarebbe motivo di condurre uno studio di non inferiorità. E noi conduciamo questo studio (invece che uno di superiorità) proprio con la speranza di dimostrare che sia non inferiore.

È proprio alla definizione di "non troppo peggiore" che bisogna porre la maggior attenzione. Abbiamo visto come ci siano dei metodi più o meno codificati per stabilire il margine di non inferiorità, ma spesso succede che venga definito a tavolino e in base alle esigenze degli sperimentatori. Inoltre, definire un margine di non inferiorità ampio permette da un lato di arruolare un minore numero di pazienti, dall'altro lato di avere risultati positivi anche per trattamenti che sarebbe discutibile definire davvero "simili". Proprio per questo, succede che gli studi di non inferiorità vengano particolarmente apprezzati dall'industria farmaceutica, perché permettono di far risparmiare nell'arruolamento di pazienti (avendo necessità di un campione in genere meno numeroso degli studi di superiorità), e perché, "giocando" con il margine di non inferiorità, è più semplice avere dei risultati positivi e immettere in commercio farmaci di efficacia non così chiara.

Oltre agli studi di non inferiorità esistono anche i cosiddetti studi di equivalenza, il cui obiettivo consiste nel dimostrare l'equivalenza tra i due trattamenti.

Questo comporta il fatto che l'ipotesi nulla sia rovesciata (H0 i due trattamenti hanno efficacia diversa). Anche qui bisogna stabilire una definizione di "equivalenza" e la numerosità campionaria è usualmente più elevata. Per questo sono trial molto rari.

Un'estensione del *CONSORT statement* permette di valutare gli studi di non inferiorità: http://www.consort-statement.org/extensions/designs/non-inferiority-and-equivalence-trials/.

Punti chiave – Studi di non inferiorità

- ✓ Gli studi di non inferiorità valutano se un nuovo trattamento è simile o non peggiore (di un margine D, detto margine di non inferiorità) rispetto ad un trattamento efficace considerato di scelta.
- ✓ In genere, si utilizza uno studio di non inferiorità quando siamo disposti ad accettare che un nuovo trattamento sia anche lievemente meno efficace del trattamento standard ma ha degli altri vantaggi (ad esempio è più economico, ha meno effetti collaterali, è più semplice da somministrare).
- ✓ Uno degli elementi critici è definire il margine di non inferiorità, che dovrebbe essere scelto a priori in base all'efficacia del trattamento standard rispetto al placebo.
- ✓ Dal margine di non inferiorità dipende anche la numerosità del campione: più il margine è ampio, minore sarà il numero di pazienti da arruolare. In generale, gli studi di non inferiorità hanno campioni meno numerosi degli studi di superiorità.
- ✓ L'analisi dei dati dovrebbe avvenire sia per intention-to-treat che per-protocol;
- ✓ I risultati degli studi di non inferiorità devono essere interpretati con cautela,
- ✓ perché spesso vengono utilizzati per "vendere" come ugualmente efficace un trattamento che invece è inferiore al trattamento standard. Quando si legge uno studio di non inferiorità è importante valutare:
  - se è sensato condurre uno studio di non inferiorità per rispondere alla domanda clinica che ci si pone;
  - se il margine di non inferiorità corrisponde davvero a una differenza di efficacia che possiamo considerare clinicamente trascurabile;
  - gli intervalli di confidenza: a maggior ragione in questi studi interpretare i risultati guardando sempre gli intervalli di confidenza.

8.4 Revisioni sistematiche e meta-analisi

Le revisioni sistematiche, che si collocano al vertice della piramide dell'evidenza, consistono in una rassegna sistematica di tutti i lavori presenti in

letteratura ("studi primari") aventi per oggetto la valutazione di efficacia di un dato intervento ben definito nei confronti di un comparatore fisso. La peculiarità di questa categoria di studi consiste nel fatto che, a differenza degli altri tipi di studio, le revisioni non sono condotte su dati originali, studiando come unità di osservazione pazienti, ma utilizzando risultati già prodotti, e utilizzando come unità di osservazione gli studi che li hanno prodotti. Quello che distingue una revisione sistematica da una revisione narrativa fatta da un esperto dell'argomento è la sistematicità nella ricerca degli studi da includere. La revisione sistematica, come ogni studio scientifico originale, deve essere replicabile. La ricerca bibliografica alla base della revisione sistematica deve essere tale da includere il più ampio numero di articoli possibile (massimizzare la sensibilità), per essere certi di non perdere nessuno studio. Nonostante questo, non è detto che tutti gli studi condotti giungano a pubblicazione. Diversi lavori mostrano che gli studi con risultato negativo (risultato non significativo), per esempio, hanno minore probabilità di essere pubblicati rispetto a quelli positivi (bias di pubblicazione). In questo caso, la revisione sistematica potrebbe portare a risultati poco affidabili. Le revisioni sistematiche possono avere al proprio interno una meta-analisi. Il termine meta-analisi definisce il metodo statistico che viene utilizzato per combinare i risultati dei singoli studi.

8.4.1 Disegno di studio

Le revisioni sistematiche sono studi secondari che "arruolano" studi primari e vengono prodotte seguendo una serie successiva di passi:

a. individuare il quesito clinico al quale si vuole rispondere;
b. identificare gli studi da includere in base alla tipologia e alle modalità di conduzione, sviluppando una strategia di ricerca bibliografica esplicita (elenco delle "parole chiave" usate e delle loro combinazioni) di massima sensibilità applicata a più fonti (Cochrane Controlled Trials Register, MEDLINE, EMBASE, riviste specialistiche su supporto cartaceo, proceedings di congressi, riferimenti bibliografici da articoli recuperati con i metodi precedenti, fonti di studi in corso e/o non pubblicati);
c. selezionare gli studi in base a predefiniti criteri di inclusione ed esclusione;
d. valutare la qualità dello studio;
e. estrarre i dati;
f. quando appropriato, analizzare i dati (meta-analisi), con particolare attenzione all'eterogeneità;
g. interpretare i risultati e fornirli in un formato facilmente interpretabile e utilizzabile dal lettore (valutare la forza dell'evidenza e l'impatto clinico).

In queste pagine ci concentreremo sull'interpretazione dei risultati di una meta-analisi e non entreremo nel dettaglio dei metodi statistici da utilizzare per effettuare meta-analisi. Per approfondimenti sul tema, rimandiamo ai vari articoli e libri di testo disponibili sull'argomento. Segnaliamo, inoltre, l'ottimo

materiale didattico ed i tutorial disponibili sul sito Cochrane (https://training.cochrane.org/online-learning).

8.4.2 Meta-analisi: interpretazione dei risultati

I risultati delle meta-analisi vengono generalmente presentati tramite il forest plot, un grafico sintetico che rappresenta ogni studio come un albero (quadrato pieno centrale che rappresenta la stima puntuale) con due rami intorno a rappresentare l'intervallo di confidenza della stima (da qui il nome di diagramma a foresta). In figura 8.5 sono rappresentati i risultati di una meta-analisi sull'utilizzo dell'aspirina, in confronto con il placebo, nella prevenzione di infarto miocardico acuto (*Am Heart J 2011;162:115-124.e2*).

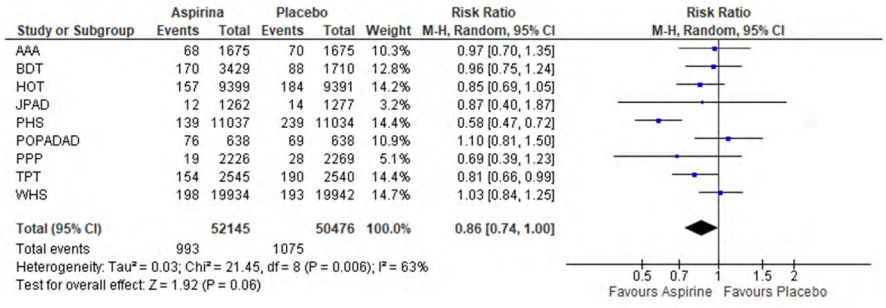

Figura 8.5 Forest plot: Aspirina vs Placebo nella prevenzione di infarto miocardico acuto. Dati tratti da *Am Heart J 2011;162:115-124.e.2.*

Nel grafico in Figura 8.5, sulle ascisse sono riportati i valori di rischio relativo (Risk Ratio, RR). La linea verticale è tracciata in corrispondenza del valore di RR = 1 (indifferenza fra i due trattamenti ovvero RR dell'ipotesi nulla). Ogni quadratino indica la stima del RR per ogni singolo studio, le barre orizzontali rappresentano l'intervallo di confidenza al 95%. La dimensione del quadratino è indicativa del peso che lo studio ha nella meta-analisi. A sinistra della linea verticale si collocano gli studi il cui risultato è a favore dell'aspirina rispetto al placebo, mentre a destra si collocano gli studi il cui risultato è a favore del placebo. Se la barra orizzontale (intervallo di confidenza) interseca la linea verticale significa che l'effetto del trattamento nel singolo studio (migliore o peggiore rispetto al placebo) non è statisticamente significativo.

Il diamante (o rombo) nella parte bassa del grafico rappresenta il risultato complessivo derivante dalla meta-analisi dell'insieme degli studi considerati. La larghezza del diamante rappresenta l'intervallo di confidenza al 95% della stima globale di RR.

9.4.3 Eterogeneità degli studi primari

Per eterogeneità degli studi primari si intende il fenomeno per il quale "n" studi effettuati per valutare l'efficacia di un trattamento in una ben definita patologia generalmente riportano "n" risultati fra loro differenti. Mentre un certo grado di eterogeneità è da considerare fisiologico, in quanto i risultati di ciascuno studio sono soggetti a variabilità casuale di campionamento, un'elevata eterogeneità fra gli studi primari solleva dubbi sul significato di una loro combinazione in un'unica stima, come si ottiene procedendo alla meta-analisi. Ciò è tanto più importante da considerare quando si possa presumere che l'eterogeneità osservata dipenda dalle condizioni in cui il singolo studio è stato condotto (modalità di reclutamento dei pazienti, modalità di somministrazione del trattamento, modalità di rilevazione degli esiti). Si parla in questo caso di eterogeneità clinica. Di questo tipo di eterogeneità occorre tenere conto sia applicando appropriati metodi di elaborazione e di analisi dei dati (ad esempio, modello ad effetti random) sia, eventualmente, rinunciando alla combinazione dei risultati qualora si ottenessero risultati non rappresentativi di alcuna specifica realtà clinica (setting) e quindi non utili o fuorvianti.

Come si può osservare dalla figura 8.5, mentre l'ampiezza degli intervalli di confidenza indica la scarsa precisione (dimensione) degli studi individuali, che rende conto anche di una notevole eterogeneità statistica fra le singole stime, la scarsa sovrapposizione fra gli intervalli stessi è espressione di risultati tra loro contrastanti e non completamente spiegabili dalla fluttuazione casuale di campionamento. Per esempio, se nel forest plot precedente confrontiamo lo studio WHS con il PHS e vediamo che i loro intervalli di confidenza non si sovrappongono, questo significa che la probabilità che, a parità delle altre condizioni, i due studi possano fornire risultati concordanti è molto bassa. Da qui il concetto che ci devono essere altri fattori (per esempio, il tipo di pazienti arruolati, il setting, il trattamento, la durata di follow up, etc) che spiegano la diversità dei risultati. Per valutare se l'eterogeneità osservata può essere spiegata solo come fenomeno statistico, si ricorre al calcolo dell'indice I^2, che esprime la percentuale di variabilità totale da attribuirsi all'eterogeneità fra studi: mentre valori di I^2 nell'ordine del 20-30% sono accettabili, valori nell'ordine del 70-75% segnalano la presenza di un eccesso di eterogeneità da spiegare (possibile eterogeneità clinica). Per un test di eterogeneità si ricorre al test Q di Cochran, che, data la sua bassa potenza, generalmente si considera significativo a partire dal p-value 0.1 (non dal tradizionale 0.05). Un'eterogeneità statisticamente significativa, facendo sospettare la presenza di diversità importanti negli studi originali, induce a cautela nel combinare e nell'applicare i risultati. Può anche essere stimolo per analizzarli con maggiore attenzione, così da identificare possibili sorgenti di variabilità, mediante delle sotto-analisi. Anche senza procedere al test formale, come abbiamo detto sopra, si può (e si deve!) valutare l'eterogeneità visivamente

analizzando il forest plot: se gli intervalli di confidenza di alcuni studi non sono fra loro sovrapposti, significa che l'eterogeneità è elevata.

8.4.4 Conclusione

In conclusione, l'evidenza ottenuta dalla revisione sistematica con meta-analisi di trial clinici randomizzati ben disegnati, ben condotti e con ridotta o assente eterogeneità clinica rappresenta il livello più elevato di informazione ottenibile in ambito scientifico per prendere decisioni pratiche nei confronti del nostro paziente. Le revisioni sistematiche hanno però due diversi obiettivi: da un lato, aiutare ad arrivare al massimo grado di "evidenza" in un determinato settore, dall'altro, in caso di elevata eterogeneità, aiutare a capire quali possono essere le cause di questa elevata eterogeneità e, quindi, in che ambito eseguire ulteriori ricerche scientifiche e come organizzarle.

Punti chiave – Revisioni sistematiche

- ✓ Le revisioni sistematiche con meta-analisi, se ben condotte, sono studi che utilizzano un approccio riproducibile, esplicito e trasparente finalizzato a minimizzare i bias nella valutazione dell'evidenza disponibile, a massimizzarne la precisione di stima e ad estenderne la generalizzabilità.
- ✓ Il risultato di una revisione sistematica di buona qualità può fornire un utile supporto per la realizzazione di buone linee guida.
- ✓ L'analisi dell'eterogeneità degli studi primari ben condotta permette di capire quali possano essere le fonti della variabilità dei risultati degli studi pubblicati.
- ✓ Attenzione al bias di pubblicazione: i risultati della revisione possono essere distorti se gli studi che segnalano l'efficacia di un trattamento vengono più facilmente pubblicati degli studi che ne segnalano l'inefficacia. Per questo sono stati sviluppati metodi statistici per valutare il fenomeno almeno presuntivamente.
- ✓ L'affidabilità dei risultati di una revisione sistematica dipende dalla qualità (assenza di bias) degli studi originali su cui si basa.
- ✓ L'eterogeneità clinica potrebbe essere tale da sconsigliare l'esecuzione della meta-analisi (mai mettere insieme mele con pere!) o rendere meno solidi i suoi risultati.

8.5 Le linee guida

Siete in Pronto Soccorso e si presenta un paziente di 28 anni con una fibrillazione atriale ad elevata frequenza ventricolare emodinamicamente stabile. Guardando attentamente l'ECG riconoscete la presenza di un fascio anomalo, suggestivo per una sindrome di Wolf Parkinson White. Chiamate il cardiologo

che consiglia la somministrazione di amiodarone endovena nel tentativo di ripristino del ritmo sinusale. Non siete tanto convinti, per fortuna finisce il vostro turno e volete imparare di più sull'argomento. Decidete di leggervi le maggiori linee guida sulla fibrillazione atriale: quelle europee, quelle americane e quelle canadesi.

Ciò che vi sorprende è che le linee guida americane considerano l'amiodarone controindicato nella sindrome di Wolf Parkinson White (2012), le europee dicono che è indicato (2014) e le canadesi non si sbilanciano! Stupiti, andate a controllare i riferimenti bibliografici e l'unico riferimento è nelle linee guida americane, ma riguarda i farmaci calcio-antagonisti! Poi guardate le ultime linee guida europee e scoprite che è stata cambiata l'indicazione, ma con un riferimento bibliografico anteriore alla pubblicazione delle linee guida analizzate ed è solo una review!

Ma il problema non è solo nelle raccomandazioni!

Ci sono spesso raccomandazioni simili, ma con riferimenti bibliografici completamente diversi, modi di graduare le evidenze eterogenei, opinioni degli esperti che cambiano nel tempo senza che ci siano state pubblicazioni sull'argomento (è interessante notare come le raccomandazioni dell'Advanced Cardiac Life Support sull'atropina nel periarresto siano completamente cambiate senza che sia stato pubblicato nulla tra due diverse versioni). Alcune linee guida considerano come apice dell'evidenza i trial clinici randomizzati controllati, ma senza considerare che le "evidenze" non dipendono solo da trial, ma dal tipo di domanda che ci si pone (per esempio, prognosi o diagnosi vs intervento) o dalla probabilità a priori dell'ipotesi testata (per fare uno studio randomizzato controllato bisogna soddisfare il concetto a priori dell'equipoise tra i due interventi, cioè che la probabilità che l'intervento sia efficace sia simile alla probabilità che non lo sia, altrimenti non sarebbe etico randomizzare i due gruppi di pazienti)

Mi gira la testa…

Come considerare allora le linee guida? Utilizzarle quotidianamente o considerarle non affidabili? E come selezionare le linee guida affidabili?

Nonostante tutti i limiti segnalati in letteratura (il conflitto di interesse, la mancanza di evidenze su molti argomenti e quindi il peso determinato dall'opinione degli esperti, l'influenza esercitata dall'autorevolezza dei relatori, la possibilità che diverse linee guida forniscano suggerimenti diversi), pensiamo che, se sviluppate in maniera rigorosa ed interpretate criticamente, siano un valido strumento e una fonte di informazioni importanti per la gestione quotidiana del paziente.

Esistono alcuni strumenti per svilupparle e per valutarne la qualità e il rigore. Uno di questi è l'AGREE II (Appraisal of Guidelines for REsearch & Evaluation), che è reperibile al sito http://www.agreetrust.org. È frutto di una collaborazione internazionale di ricercatori finalizzata a migliorare qualità ed efficacia delle linee guida, attraverso lo sviluppo di un documento condiviso che

funga da riferimento per lo sviluppo, la stesura e l'interpretazione critica delle linee guida. L'AGREE prevede una checklist per valutare la qualità di una linea guida, fornire un modello metodologico per chi scrive una linea guida, aiutare le autorità sanitarie nella scelta di quale linea guida raccomandare e aiutare il medico clinico a valutare una linea guida prima di applicarla sul paziente. La checklist prevede 24 items suddivisi in 6 aree:

1. obiettivo e motivazione;
2. coinvolgimento delle parti in causa;
3. rigore nell'elaborazione;
4. chiarezza e presentazione;
5. applicabilità;
6. indipendenza editoriale.

Inoltre, le diverse società scientifiche utilizzano vari sistemi per classificare la qualità/livello delle evidenze e la forza delle raccomandazioni. Anche se con diversa denominazione, in ogni linea guida troviamo:

- livello qualitativo di evidenza, che può essere più o meno elevato, in funzione della fonte: revisione sistematica con meta-analisi, più trial randomizzati controllati, un solo trial o semplice parere degli esperti;
- classe di raccomandazione, una procedura/trattamento è fortemente consigliato, moderatamente consigliato o sconsigliato.

Negli anni 2000 è nata una collaborazione internazionale, il gruppo di lavoro GRADE (http://www.gradeworkinggroup.org), per tentare un approccio trasparente e condiviso all'analisi della qualità delle evidenze e alla forza delle raccomandazioni presenti nelle linee guida.

Secondo i criteri GRADE le raccomandazioni sono valutate in base a:

- qualità dell'evidenza (la misura in cui posso fidarmi che la stima di un certo effetto o associazione sia corretta). Sono previsti quattro gradi di evidenza:
  - elevata;
  - moderata;
  - bassa;
  - molto bassa;
- forza della raccomandazione (la misura in cui posso fidarmi che l'aderenza ad una certa raccomandazione porti più benefici che danni) che è definita:
  - forte: sono ragionevolmente sicuro che l'adesione alla raccomandazione comporti più benefici che rischi (o viceversa);
  - debole: i benefici della raccomandazione possono o meno superare i rischi.

In questo modo, ad esempio, l'utilizzo dei diuretici nello scompenso cardiaco risulterebbe in un'indicazione di bassa evidenza (non vi sono studi randomizzati e controllati sull'argomento, poiché non sarebbe etico effettuarli), ma con una forte raccomandazione nell'utilizzarli (non ci sono dubbi che l'utilizzo dei diuretici nello scompenso cardiaco dia un beneficio). Non tutte le raccomandazioni

devono essere infatti sostenute da trial randomizzati e controllati. È molto noto un articolo provocatorio apparso sul *BMJ* in cui gli autori sostenevano che non vi sono studi sperimentali che mostrino l'efficacia del paracadute nel buttarsi dall'aereo e auspicavano che i sostenitori dell'Evidence Based Medicine (EBM, medicina basata sull'evidenza) partecipassero come volontari ad un possibile studio a riguardo! (*BMJ 2003;327:1459*)

In questa giungla di evidenze e raccomandazioni, un approccio equilibrato potrebbe essere quello di conoscere ed analizzare a fondo le linee guida relative agli argomenti di cui ci occupiamo per poterne rilevare gli aspetti di forza e di debolezza e quindi applicarle in modo critico. Nel caso di patologie che incontriamo più raramente, la critica alle linee guida non potrà che essere più superficiale. Non esistendo evidenze che le linee guida di alcune società scientifiche siano migliori di altre, un criterio potrebbe essere quello di valutare l'aggiornamento e la qualità delle evidenze fornite dalle diverse linee guida utilizzando strumenti quali *l'AGREE* e il *GRADE*. I criteri minimi che una linea guida dovrebbe avere per essere valutata sono:

- la presenza di una ricerca bibliografica sistematica sull'argomento;
- l'esplicitazione dei metodi con cui gli autori hanno valutato le evidenze e consigliato le raccomandazioni;
- l'esplicitazione dei metodi che hanno portato a decidere chi coinvolgere tra gli autori (e che vengano rappresentati tutti i possibili aspetti della patologia in questione, per esempio che siano coinvolte diverse specialità mediche, associazioni di pazienti eccetera);
- l'esplicitazione dei tempi per l'aggiornamento della linea guida.

Punti chiave – Linee guida

✓ Le linee guida sono strumenti esplicitamente creati per indirizzare i medici nella gestione di particolari situazioni cliniche concentrando le evidenze disponibili e il parere degli esperti per applicarle al singolo paziente;

✓ Possono essere un ottimo strumento per migliorare ed uniformare la pratica clinica quotidiana.

✓ Come per i trials clinici, vi è il rischio di bias: le linee guida non sono esenti da errori e non dovrebbero essere applicate acriticamente al singolo paziente. In particolare i conflitti di interesse, l'influenza esercitata dall'autorevolezza di singoli autori delle linee guida, la mancanza di rigore scientifico e di una metodologia chiara e condivisa per la loro scrittura possono introdurre degli importanti bias.

✓ Le linee guida che riguardano argomenti che conosciamo approfonditamente dovrebbero essere conosciute, analizzate ed applicate in modo critico. Per quanto riguarda le patologie che vediamo con minore frequenza, le linee gui-

da possono essere uno strumento utile per la pratica clinica, da approfondire e confrontare caso per caso con altre fonti.

✓ Strumenti utili per analizzare criticamente le linee guida (ed eventualmente per vedere come sono state scritte) possono essere l'AGREE e il GRADE (vedi appendice).

Bibliografia consigliata

Studi di prognosi

Altman DG, Bland JM. Time to event (survival) data. *BMJ*. 1997;317:468-469.

Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605.

Bland J M, Altman D G. Survival probabilities (the Kaplan-Meier method) *BMJ*. 1998; 317 :1572.

Bland JM., Altman DG. The log-rank test. *BMJ*. 2004; 328:1073-1073.

Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606.

Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ*. 2009;338:b604.

Studi osservazionali di intervento

Agoritsas T, Merglen A, Shah ND, O'Donnell M, Guyatt GH. Adjusted Analyses in Studies Addressing Therapy and Harm: Users' Guides to the Medical Literature. *JAMA*. 2017;317(7):748-759.

Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011;46(3):399-424.

Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ*. 2019;367:l5657.

Haukoos JS,. Lewis RJ The Propensity Score *JAMA*. 2015;314(15):1637-1638.

Streiner DL, Norman GR. The pros and cons of propensity scores. *Chest*. 2012;142(6):1380-1382.

Studi di non inferiorità.

D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statist Med.* 2003;22:169–86.

Garattini S, Bertelé V. Non-inferiority trials are unethical because they disregard patients' interests. *Lancet*. 2007;370: 1875–77.

Hwang IK, Morikawa T. Design issues in noninferiority/equivalence trials. *Drug Information J*. 1999;33:1205-18.

Mulla SM, Scott IA, Jackevicius CA, et al. How to Use a Noninferiority Trial. *JAMA*. 2010;308(24):2605-2611.

Piaggio G, Elbourne DR, Pocock SJ, et al. Reporting of Noninferiority and Equivalence Randomized Trials. *JAMA*. 2012;308(24):2594-2604.

Pocock SJ. The pros and cons of noninferiority trials. *Fundam Clin Pharmacol.* 2003;17: 483–490.

Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med.* 2000;1:19–21.

Revisioni sistematiche e meta-analisi

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.

Ioannidis J P A, Patsopoulos N A, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007; 335 :914.

Murad MH, Montori VM, Ioannidis JPA, et al. How to Read a Systematic Review and Meta-analysis and Apply the Results to Patient Care: Users' Guides to the Medical Literature. *JAMA*. 2014;312(2):171–179.

Sedgwick P. Meta-analyses: heterogeneity and subgroup analysis. *BMJ*. 2013; 346 :f4040

Sedgwick P. Meta-analyses: what is heterogeneity? *BMJ*. 2015; 350 :h1435

Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*. 1994 Nov 19;309(6965):1351-5.

Linee guida

Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328(7454):1490.

Gibbons RJ, Antman EM, Smith SC. Has guideline development gone astray? No. *BMJ*. 2010;340:c343.

Grol R. Has guideline development gone astray? Yes. *BMJ*. 2010;340:c306.

9. Decisioni cliniche e soglie decisionali

Un uomo di 80 anni, affetto da broncopneumopatia cronica ostruttiva (BPCO), decadimento cognitivo e obesità, arriva in Pronto Soccorso per dispnea e dolore di tipo pleuritico a livello toracico posteriore dx, presenti da circa 3 ore. I familiari riferiscono tosse ma non febbre, negano traumi, e riferiscono che nella settimana precedente il paziente è rimasto spesso a letto. All'esame obiettivo rilevati: lievi edemi declivi bilaterali, murmure vescicolare ridotto, frequenza cardiaca di 110 bpm aritmica, saturazione periferica di ossigeno 92%, pressione arteriosa 135/80 mmHg, all'ECG presenza di fibrillazione atriale; la radiografia del torace suggerisce una polmonite sinistra.

Che diagnosi fate? Polmonite, BPCO riacutizzata, scompenso cardiaco, embolia polmonare? Fareste degli altri esami per confermare o escludere la vostra diagnosi? Che terapia fareste?

9.1 Trattare o non trattare?

Spesso come medici dobbiamo decidere quali indagini diagnostiche fare o se iniziare un trattamento senza essere sicuri della diagnosi "vera". Questo tipo di scelta, che facciamo quotidianamente in modo più o meno conscio, implica in realtà molte conoscenze integrate in un percorso decisionale complesso; dovremmo infatti considerare la probabilità pre-test di malattia di quel paziente, la variazione di questa probabilità in base ai risultati dei test diagnostici (probabilità post-test), i rischi potenziali di tali test per il nostro paziente, la gravità della malattia, i rischi e i benefici del trattamento. In alcuni casi, quando la malattia è grave e il trattamento ha pochi effetti collaterali, la scelta è facile (ad esempio la somministrazione di acido folico alle gravide per prevenire spina bifida e difetti del tubo neurale). In molti casi, però, il bilancio costi/benefici delle scelte è più complesso.

Per cominciare, ammettiamo che una nuova ricerca dimostri in modo inequivocabile che mangiare un pompelmo al giorno, nel post-infarto, annulli la mortalità ad un anno portandola, ad esempio, dal 20% a zero.

Ogni medico, edotto di ciò, non dovrebbe avere remore a prescrivere ai pazienti che hanno avuto un infarto miocardico di consumare un pompelmo al giorno. In questo modo, verrebbero evitati 20 decessi per 100 trattati, anche se in 80 casi la prescrizione sarebbe inutile in quanto i pazienti non avrebbero avuto l'evento. Non potendo identificare a priori quelli che beneficeranno della cura a base di pompelmo, il medico si limiterà a considerare che, in media, ogni 5 pazienti trattati, 1 avrà un notevole beneficio (salva la vita) che, senza trattamento, non ci sarebbe stato. Questo è espresso dal valore di NNT (number

needed to treat, NNT) per il trattamento considerato, che è appunto pari a 5 (ogni 100 trattati, 20 salvati: 1 vita salvata corrisponde mediamente a 5 trattati). Elemento importante è che, presumibilmente, nessuno dei pazienti trattati inutilmente con il pompelmo avrà dei danni.

Ammettiamo ora che una successiva ricerca, altrettanto ben condotta, dimostri in modo altrettanto inequivocabile che la somministrazione quotidiana di pompelmo induce un'allergia gravissima che, nel 20% dei casi, porta al decesso. A questo punto, il medico dovrà concludere che, se non è possibile distinguere a priori il paziente che avrà un danno dal pompelmo da quello che ne avrà un vantaggio, la "dieta pompelmo" sarà da evitare.

Nella realtà, però, non si può essere assolutamente certi della diagnosi. Da questo deriva la necessità di impostare la decisione clinica considerando la probabilità di malattia, facendo riferimento alla soglia decisionale di indifferenza che ora definiremo.

9.2 La soglia decisionale

La soglia decisionale di indifferenza è la probabilità di malattia (p*) per cui il beneficio del trattamento è identico agli effetti collaterali di questo. In altre parole, se trattassimo tutti i pazienti con quella probabilità di malattia salveremmo un numero uguale di pazienti rispetto a quelli che uccideremmo per gli effetti collaterali del nostro farmaco.

Un esempio è rappresentato dal percorso diagnostico-terapeutico da seguire in caso di sospetto clinico di trombo-embolia polmonare (TEP).

È documentato che un paziente affetto da una forma medio-grave di TEP acuta, non trattata, ha un rischio di morte del 30%. È altresì documentato che tale paziente, tempestivamente trattato con terapia anticoagulante, ha un rischio di morte dell'8%. Data una riduzione del 22% del rischio di morte nei pazienti affetti da TEP che vengono trattati, possiamo approssimare a 5 il valore di NNT (NNT = 100/(30-8)).

Il trattamento con eparina comporta però dei rischi. Il danno prodotto è un sanguinamento maggiore che si può verificare in circa il 10% dei trattati e può portare a decesso con un rischio del 10%, così che per ogni 100 trattamenti, si potrebbe avere 1 decesso come danno iatrogeno (NNH nei pazienti non affetti da TEP è quindi 100/(1-0) = 100). NNT (calcolato per morte evitata) e NNH (calcolato per decesso indotto da sanguinamento maggiore) sono fra loro direttamente confrontabili.

Il bilancio fra Beneficio (B) del trattamento (1 vita salvata ogni 5 trattati affetti da TEP: B = 1/5 = 1/NNT), che la decisione di trattare assegna a tutti i pazienti con TEP, e Costo (C) del trattamento (1 vita perduta ogni 100 trattamenti effettuati: C = 1/100 = 1/NNH), che la decisione di trattare infligge a tutti i pazienti, si può esprimere con il rapporto C/B, che risulta eguale al

rapporto (1/NNH)/(1/NNT), ovvero NNT/NNH e quindi, numericamente, 5/100: 5% è quindi la probabilità di malattia per cui il beneficio dato dal trattamento è uguale al costo. Se la probabilità di presenza di TEP del paziente fosse maggiore (e non avessimo altri esami), converrebbe trattarlo con eparina, se la probabilità di embolia polmonare fosse inferiore al 5%, non converrebbe trattarlo, perché i danni del trattamento sarebbero superiori ai benefici. Solo nel caso in cui NNT è molto piccolo rispetto a NNH, allora il rapporto NNT/NNH può essere utilizzato come stima di p*; nel caso tale rapporto non fosse basso, si può utilizzare la formula riportata nel box 9.1.

Box 9.1
Una formula può aiutare

In generale, è possibile ricavare il valore soglia p* con una semplice formula senza eccessive approssimazioni. Se alla quota di malati *p* si assegna un beneficio di entità *B* e alla quota di non malati *(1-p)* si assegna un danno di entità *C*, allora il valore *p* per il quale risulta vera l'eguaglianza:

$p \cdot B = (1-p) \cdot C$ (trattando, tutti i benefici procurati compensano i danni indotti)

è proprio il valore soglia p*, che quindi risulta pari a:

$$p^* = \frac{C}{(C+B)} = \frac{(\frac{C}{B})}{(\frac{C}{B}+1)} = \frac{\text{NNT}}{(\text{NNT}+\text{NNH})} = \frac{\frac{\text{NNT}}{\text{NNH}}}{\frac{\text{NNT}}{\text{NNH}}+1}$$

Per l'esempio svolto nel testo risulta ora facile calcolare:

$$p^* = \frac{\frac{1}{20}}{\frac{1}{20}+1} = \frac{1}{(1+20)} = \frac{1}{21} = 0.048$$

Spesso il medico, mentre è convinto da un discorso di popolazione del tipo:

"Se di 100 sospetti di TEP, 5 o più sono affetti, allora si daranno più benefici che danni al gruppo trattando tutti; viceversa, se di 100 sospetti meno di 5 sono realmente affetti, allora si daranno più benefici che danni al gruppo non trattando nessuno."

resta piuttosto scettico dalla formulazione:

"Se per 1 singolo paziente sospettiamo una TEP con probabilità 5% o più, è ragionevole trattarlo con eparina, se invece la probabilità di malattia è inferiore al 5% è ragionevole evitargli il trattamento."

pensando che ciò che vale per la popolazione (approccio epidemiologico) sia diverso da ciò che vale per l'individuo (approccio clinico), guardando con scetticismo a quella disciplina che va sotto il nome di Epidemiologia Clinica. In realtà, proprio l'Epidemiologia Clinica ha portato alla messa a punto di strumenti

fondamentali per la clinica quali le "clinical prediction rules" che, definendo algoritmi per il calcolo di appropriati punteggi (scores), permettono di attribuire una probabilità ad ogni paziente sospettato di avere una specifica malattia (vedi capitolo prognosi).

Ritornando al nostro esempio, ogni paziente al quale lo score clinico assegna una probabilità di TEP superiore a 4.8% va trattato con eparina, mentre ogni paziente con probabilità assegnata di TEP inferiore a 4.8% non va trattato. Se un paziente avesse esattamente una probabilità pari a 4.8%, sarebbe indifferente la scelta fra trattarlo e non trattarlo, perché in entrambi i casi il beneficio eguaglierebbe il danno: da qui il termine di soglia di indifferenza per il valore p^*.

9.3 Utilizzo della soglia decisionale nella pratica clinica

Nella pratica clinica, però, disponiamo di altre fonti di informazione circa il reale stato del paziente. Si tratta di vari test di laboratorio o di esami di imaging clinico di varia accuratezza: test molto sensibili, anche se spesso poco specifici, come il D-dimero, utili per escludere la malattia sospettata qualora risultassero negativi (test SnNOut), o di test molto specifici anche se spesso non sensibili, come l'ecografia compressiva degli arti inferiori, utili a confermare il sospetto di malattia qualora risultassero positivi (test SpPIn). Rari sono i test utili tanto se negativi (di elevata sensibilità) quanto se positivi (di elevata specificità), come ad esempio l'angiografia polmonare. Questi ultimi generalmente sono test invasivi, la cui esecuzione può comportare di per sé un rischio di danni per il paziente.

Per un test diagnostico si paga sempre qualche cosa:

- per l'elevata sensibilità, in termini di concomitante bassa specificità;
- per l'elevata specificità, in termini di concomitante bassa sensibilità;
- per l'elevata accuratezza complessiva (ottima sensibilità e specificità), in termini di rischio clinico (esami invasivi).

Sicuramente vi ricordate che un modo utile di esprimere la quantità di informazione fornita da un test diagnostico, se positivo o negativo, è il suo rapporto di verosimiglianza o likelihood ratio (vedi paragrafo rapporti di verosimiglianza):

- se test positivo: LR+ = Sensibilità/(1-Specificità);
- se test negativo: LR– = (1-Sensibilità)/Specificità

Nella Tabella 9.1 è riportata l'utilità di alcuni test per la diagnosi di embolia polmonare acuta.

Test disponibili per diagnosi di Embolia Polmonare Acuta	LR +	LR–
D-dimero	1	0.08
Ecografia compressiva arti inferiori	13	0.63
TC torace	24	0.11
Scintigrafia polmonare (alta probabilità)	13	-
Scintigrafia polmonare (normale)	-	0.4
Angiografia	98	0.02

Tabella 9.1 Valori di LR+ e LR- per alcuni test utilizzati nella diagnosi di embolia polmonare acuta[12].

Maggiore è il valore di LR+ di un test, tanto più il suo esito positivo incrementerà la probabilità a priori di malattia (o probabilità pre-test, cioè la probabilità di malattia che attribuisco al singolo paziente prima di sottoporlo al test, calcolata in base ad uno score, ove disponibile, o attribuita arbitrariamente in base all'epidemiologia e alla clinica). Specularmente, minore è il valore di LR- di un test, tanto più il suo esito negativo ridurrà la stessa probabilità a priori di malattia. Un test perfetto dovrebbe avere LR+ tendente a infinito, LR– tendente a zero.

Come già abbiamo visto nel capitolo 5, uno strumento visuale utile a rappresentare come la probabilità pre-test viene trasformata dal risultato del test diagnostico in una probabilità post-test è dato dal nomogramma di Fagan, che si può anche utilizzare "all'inverso", ovvero per derivare per quali probabilità a priori, conoscendo il rapporto di verosimiglianza del test e la soglia di indifferenza (p*), si può escludere la patologia sospettata. Si parte cioè dalla soglia di indifferenza come probabilità post-test e, passando dal rapporto di verosimiglianza del test, si ottiene la probabilità pre-test che consente di escludere la malattia, se il test è negativo, o confermarla, se è positivo. Si possono così costruire gli intervalli di probabilità pre-test entro i quali un definito test diagnostico è utile, calcolando una coppia ulteriore di soglie decisionali: P_A, soglia di accertamento, e P_T, soglia di trattamento, in base alle quali stabilire la seguente strategia decisionale più articolata rispetto alla precedente, basata su di una sola soglia e nessuna informazione del test:

1 Roy PM, Colombet I, Durieux P, Chatellier G, Sors H, Meyer G. Systematic review and meta-analysis of strategies for the diagnosis of suspected pulmonary embolism. *BMJ*. 2005 Jul 30;331(7511):259

2 Hayashino Y, Goto M, Noguchi Y, Fukui T. Ventilation-perfusion scanning and helical CT in suspected pulmonary embolism: meta-analysis of diagnostic performance. *Radiology*. 2005;234(3):740-8.

- un paziente al quale si assegni una probabilità pre-test inferiore a P_A non andrà sottoposto a test né tanto meno a trattamento (infatti, la positività al test non porterà la probabilità post test al di sopra della soglia decisionale);
- un paziente al quale si assegni una probabilità pre-test maggiore di P_T non andrà sottoposto a test, ma direttamente trattato (infatti la negatività al test non porterà la probabilità post test al di sotto della soglia decisionale);
- un paziente al quale si assegni una probabilità pre-test compresa nell'intervallo fra P_A e P_T andrà sottoposto al test e:
  - se positivo, trattato;
  - se negativo, non trattato.

Per il calcolo della soglia di accertamento (P_A) e quella di trattamento (P_T) si veda l'appendice.

Possiamo quindi assegnare ai diversi test, in base ai loro valori di LR, gli intervalli di utilità riportati in Tabella 9.2.

Test disponibili per diagnosi di Embolia Polmonare Acuta	P_A	P_T
D-dimero	4.8%	38.5%
Ecografia compressiva arti inferiori	0.38%	7.3%
TC torace	0.21%	31.2%
Scintigrafia polmonare (alta probabilità)	0.38%	-
Scintigrafia polmonare (normale)	-	11.1%
Angiografia	0.05%	71.4%

Tabella 9.2 Valori di soglia di accertamento (P_A) e soglia di trattamento (P_T) per alcuni test utilizzati singolarmente nella diagnosi di embolia polmonare acuta.

Per controllare i risultati riportati nella tabella 9.2 è facile constatare, usando ancora una volta il Nomogramma di Fagan e i valori di LR della tabella 9.1, che un test D-dimero negativo porta una probabilità pre-test del 38.5% al di sotto della soglia p^* del 4.8%, così come un test TC negativo porta al 4.8% una probabilità pre-test del 31.2%.

Analogamente si può controllare che un test TC positivo porta al 4.8% una probabilità pre-test del 2.1 per 1000 e un test agiografico positivo porta una bassissima probabilità pre-test del 5.1 per 10000 al valore post test sempre del 4.8% (soglia di indifferenza o p^*).

Nell'attività clinica quotidiana abbiamo poi altre numerose fonti di incertezza rispetto alle stime della probabilità pre-test di malattia, dell'accuratezza diagnostica dei test, dei benefici del trattamento e degli effetti collaterali di questo (per esempio, in molte situazioni non è opportuno ragionare solo in termini di sopravvivenza). Questo per dire che la soglia decisionale non deve essere

considerata come un valore preciso, ma più come un valore indicativo da considerare nella strategia gestionale del paziente.

Per concludere, può essere utile pensare al valore di soglia decisionale poiché, se molto basso (benefici del trattamento molto superiori ai rischi), la maggior parte dei nostri pazienti avrà una probabilità pre-test bassa e gli sforzi diagnostici saranno volti ad escludere la patologia, e basterà anche solo un test positivo per superare la soglia. Viceversa, se il valore di soglia decisionale risulta molto alto (rischi di un trattamento molto superiori ai benefici), visto che la probabilità di malattia per cui vale la pena trattare sarà alta, avremo bisogno di grande sicurezza per decidere il trattamento.

Punti chiave

- ✓ Di fronte ad un sospetto diagnostico occorre declinarlo in termini di probabilità di malattia;
- ✓ La soglia decisionale di indifferenza è la probabilità di malattia (p*) per cui il beneficio del trattamento è identico agli effetti collaterali;
- ✓ La soglia di accertamento per ogni esame è la probabilità al di sotto di cui il test è inutile (il suo risultato positivo non porta la probabilità post test al di sopra della soglia di indifferenza);
- ✓ La soglia di trattamento e la probabilità di malattia al di sopra di cui il test è inutile: in ogni caso, vale la pena di trattare (i benefici del trattamento sono maggiori dei rischi);
- ✓ Il range di probabilità tra soglia di accertamento e soglia di trattamento è la probabilità pre-test per cui è utile eseguire il test (la positività o la negatività del test portano la probabilità al di sopra o al di sotto della soglia di indifferenza);
- ✓ Nonostante il concetto di soglia decisionale sia per molti versi astratto, ha senso applicarlo ad ogni patologia per capire meglio come orientare la nostra decisione clinica.

Bibliografia consigliata

Casazza G, Costantino G, Duca P. Clinical decision making: an introduction. *Intern Emerg Med.* 2010;5(6):547-52.

Weinstein MC, Fineberg HV. *L'analisi della decisione in medicina clinica.* Franco Angeli Editore – 2008.

10. Decidere in medicina[1]

10.1 Decisione ed errore

Fino ad ora abbiamo trattato numeri e certezze. Abbiamo considerato come valutare gli studi clinici e come trasferirne i risultati alla gestione dei nostri pazienti, come se i numeri trattati fossero davvero "reali". La realtà, però, è molto più complessa e le certezze sui numeri sono molto poche. In questo contesto, il medico si deve muovere in un ambito di incertezza in cui l'errore è sempre presente. In diversi studi i ricercatori hanno chiesto ai medici quale percentuale di errore ritenessero che fosse "accettabile" e la risposta più frequente è stata: inferiore all'1-2%.

Il medico è un essere umano che quotidianamente prende decisioni cercando di scegliere nella maniera più corretta possibile. In tal senso, il compito del medico non è troppo diverso da quello di altri professionisti che lavorano in settori differenti. In molti di questi settori il processo decisionale ha raggiunto livelli di efficienza estremamente elevati.

William E. Deming (ingegnere e statistico statunitense, spesso citato quando si parla di processi decisionali) ha affermato che, se dovessimo tollerare di convivere con un livello di efficienza del 99.9%, avremmo 2 atterraggi a rischio al giorno nel solo aeroporto O'Hare di Chicago e ogni ora ci sarebbero 16000 recapiti postali falliti e 32000 assegni bancari prelevati dal conto sbagliato.

In medicina la storia è ben diversa. In un noto editoriale pubblicato su *BMJ* alcuni anni fa, D. Berwick e L. Leape hanno provato ad immaginare che cosa accadrebbe se un settore notoriamente efficiente come quello dell'aviazione fosse caratterizzato dal tasso di errori attualmente riscontrato in ambito medico. L'editoriale cominciava così: «*Ladies and gentlemen, welcome aboard Sterling Airline's Flight Number 743, bound for Edinburgh. This is your captain speaking. Our flight time will be two hours, and I am pleased to report both that you have a 97% chance of reaching your destination without being significantly injured during the flight and that our chances of making a serious error during the flight, whether you are injured or not, is only 6.7%. Please fasten your seatbelts, and enjoy the flight. The weather in Edinburgh is sunny.*»[2] Il messaggio lapidario che emerge da questo editoriale è chiaro: i margini di errore riscontrati in ambito medico non sarebbero accettabili in altri settori (e pochi di noi salirebbero su quell'aereo…).

Da tempo ormai il mondo della medicina si interroga sulle ragioni di questa differenza, spesso confrontandosi direttamente con ambiti lavorativi più

1 Questo capitolo è stato curato da Fabrizio Elia.

2 Berwick DM, Leape LL. Reducing errors in medicine. *BMJ*. 1999; 319(7203): 136-7

efficienti e meno inclini all'errore. Che differenza c'è tra un medico ed un pilota di linea? Perché i medici sbagliano tanto e perché è così complicato prendere decisioni cliniche? Si potrebbero fornire molte risposte diverse a questa domanda, ma una delle più efficaci arriva da Atul Gawande in uno dei suoi libri di maggior successo. Gawande, ispirato dal lavoro di due studiosi della complessità (Brenda Zimmerman e Sholom Gouberman), sostiene che esistono: «… *three different types of problems in the world: the simple (baking a cake from a mix), the complicated (sending a rocket to the moon), and the complex (raising a child)*»[3]. Mandare in aria un razzo o pilotare un aereo rientrano tra i problemi complicati, prendere decisioni sul paziente rientra tra i problemi complessi. I problemi complessi sono quelli in cui il numero di variabili da prendere in considerazione è estremamente elevato ed in cui il livello di incertezza è altissimo.

I sistemi biologici, oggetto delle decisioni mediche, sono complessi per definizione anche se vengono quasi sempre interpretati in maniera grossolana e semplicistica, compiendo su di essi azioni altrettanto grossolane e semplicistiche. I pazienti non sono realtà statiche, ma soggetti le cui condizioni variano in maniera dinamica anche in relazione alle decisioni mediche. Il processo decisionale parte spesso da poche informazioni (peraltro, in alcuni casi parzialmente contraddittorie, non dirimenti o addirittura fuorvianti) e si realizza secondo "action-feedback loops" dove l'effetto delle decisioni prese genera nuove informazioni sulla base delle quali prendere ulteriori decisioni. Il fatto di dover prendere decisioni a partire da informazioni incomplete e di doverlo fare spesso in tempi rapidi implica che il medico si trovi ad utilizzare una sorta di "tetris strategy", in cui soluzioni abbastanza corrette prese rapidamente valgono più della soluzione migliore presa troppo lentamente. A complicare la questione, il medico si trova frequentemente ad affrontare più problemi complessi contemporaneamente. Pat Croskerry[4], riferendosi a questa situazione (ed in particolare a quella del medico di Pronto Soccorso), paragona il medico ad un acrobata del circo che deve far girare dei piatti sui bastoni senza farli mai rallentare o cadere.

Difficilmente il processo decisionale si presta quindi a modelli e categorizzazioni e, quando applicato nel mondo reale, diventa quello che è stato definito "flesh and blood" decision-making. Nel mondo reale spesso è necessario affrontare problemi per i quali le potenziali soluzioni possono essere parziali, ambigue, con obiettivi vaghi e qualche volta contraddittori ma con effetti spesso irreversibili. Non a caso, quindi, alcuni autori hanno coniato il termine di "wicked problem".

---

3 Gawande A. *The checklist manifesto: how to get things right*. Metropolitan Books, 1° edition 2010: pag 48

4 Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med.* 2003;78:775-80.

Al centro delle decisioni mediche ci sono ovviamente degli esseri umani, i pazienti, e questa rappresenta forse la maggiore differenza tra il medico ed altri professionisti che lavorano in ambiti differenti. I pazienti influenzano le decisioni mediche (nel bene o nel male) e con i pazienti il medico compie (o dovrebbe compiere) scelte condivise. I pazienti subiscono gli effetti delle decisioni mediche e, in questo senso (riprendendo la metafora circense), il medico da giocoliere si trasforma piuttosto in lanciatore di coltelli.

Con queste premesse, risulta evidente che ragionare sui processi decisionali in medicina è estremamente importante. Nei decenni passati abbiamo vissuto un cambiamento epocale con l'avvento dell'evidence based medicine. La ricerca degli strumenti migliori sulla base delle evidenze a disposizione ha rappresentato un enorme passo in avanti nella cura dei pazienti. Purtroppo, ancora oggi gran parte delle evidenze fornite dai trial clinici randomizzati controllati è incompleta, inconclusiva, assente o troppo datata. Tali evidenze devono essere integrate nel processo decisionale e completate dal ragionamento clinico per consentirci di spostare l'attenzione dal "paziente medio" degli RCT al "paziente individuale" della vita reale, e di fornire a quest'ultimo le migliori cure possibili. Proprio tenendo presente le potenzialità, ma anche i limiti, della EBM, il ragionamento e la decisione clinica assumono un valore enorme. L'interesse per questi ultimi è progressivamente aumentato e ciò è testimoniato dal crescente numero di lavori scientifici prodotti nel corso dell'ultimo decennio.

10.2 Come funziona il ragionamento clinico?

In che modo i medici prendono decisioni e come funziona il ragionamento clinico?

Il ragionamento clinico può essere rappresentato come un processo di valutazione di ipotesi incerte (le possibili diagnosi) alla luce dell'acquisizione di informazioni imperfette (i dati clinici) culminante in una decisione clinica.

Tale processo inizia a partire dall'individuazione del contesto clinico o del problema clinico (problem space) e comporta l'elaborazione di ipotesi diagnostiche compatibili. L'acquisizione di informazioni e dati clinici accompagna tutte le fasi del processo e sulla base di tali dati le ipotesi possono essere testate e confermate. Al termine di questo processo il medico è in grado di prendere decisioni associandole eventualmente a etichette diagnostiche (più o meno definitive). Le decisioni cliniche possono funzionare come ulteriori test e, sulla base dei loro effetti, le ipotesi e le etichette diagnostiche possono venire ulteriormente controllate e modificate.

Per indicare questo percorso, in letteratura vengono utilizzati indifferentemente i termini ragionamento clinico e ragionamento diagnostico. Riteniamo utile privilegiare il primo rispetto al secondo, in quanto l'obiettivo del ragionamento è quello di arrivare a una decisione piuttosto che a una diagnosi. La

diagnosi è un'etichetta utilizzata a un certo momento nel corso del ragionamento clinico per identificare una malattia o uno stato morboso. Ma il medico può essere chiamato a prendere decisioni (trattare o non trattare, come trattare, eseguire ulteriori test oppure non eseguirli) quando un'etichetta diagnostica non è stata ancora posta.

Il ragionamento clinico risulta essere un processo aperto e dinamico che può essere affrontato con modalità differenti. Tali modalità caratterizzano i processi decisionali in qualsiasi ambito del ragionamento umano e sono state oggetto di studio delle scienze cognitive negli ultimi decenni. Ciò che le scienze cognitive hanno messo in luce è che, quando si tratta di fare delle scelte e prendere decisioni, l'essere umano è in grado di utilizzare due strategie cognitive differenti e in buona parte indipendenti: un pensiero veloce o non analitico, che ci permette di reagire in maniera automatica a differenti situazioni in brevissimo tempo, e un pensiero lento o analitico, che ci permette di dare risposte razionalmente fondate. Questo modello interpretativo del ragionamento umano è definito *dual process framework*. Mentre la prima modalità di pensiero (definita Sistema 1) è sempre attiva e scarsamente controllabile, l'attivazione della seconda modalità (definita Sistema 2) richiede tempo e sforzo. Poiché siamo esseri tendenzialmente economici, il pensiero veloce spesso si impone su quello lento, funzionando come una sorta di pilota automatico.

Contrariamente a presupposti teorici ancora diffusi, le nostre capacità cognitive sono limitate (Herbert Simon ha utilizzato per primo il termine di bounded rationality): la nostra memoria è fallibile, la capacità di ottenere informazioni è limitata e la capacità di processare tali informazioni è scarsa. Per questo motivo il pensiero veloce si è imposto come strumento decisivo, acquisito nel corso dell'evoluzione, che ci ha consentito di sopravvivere permettendoci di prendere decisioni rapide utilizzando il minor numero di informazioni a nostra disposizione.

Una delle caratteristiche del pensiero veloce è quella di servirsi di regole intuitive, dette euristiche. Si tratta di scorciatoie cognitive, anche definite come regole del pollice (rules of thumb), ovvero regole veloci e frugali (fast and frugal) in grado di farci prendere decisioni nella maniera più economica possibile. Utilizziamo tali regole in maniera automatica ogni qual volta dobbiamo prendere delle decisioni e nella maggior parte dei casi queste regole funzionano. Anche in ambito medico, esse ci consentono di fare scelte corrette nella maggior parte dei casi. Purtroppo, quando ci troviamo nei casi che sfuggono all'euristica che stiamo utilizzando, il pensiero veloce ci può portare a compiere azioni che non controlliamo, rischiando di provocare eventi avversi. È in questi casi che si producono gli errori cognitivi (o bias). Tali errori sono sistematici e prevedibili: come le illusioni ottiche, ci spingono (quasi) tutti nella stessa direzione.

La caratteristica sorprendente degli errori cognitivi è proprio la prevedibilità e la sistematicità con cui essi si verificano, indipendentemente da cultura, grado di istruzione, età, sesso, esperienza ed estrazione. Una classica illustrazione è rappresentata dal problema seguente: "una pallina e una racchetta da tennis costano 1 euro e 10 centesimi; la racchetta costa 1 euro più della pallina: quanto costa la pallina?". Praticamente a tutti viene in mente una risposta istantanea, 10 centesimi. È sbagliata (la risposta corretta è 5 centesimi), ma molti intuitivamente l'accettano. Chi non sbaglia è stato educato a riflettere e a formalizzare i problemi utilizzando un approccio analitico. Queste caratteristiche cognitive sono proprie di tutti gli esseri umani e quindi anche dei medici.

10.3 Pensiero lento o pensiero veloce?

Il ragionamento clinico affrontato in maniera lenta ed analitica si avvicina al metodo ipotetico-deduttivo ed in molti casi è stato descritto come tale. Le diverse ipotesi diagnostiche generate sulla base del contesto clinico vengono considerate più o meno probabili in funzione della loro maggiore o minore diffusione nella popolazione cui il paziente appartiene per età, sesso e sintomatologia. Il medico raccoglie informazioni aggiuntive, che la conoscenza clinica identifica come maggiormente indicative, di alcune delle possibili diagnosi a discapito di altre e applica test clinici. L'iniziale plausibilità delle patologie considerate e la rilevanza degli ulteriori dati raccolti permetteranno così di "aggiornare" la probabilità delle diverse ipotesi diagnostiche, fino al punto in cui qualcuna di esse avrà raggiunto un livello di affidabilità sufficiente per orientare l'azione.

Il fondamento scientifico di tale processo è costituito dal teorema di Bayes e dalla teoria dell'utilità attesa. Il primo, come abbiamo già descritto in buona parte del testo, ci consente di definire l'impatto delle nuove informazioni acquisite sulla stima della probabilità di una certa patologia. Conoscendo la probabilità pre-test di una certa malattia e il valore informativo del test, siamo in grado di definire la probabilità di malattia in base al risultato del test stesso. Secondo la teoria dell'utilità attesa, se tale probabilità supera la soglia decisionale di trattamento (il livello di probabilità al di sopra del quale i benefici del trattamento sono superiori ai danni) è indicato trattare il paziente. Al contrario, se si trova al di sotto di tale soglia, è indicato effettuare altri test diagnostici per innalzare ulteriormente il grado di certezza diagnostica.

Evidentemente l'applicazione di questa metodologia risulta complessa e richiede un enorme sforzo cognitivo. Ecco perché nella maggior parte dei casi il medico affronta i problemi clinici utilizzando la modalità rapida ed automatica. L'utilizzo delle euristiche rappresenta una strategia economica e rapida per risolvere problemi clinici. Le euristiche funzionano come scorciatoie all'interno del percorso decisionale, consentendoci di saltare alcune delle tappe o di

velocizzarle. Il numero di euristiche conosciute e utilizzate è progressivamente cresciuto nel corso degli anni. L'applicazione di tali strumenti nella pratica clinica in contesti differenti è stato ampiamente discusso in letteratura e un elenco sistematico è disponibile in alcune rassegne recentemente pubblicate. Nella maggior parte di queste pubblicazioni emerge il ruolo delle euristiche come fonte di errori cognitivi, assegnando a questi strumenti una connotazione negativa. La realtà è che, sebbene in alcune situazioni possano effettivamente produrre degli errori, nella maggior parte dei casi le euristiche funzionano e garantiscono buoni risultati.

È verosimile che nel corso del ragionamento clinico (così come nella vita di tutti i giorni) la modalità rapida si imponga su quella lenta funzionando come una sorta di pilota automatico. In caso di necessità, tuttavia, ciascun individuo è in grado di ricorrere alla modalità lenta. Nessuna delle due modalità si è finora dimostrata superiore all'altra e la validità di ciascuna dipende dai contesti clinici e dall'esperienza dell'operatore nell'affrontare un certo problema clinico. Di fronte a problemi clinici ben noti, il medico tenderà con maggiore probabilità ad utilizzare la modalità rapida, di fronte a problemi clinici meno conosciuti, utilizzerà con maggiore probabilità la modalità lenta. Entrambe, in ogni caso, sono potenziali fonti di errori. Occorre, inoltre, dire che alcune strategie operative sono il risultato del tentativo di unire i vantaggi del sistema lento (la sistematicità) con quelli del sistema rapido (la rapidità e l'economicità). Gli algoritmi decisionali, ben noti a chi si occupa di Medicina d'Urgenza, appartengono a questo gruppo di strategie. Essi utilizzano alberi decisionali che, partendo da un argomento, sviluppano una serie di domande a cascata con risposte spesso dicotomiche che conducono alla domanda successiva. Risultano estremamente utili quando il contesto clinico prevede una diagnostica differenziale limitata, nelle situazioni in cui è necessario prendere decisioni rapide e in cui è importante non dimenticare alcune diagnosi fondamentali (gli algoritmi dei corsi ALS ed ATLS rappresentano un buon esempio).

10.4 Bias cognitivi

Una paziente di 80 anni viene portata in PS lamentando astenia, malessere, nausea e difficoltà alla stazione eretta. Viene riferita una sintomatologia simil-influenzale risalente a circa 10 giorni prima.

All'arrivo la paziente è normotesa, apiretica, con scambi respiratori adeguati.

L'anamnesi è significativa per la presenza di una fibrillazione atriale permanente non sottoposta a terapia anticoagulante (per scelta della paziente) ed una arteriopatia polidistrettuale. Nel corso dell'ultimo anno la paziente è stata ricoverata in seguito a due episodi di polmonite.

Gli esami ematochimici rilevano un modesto incremento degli indici di flogosi. L'ECG evidenzia una FA a media penetranza ventricolare. In considerazione

dell'incremento degli indici di flogosi e dei recenti ricoveri per polmonite, la paziente viene sottoposta a radiografia del torace eseguita a decubito supino. Il radiologo segnala un sospetto addensamento alla base di sinistra.

Nell'ipotesi diagnostica di polmonite viene iniziato un trattamento con liquidi ed antibiotici. Nelle ore successive, due medici differenti, pur non rilevando febbre né sintomi o segni polmonari suggestivi, confermano la diagnosi e proseguono la terapia in corso.

Circa 16 ore dopo l'arrivo, si verifica un nuovo episodio sincopale associato a nausea e malessere. All'esame obiettivo viene rilevata una pressione di 110/50 a destra, mentre a sinistra la pressione risulta non rilevabile. A questo punto, nell'ipotesi diagnostica di dissecazione aortica, viene richiesta una TC encefalo-torace-addome con mdc. Tale esame, tuttavia, esclude la dissecazione e non rileva addensamenti polmonari compatibili con focolai flogistici.

La paziente viene comunque ricoverata in medicina d'urgenza e sottoposta a monitoraggio. All'arrivo, il medico d'urgenza conferma l'asimmetria di pressione, ma rileva anche la presenza di un arto superiore più freddo rispetto al controlaterale in assenza di dolore o deficit neurologici. In considerazione di questo dato, viene chiesto al radiologo di riesaminare le immagini TC; alla rivalutazione viene evidenziata l'occlusione dell'arteria succlavia di sinistra. La paziente viene pertanto portata in sala operatoria e sottoposta ad embolectomia con ripristino di un circolo adeguato. Dopo l'intervento, in considerazione della ripresa di perfusione dell'arto e del provvisorio miglioramento clinico, viene riportata in reparto.

Nei giorni successivi tuttavia la paziente continua a lamentare malessere con difficoltà alla stazione eretta e riferisce inoltre diplopia. Nell'ipotesi di un evento ischemico a carico del circolo posteriore viene ripetuta una TC encefalo. Tale esame rileva la presenza di una lesione ischemica a livello cerebellare. Rivedendo le immagini della prima TC, l'area ischemica risultava già presente, anche se meno demarcata rispetto alle immagini ottenute successivamente. La paziente viene quindi trasferita in neurologia con diagnosi di evento ischemico cerebellare di origine cardio-embolica associato ad embolismo dell'arteria succlavia.

10.4.1 Ancore e conferme

Il caso clinico descritto mette in luce una serie di errori cognitivi frequentemente osservabili durante il ragionamento clinico. La prima diagnosi di polmonite viene formulata sulla base di elementi diagnostici deboli ed in assenza di febbre oltre che di sintomi e reperti polmonari suggestivi. Il fatto che la paziente sia già stata ricoverata recentemente per tale patologia spinge il medico ad orientarsi verso questo tipo di diagnosi. Rimane quindi vittima del posterior probability bias che consiste nel considerare più probabile una certa patologia in considerazione del fatto che si è già verificata in precedenza.

Un ulteriore errore in cui sono incorsi i medici nella gestione di questa paziente è l'anchoring, ovvero la tendenza ad imprimere un certo valore ad alcuni indizi diagnostici iniziali in base ai quali viene formulata la diagnosi, senza mettere in discussione quest'ultima nel corso del processo diagnostico (quando nuovi elementi diagnostici vengono alla luce). Nel caso della nostra paziente, i dati anamnestici (due recenti ricoveri per polmonite) e quelli di laboratorio hanno funzionato da ancora spingendo verso la diagnosi di polmonite. È utile notare che tre medici in successione sono caduti vittima dell'ancoraggio, non tenendo conto in maniera opportuna di altre ipotesi in competizione. L'ancoraggio rappresenta un'euristica estremamente potente e comunemente utilizzata. In un'analisi effettuata tra specializzandi di medicina interna, l'ancoraggio è risultato il bias cognitivo più frequente tra quelli indagati. I medici rischiano di caderne vittime non solo durante il processo diagnostico. L'ancoraggio può manifestarsi nella valutazione dei sintomi di un paziente nel momento in cui ci siamo già fatti un'idea della patologia sottostante (sottostimando per esempio l'entità del dolore) oppure può verificarsi nel momento in cui il medico deve decidere che tipo di trattamento mettere in atto. In un noto studio dell'associazione nazionale pediatrica americana veniva chiesto a 20 pediatri di visitare 400 bambini non ancora sottoposti a tonsillectomia ed indicare quali di questi avessero necessità di essere sottoposti a tale intervento. L'intervento veniva consigliato al 45% dei bambini. I bambini che non avevano ricevuto indicazione venivano sottoposti ad una seconda visita presso un medico differente ed in questo caso il 46% dei pazienti riceveva l'indicazione alla tonsillectomia. I bambini per i quali nelle due visite precedenti non era stato ritenuto necessario l'intervento venivano sottoposti ad un terzo consulto; al 44% di loro veniva consigliata la tonsillectomia. Probabilmente i pediatri erano ancorati all'idea comune che circa il 50% dei pazienti con tonsilliti ricorrenti in età pediatrica necessitino di tonsillectomia e non hanno saputo ridiscutere il proprio approccio terapeutico alla luce dei dati clinici subentranti. In ultimo, l'ancoraggio può giocare un ruolo significativo anche nel processo di ricerca di informazioni mediche.

Un ulteriore errore cognitivo strettamente correlato all'anchoring è il confirmation bias, ovvero la tendenza a cercare dati che consentano di confermare la diagnosi piuttosto che di rigettarla. Nel caso della nostra paziente vengono effettuati esami mirati a confermare l'ipotesi iniziale. Il reperto riscontrato alla lastra del torace è servito a supportare l'ipotesi di polmonite, senza considerare i limiti del radiogramma toracico effettuato a paziente supino ed in una singola proiezione.

10.4.2 Questione di disponibilità

Rilevando l'asimmetria di pressione, il medico di turno pone il sospetto di dissecazione aortica, non considerando l'ipotesi di un'ischemia dell'arto

superiore (anche dopo aver ricevuto il referto della TC). Cade probabilmente vittima dell'availability bias, un comune errore cognitivo che consiste nel considerare una diagnosi come più probabile in base alla relativa facilità con cui esempi simili vengono in mente.

In un noto film italiano di qualche anno fa, il regista Nanni Moretti narra una vicenda occorsagli nella vita reale. Vittima per mesi di un intenso prurito, si reca in visita presso numerosi studi di dermatologia. Tutti i dermatologi consultati non risultano in grado di scovare la causa del problema, che si rivelerà essere un linfoma di Hodgkin. La maggior parte di loro incorre nell'availability bias. I dermatologi visitano comunemente pazienti con prurito e nella maggior parte dei casi tale sintomo è dovuto a problemi dermatologici. Essi sanno che il linfoma ed altre patologie internistiche possono causare prurito, ma nella maggior parte dei casi i malati affetti da tali patologie non arrivano alla loro attenzione. Tali cause risultano pertanto meno accessibili nella loro memoria. L'availability bias è, in effetti, un errore tipico dei medici specialisti, ma può verificarsi anche nei casi in cui una certa patologia sia stata recentemente riconosciuta o studiata o nei casi in cui una diagnosi passata abbia prodotto un forte impatto sul medico (per esempio, una diagnosi il cui l'avvenuto o il mancato riconoscimento ha influito sulla prognosi del paziente). Anche la pressione dei mass media può indurre un availability bias. Esiste una correlazione tra la copertura mediatica riguardante una patologia (tipicamente, un'infezione a facile diffusione) e l'incremento nell'utilizzo di test diagnostici specifici per quella malattia. Più un evento clinico è oggetto di attenzione da parte dei mass media, più facilmente sarà accessibile nella memoria dei medici.

Tornando al caso della nostra paziente, possiamo dire che la dissecazione aortica è una diagnosi alla quale il medico d'urgenza è preparato e che lo tiene costantemente all'erta a causa delle conseguenze potenzialmente letali di tale patologia. Al contrario, l'ischemia acuta dell'arto superiore è un evento meno frequente e meno accessibile nella memoria (soprattutto quando si presenta in assenza di sintomi e segni specifici come il dolore ed i deficit neurologici).

10.4.3 L'importanza della cornice

Il radiologo che interpreta la prima TC viene influenzato dal quesito clinico che gli viene posto (escludere o confermare la dissecazione aortica). Quando il giudizio clinico viene influenzato dalle modalità con cui una situazione clinica viene presentata, può verificarsi un errore cognitivo noto come framing effect. Tale errore è ben illustrato dal "mito del ragno fantasma", noto ai medici d'urgenza. Un paziente si presenta in pronto soccorso con una lesione cutanea di origine non ben definita (ascesso, foruncolo, etc. etc.), sostenendo di essere stato punto da un ragno. Il paziente tuttavia non si è reso conto della puntura, i familiari negano di aver visto un ragno e l'abitazione in cui il paziente vive non è abitualmente popolata da ragni. Nonostante manchi qualsiasi evidenza,

il paziente viene dimesso con la diagnosi di puntura di insetto. In questo caso, è il paziente a spingere il medico verso la diagnosi sbagliata, in altre situazioni sono i colleghi stessi a mettere fuori strada il medico con racconti distorti di un caso clinico. Una categoria particolarmente a rischio è rappresentata proprio dai radiologi, i quali ricevono la richiesta di esecuzione di un esame strumentale accompagnata da una sintesi clinica o da un quesito diagnostico. Una sintesi clinica scorretta od un quesito mal posto rischiano tuttavia di portare fuori strada il radiologo, che focalizzerà la propria attenzione sui dettagli sbagliati o interpreterà i dati in maniera scorretta.

Anche l'impatto degli studi clinici può variare a seconda di come i dati vengono presentati. Presentare la riduzione di mortalità associata ad un farmaco in termini di rischio relativo piuttosto che di differenza di rischio assoluto consente di amplificare gli effetti del farmaco agli occhi del lettore di uno studio clinico. In maniera similare, l'utilizzo di immagini ad elevato impatto (per esempio quelle tratte da una risonanza magnetica funzionale del cervello), piuttosto che di banali grafici a barre, consente di incorniciare i dati clinici rendendoli illusoriamente più significativi agli occhi del lettore. Nel caso della nostra paziente, il radiologo, influenzato dal quesito clinico, esclude correttamente la dissecazione aortica, ma non pone attenzione a ciò che sta attorno, ovvero alla presenza di una occlusione dell'arteria succlavia.

10.4.4 Soddisfatti dalla ricerca

L'approccio utilizzato dal radiologo riflette anche un altro errore cognitivo, il search satisfying bias (o satisfaction of search). Tale errore viene generato dalla tendenza a cessare la ricerca quando si è trovato qualcosa. Purtroppo, ciò che emerge all'inizio del percorso diagnostico spesso rappresenta solo una parte del problema o addirittura risulta essere un riscontro accidentale o fortuito. Un esempio di tale errore è ben rappresentato dal detto comune secondo cui la frattura più spesso dimenticata dai radiologi è sempre la seconda. Questo avviene perché, una volta riscontrata la prima frattura, il radiologo si accontenta e sospende la ricerca.

10.4.5 Una fine prematura

Anche la diagnosi di ischemia cerebellare, i cui sintomi hanno condotto la malata in pronto soccorso, ha subito un ritardo di alcuni giorni. Dopo aver effettuato la diagnosi di occlusione dell'arteria succlavia, i medici hanno chiuso prematuramente il caso (la premature closure costituisce un altro errore cognitivo), imputando tutta la sintomatologia a tale evento clinico.

Punti chiave

- ✓ Esistono problemi semplici (cucinare una torta), complicati (mandare un razzo sulla luna) o complessi (crescere un figlio).
- ✓ Il nostro cervello utilizza due modi di pensiero diversi il sistema 1, intuitivo e veloce, e il sistema 2, analitico, razionale.
- ✓ Nella vita quotidiana, come nella medicina, utilizziamo per la maggior parte del tempo il sistema 1.
- ✓ Il sistema 1, nonostante sia molto efficiente, può essere fonte di errore e gravato da bias cognitivi.
- ✓ Nel ragionamento clinico dobbiamo essere consapevoli dei limiti dei dati disponibili e del modo di funzionamento del nostro cervello per ridurre le possibilità di errore.

Bibliografia consigliata

American Child Health Association Research Division. *Physical Defects: The Pathway to Correction*. New York, NY: American Child Health Association; 1934.

Ashman CJ, Yu JS, Wolfman D. Satisfaction of search in osteoradiology. *AJR Am J Roentgenol.* 2000;175:541-4.

Berwick DM, Leape LL. Reducing errors in medicine. *BMJ*. 1999; 319(7203): 136-7.

Berbaum KS, Schartz KM, Caldwell RT, Madsen MT, Thompson BH, Mullan BF, Ellingson AN, Franken EA Jr. Satisfaction of search from detection of pulmonary nodules in computed tomography of the chest. *Acad Radiol.* 2013;20:194-201.

Brezis M, Halpern-Reichert D, Schwaber MJ. Mass media-induced availability bias in the clinical suspicion of West Nile fever. *Ann Intern Med.* 2004;140:234-5.

Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med.* 2003;78:775-80.

Crupi V, Gensini GF, Motterlini M. *La dimensione cognitiva dell'errore in Medicina*. Franco Angeli Editore 2006.

Fandel TM, Pfnur M, Schafer SC, Bacchetti P, Mast FW, Corinth C, Ansorge M, Melchior SW, Thüroff JW, Kirkpatrick CJ, Lehr HA. Do we truly see what we think we see? The role of cognitive bias in pathological interpretation. *J Pathol.* 2008; 216: 193–200.

Gawande A. *The checklist manifesto: how to get things right*. Metropolitan Books, 1° edition 2010.

Gigerenzer G (2008) *Gut feelings: the intelligence of the unconscious*. Viking Press, New York.

Gladwell M (2005) *Blink: The power of thinking without thinking*. Little, Brown and Co, New York.

Groopman J. *How doctors think*. Houghton Mifflin Company. 1° edition 2007.

Kahneman D (2011) *Thinking fast and slow*. Macmillan, New York.

Kahneman D, Frederick S (2002) Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich T, Griffin DW, Kahneman D (eds) *Heuristics and biases*. Cambridge University Press, New York, pp 49–81.

Lau AY, Coiera EW. Do people experience cognitive biases while searching for information? *J Am Med Inform Assoc*. 2007;14:599-608.

Leape LL. Error in medicine. *JAMA*. 1994; 272 (23): 1851-1857.

Mamede S, van Gog T, van den Berge K, Rikers RM, van Saase JL, van Guldener C, Schmidt HG. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA*. 2010;304:1198-203.

Marewski JN, Gigerenzer G (2012) Heuristic decision making in medicine. *Dialogues Clin Neurosci*. 14:77–89.

McCabe DP, Castel AD. Seeing is believing: the effect of brain images on judgments of scientific reasoning. *Cognition*. 2008;107:343-52.

Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Acad Med*. 2017;92(1):23-30.

Ogdie AR, Reilly JB, Pang WG, Keddem S, Barg FK, Von Feldt JM, Myers JS. Seen through their eyes: residents' reflections on the cognitive and contextual components of diagnostic errors in medicine. *Acad Med*. 2012;87:1361-7.

Perneger TV, Agoritsas T. Doctors and patients' susceptibility to framing bias: a randomized trial. *J Gen Intern Med*. 2011;26:1411-7.

Phua DH, Tan NC. Cognitive aspect of diagnostic errors. *Ann Acad Med Singapore*. 2013;42:33-41.

Pines JM. Profiles in patient safety: confirmation bias in emergency medicine. *Acad Emerg Med*. 2006;13:90-4.

Reason J. *Human error*. 1990. Cambridge University Press, New York.

Rittel HWJ, Webber MM. Dilemmas in a general theory of planning. *Policy Sci* 1973; 4:155–169.

Riva P, Rusconi P, Montali L, Cherubini P. The influence of anchoring on pain judgment. *J Pain Symptom Manage*. 2011;42:265-77.

Schmidt HG, Mamede S, van den Berge K, van Gog T, van Saase JL, Rikers RM. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Acad Med*. 2014;89:285-91.

Self WH, Courtney DM, McNaughton CD, Wunderink RG, Kline JA. High discordance of chest x-ray and computed tomography for detection of

pulmonary opacities in ED patients: implications for diagnosing pneumonia. *Am J Emerg Med*. 2013;31:401-5.

Sniderman AD, LaChapelle KJ, Rachon NA, Furberg CD. The necessity for clinical reasoning in the era of evidence-based medicine. *Mayo Clin Proc*. 2013; 88(10): 1108-14.

Sternbach G. The phantom spider and other myths. *J Emerg Med*. 2012;42:457-8.

Stripe SC, Best LG, Cole-Harding S, Fifield B, Talebdoost F. Aviation model cognitive risk factors applied to medical malpractice cases. *J Am Board Fam Med* 2006; 19 (6):627–632.

Wears RL. What makes diagnosis hard? *Adv Health Sci Educ Theory Pract*. 2009; 14(Suppl 1):19–25.

Wegwarth O, Gaissmaier W, Gigerenzer G. Smart strategies for doctors and doctors-in-training: heuristics in medicine. *Med Educ*. 2009;43:721–728.

11. Conclusioni

11.1 Incertezza e sopravvivenza

Abbiamo iniziato questo testo partendo dalla ricerca bibliografica, da domande che emergono quotidianamente dalla nostra pratica clinica e abbiamo cercato di illustrare come fare per ricercare e interpretare al meglio le informazioni disponibili, per poi applicarle in modo critico al singolo paziente. Abbiamo finito parlando di errori cognitivi e decisione clinica.

Esistono molti libri di "statistica" o epidemiologia clinica sicuramente scritti meglio e più completi di questo. Spesso, coloro che si occupano di metodologia sono metodologi che non gestiscono quotidianamente pazienti, mentre coloro che si occupano di clinica sono medici un po' a disagio quando hanno a che fare con formule e numeri. Il nostro intento è stato quello di integrare i due punti di vista, quello statistico/metodologico e quello clinico. Siamo convinti che per fare bene la clinica serva una solida base metodologica, e, viceversa, per occuparsi di metodologia in questo ambito si debba sempre tenere presente che l'obiettivo finale è clinico: il benessere del paziente. Per questi motivi, abbiamo deciso di strutturare i vari capitoli partendo sempre da un caso clinico e di illustrare i concetti di metodologia collegati al caso di partenza. Abbiamo, inoltre, deciso di non essere esaustivi, ma di restare ad un livello superficiale, rimandando i lettori più desiderosi di approfondimento ai vari riferimenti riportati nella bibliografia consigliata. L'obiettivo principale del nostro volume è quello di favorire un approccio critico alla letteratura scientifica e alla decisione clinica. L'approccio critico, per definizione, mette in crisi tutte le certezze. L'assenza di certezze è la costante che caratterizza l'attività clinica quotidiana: l'unica certezza che abbiamo è che faremo degli errori. A volte ci renderemo conto di questi errori, a volte non ce ne accorgeremo ma, talvolta, qualcun altro ce li farà notare. L'insegnamento più importante che possiamo trarre è quello di metterci in discussione e di imparare sempre dai nostri errori. Insieme a questo, un po' di umiltà, gentilezza e ironia certamente non guastano! Questo per dire che la costante assoluta della professione medica è proprio l'incertezza, l'insicurezza di ogni scelta. A questo si può rispondere in tre diversi modi: innanzitutto, facendo finta che l'incertezza non esista, e quindi fingendo una sicurezza non possibile; oppure rimandando ad altri le decisioni; infine, ed è la scelta che riteniamo più corretta, imparando a convivere con l'incertezza, apprezzando la bellezza di non avere sempre risposte scontate e di quanto si debba usare l'intelligenza per affrontare situazioni complesse in modo sensato. E possedere qualche strumento metodologico essenziale per la valutazione critica dell'evidenza può essere

un ottimo punto di partenza nell'affrontare tutta questa incertezza. Apprezzare l'incertezza significa riconoscere i nostri limiti e ridurre la possibilità di burn out per il futuro. Con tutta questa incertezza, serve davvero essere un po' ironici e, soprattutto, gentili!

Bibliografia consigliata

Schulz CM. Oggi ho preso 120 decisioni... tutte sbagliate! *Celebrate Peanuts 60 years*. Vol. 7 Dalai Editore 2010 ISBN: 8860737478.

Ringraziamenti

Si ringraziano per il prezioso supporto:

Elisa Ceriani, Simone Birocchi, Giulia Cernuschi, Franca Dipaola, Piergiorgio Duca, Fabrizio Elia, Francesca Perego, Gian Marco Podda, Maria Teresa Pugliano, Anna Maria Rusconi, Monica Solbiati, Federica Tordato e il Gruppo di Autoformazione Metodologica (GrAM).

Appendice

a. Griglie per la valutazione degli studi clinici

È possibile trovare in letteratura una serie di strumenti che hanno lo scopo di aiutare chi esegue gli studi clinici (siano essi di efficacia o di accuratezza diagnostica) a riportare in maniera corretta e chiara i risultati, e forniscono al lettore una griglia di lettura che permette di valutare la qualità dello studio e di capire se sono presenti bias più o meno evidenti.

Segnaliamo il sito di EQUATOR Network (http://www.equator-network.org/), un'iniziativa internazionale che promuove l'affidabilità e il valore della letteratura medica, promuovendo una stesura trasparente e accurata degli studi.

Sul sito è presente molto materiale per autori, editori e revisori sull'argomento e c'è una pagina con link a siti di strumenti per riportare in maniera corretta e chiara i risultati di uno studio o fornire ai lettori una griglia di lettura e vari siti di metodologia, statistica e organizzazioni internazionali.

Esempi di griglie utili per valutare la qualità del reporting degli studi sono:

- STARD (Standards for Reporting of Diagnostic Accuracy)
- CONSORT (CONsolidated Standards of Reporting Trials)
- AGREE II (Appraisal of Guidelines Research&Evaluation in Europe)
- GRADE (Grading of Recommendations Assessment, Development and Evaluation)
- STROBE (STrengthening the Reporting of OBservational studies in Epidemiology)
- PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)

b. Statistica k di Cohen

La statistica *k* (nota come *k* di Cohen) si applica allo studio di variabili categoriche, nominali (esempio: patologia presente/assente) o ordinali (esempio: patologia lieve/moderata/severa), comunque mutuamente esclusive. Le variabili continue possono essere analizzate in modo più appropriato con altri metodi.

Il coefficiente *k* esprime la concordanza corretta per il caso (cioè oltre la proporzione di concordanza attesa per effetto del caso):

$$\kappa = \frac{P_O - P_C}{1 - P_C}$$

dove P_o = proporzione di concordanza osservata (o globale);
P_c = proporzione di concordanza attesa per effetto del caso;
$1 - P_c$ = proporzione di concordanza massima non dovuta al caso;

Nella figura A.1 viene tradotto in veste grafica lo stesso concetto: la concordanza potenziale esprime la perfetta descrizione della realtà da parte, per esempio, di due operatori (in una realtà ideale, tutte le radiografie del torace individuano correttamente la polmonite e tutti gli operatori che osservano la radiografia giungono alla stessa conclusione). La concordanza potenziale è formata dalla concordanza legata al caso (ad esempio, due operatori individuano le stesse polmoniti scrivendo a caso il referto) sommata alla potenziale concordanza al di là del caso (ad esempio, due operatori leggono correttamente la radiografia del torace, perché abbiamo la fortuna di avere nel nostro ospedale i due campioni del mondo a pari merito di lettura delle radiografie del torace, che non sono comunque perfetti, altrimenti avremmo una concordanza legata al caso pari a zero). Tuttavia, nella realtà clinica, ciò non succede praticamente mai, pertanto la concordanza osservata è rappresentata da un segmento più piccolo rispetto a quello della concordanza potenziale. Quindi, mentre la concordanza legata al caso rimane fissa (per quale ragione il caso dovrebbe cambiare il suo comportamento?) quella che si riduce, in misura variabile, è la concordanza che si raggiunge effettivamente al di là del caso.

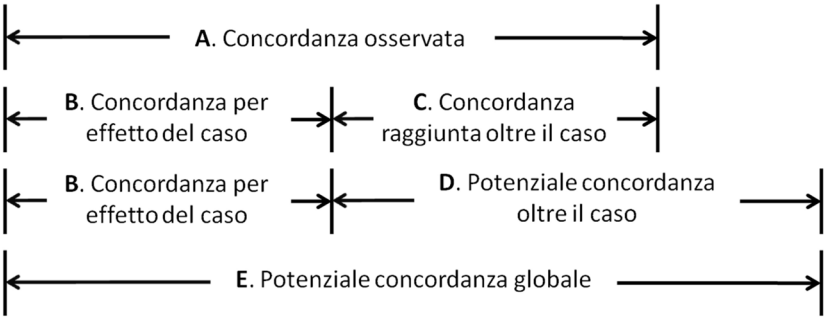

Figura A.1 Relazione tra k e la concordanza globale e casuale (adattato da Phys Ther 2005;85:257-268).

Le misurazioni di due medici relativamente ad una scala nominale a due categorie (esame positivo/negativo per malattia) possono essere rappresentate in una tabella di contingenza 2x2.

In tabella A.1 possiamo vedere il giudizio di 2 medici su 39 radiografie del torace per presenza di segni di scompenso cardiaco.

		Medico 1		
		Positivo	**Negativo**	**Totale**
Medico 2	**Positivo**	(a) 22	(b) 2	(PosM2) 24
	Negativo	(c) 4	(d) 11	(NegM2) 15
	Totale	(PosM1) 26	(NegM1) 13	(n) **39**

Tabella A.1 Valutazione di 39 radiografie del torace da parte di 2 medici.

(a) e (d) indicano i casi (positivi e negativi) su cui i 2 medici concordano
(b) e (c) indicano i casi su cui i 2 medici sono in disaccordo

$$P_o = \frac{a+d}{n} = \frac{22+11}{39} = 0.84$$ (proporzione di concordanza osservata)

La proporzione di concordanza attesa per effetto del caso si calcola presumendo che ogni medico concordi con l'altro, per ciascuna categoria (positivo o negativo), in una proporzione pari a quella che esiste tra il suo giudizio (positivo o negativo) rispetto al totale.

In pratica, nell'esempio precedente, il primo medico ha refertato come positive 24/39 radiografie del torace, il che corrisponde al 61.5% degli esami refertati. Se operasse solo il caso, ci aspetteremmo che, se il primo medico refertasse le 26 radiografie del torace che il secondo medico ha refertato come positive per scompenso, la proporzione di giudizio positivo in questo sottoinsieme di 26 radiografie sarebbe analoga alla percentuale di radiografie positive rispetto al totale delle radiografie, perciò 61.5x26/100=16.

Lo stesso accade per le radiografie refertate come negative: 15/39=38.5% è la percentuale di esami giudicati negativi dal primo medico rispetto al totale; lo stesso medico per effetto del caso giudicherà negativa un'analoga percentuale radiografie refertate come negative dal secondo operatore, perciò 38.5x13/100=5.

Sommando poi il numero di esami in cui i due operatori concordano per effetto del caso e dividendo la somma per il numero totale delle radiografie, si ottiene la proporzione di concordanza attesa per effetto del caso.

Tradotto in formula:

$$P_C = \frac{(\text{PosM1}\times\text{PosM2})/n+(\text{NegM1}\times\text{NegM2})/n}{n} = \frac{(26\times 24)/39 + (13\times 15)/39}{39} = \frac{16+5}{39} = 0.53$$

(proporzione di concordanza attesa per effetto del caso)

$$\kappa = \frac{P_O - P_C}{1 - P_C} = \frac{0.84 - 0.53}{1 - 0.53} = 0.66$$

Il coefficiente ***k*** può essere utilizzato anche per scale con più di due categorie nominali o ordinali (ad esempio, dolore assente/lieve/moderato/severo); in questo caso, andrà considerata la natura gerarchica delle categorie in esame, poiché il disaccordo tra categorie vicine nella scala (ad esempio, tra dolore assente e lieve) è meno grave del disaccordo tra categorie distanti nella scala (ad esempio, tra dolore assente e moderato). Per rappresentare il grado di disaccordo, il coefficiente ***k*** può essere "pesato".

c. Odds ratio

Odds e probabilità sono due modi di esprimere la propensione al verificarsi di un evento. Gli odds vengono utilizzati soprattutto nel mondo delle scommesse. L'odds è definito come il rapporto tra numero di eventi favorevoli e numero di eventi non favorevoli, mentre la probabilità sappiamo che è il rapporto fra numero di eventi favorevoli e numero di eventi possibili. Odds e probabilità risultano così in corrispondenza biunivoca fra di loro: data una probabilità (p) se ne ricava facilmente l'odds facendo p/(1-p) e dato un valore di odds se ne ricava facilmente il corrispondente valore di probabilità facendo odds/(odds+1).

		TVP		
		Sì	**No**	**Totale**
Scompenso cardiaco	**Sì**	32	130	**162**
	No	742	7610	**8352**
	Totale	**774**	**7740**	**8514**

Tabella A.2 Scompenso cardiaco e trombosi venosa profonda (TVP) in 2729 pazienti.

Prendiamo l'esempio dello studio caso-controllo riportato nel testo. Abbiamo visto che possiamo calcolare la probabilità ("rischio") di essere scompensato per i casi e per i controlli, rispettivamente pari a 32/774 e 130/7740. Ora, la frequenza di scompensati fra casi e controlli può essere paragonata, oltre che mediante confronto delle rispettive percentuali (vale a dire scompensati sul totale), mediante confronto dei rispettivi odds (scompensati su non scompensati). Nel nostro esempio, 32/742 e 130/7610 rappresentano appunto gli odds di scompenso fra casi e controlli. Gli odds possono, quindi, essere visti in un certo senso come probabilità riscalate. Una probabilità, come abbiamo visto in precedenza, è un numero compreso fra 0 ed 1 che quantifica la fiducia nel verificarsi di un evento. Quando stimiamo la

probabilità del verificarsi di un evento mediante l'approccio frequentista, non facciamo altro che calcolare la percentuale (frequenza relativa) dei soggetti nei quali si è verificato l'evento. Un odds, così come una probabilità, misura il grado di fiducia che nutriamo nel verificarsi di un evento. A differenza della probabilità, l'odds è però un numero che varia fra 0 e + e quando stimiamo un odds non facciamo altro che rapportare il numero di soggetti con evento al numero di soggetti senza evento. Tornando al nostro esempio: 32/774=0.041 (probabilità) significa che su 100 soggetti complessivi del nostro campione, 4.1 sono scompensati; 32/742=0.043 (odds) significa invece che, nel campione, per ogni 100 soggetti non scompensati ce ne sono 4.3 scompensati. Come vedremo più avanti, utilizzando il termine odds come sinonimo di rischio (abbiamo appunto appena visto che l'odds è una stima di rischio differente dalla probabilità) possiamo calcolare una misura di associazione analoga al RR.

Come abbiamo visto, l'odds può anche essere visto come il rapporto tra la probabilità p di un evento e la probabilità (1-p) dell'evento complementare:

$$\text{odds} = \frac{p}{(1-p)}$$

Viceversa, a partire dall'odds, la probabilità può essere calcolata come:

$$p = \frac{\text{odds}}{(\text{odds} + 1)}$$

L'odds ratio (OR) è il rapporto tra due odds: tornando al nostro esempio, è il rapporto tra gli odds (che possiamo vedere come un "rischio") di scompenso fra i casi e gli odds ("rischio") di scompenso fra i controlli. Il suo valore ci dice quanto è più (o meno) frequente osservare scompensati fra i casi rispetto ai controlli. Grazie ad alcune proprietà matematiche degli odds, si può vedere come il rapporto fra il "rischio" (odds) di scompenso fra i casi di TVP e il "rischio" (odds) di scompenso fra i controlli non TVP sia esattamente identico al rapporto fra il "rischio" (odds) di TVP fra gli scompensati ed il "rischio" (odds) di TVP fra i non scompensati. Questo ci permette quindi di interpretare l'OR esattamente come se fosse un RR.

L'odds ratio (OR) è il rapporto tra odds: il rapporto tra l'odds di un evento in un gruppo (esposizione nei casi, ad esempio) e l'odds dello stesso evento in un secondo gruppo (esposizione fra i controlli). Può essere utilizzato per quantificare un'associazione tra una variabile indipendente ("causa": esposizione a un fattore di rischio, una terapia o una strategia preventiva; nel nostro esempio, la presenza di scompenso cardiaco) e una variabile dipendente ("effetto": malattia; nel nostro esempio, l'insorgenza di TVP).

		Malattia	
		Sì (casi)	**No (controlli)**
Esposizione	**Sì (esposti)**	A	B
	No (non esposti)	**C**	D

Tabella A.3 tabella 2x2 di esposizione e malattia.

In riferimento alla Tabella A.3, ricordiamo che negli studi prospettici il rischio relativo (RR), ovvero il rapporto tra il rischio di malattia negli esposti e nei non esposti, è dato da:

$$\text{RR} = \frac{\text{rischio esposti}}{\text{rischio non esposti}} = \frac{\frac{A}{(A+B)}}{\frac{C}{(C+D)}}$$

In uno studio retrospettivo abbiamo visto che non possiamo calcolare il RR, ma ha senso calcolare l'OR, definito come rapporto fra odds di malattia fra gli esposti e odds di malattia fra i non esposti: rispettivamente (A/B) e (C/D):

$$\text{OR di outcome (malattia)} = \frac{\text{odds di malattia esposti}}{\text{odds di malattia non esposti}} = \frac{\frac{A}{B}}{\frac{C}{D}} = \frac{A}{B} \times \frac{D}{C} = \frac{\text{A} \times \text{D}}{\text{B} \times \text{C}}$$

L'OR può venire facilmente calcolato facendo il rapporto fra i prodotti incrociati delle celle della tabella 2x2 AxD e BxC. Da rilevare che, quando l'outcome è raro (indicativamente con un'incidenza minore dell'1%), come capita spesso in studi di coorte, A e C sono piccoli rispetto a B e D, perciò A/B e C/D sono buone approssimazioni di A/(A+B) e di C/(C+D), da cui risulta anche che OR è una buona approssimazione del più direttamente interpretabile RR.

L'OR è stato introdotto come misura di associazione (tra esposizione-trattamento e outcome) negli studi in cui non è nota la prevalenza di malattia, ovvero negli studi caso-controllo, quando non è possibile calcolare l'incidenza (rischio) di una data patologia nei due gruppi di esposizione, ma è possibile stimare l'esposizione nei due gruppi di condizione di salute (malati=casi; non malati=controlli).

Per rispondere alla nostra domanda su qual è il rischio di TVP nei soggetti affetti da scompenso, non possiamo calcolare l'incidenza di TVP nei pazienti con e senza scompenso cardiaco, ma possiamo stimare la prevalenza di scompenso nei pazienti con e senza TVP. Il ricorso all'odds ci svincola in questo caso dal dover conoscere la prevalenza di malattia.

Gli studi caso-controllo sono utili per studiare outcome rari. Nel caso di eventi rari, uno studio prospettico richiederebbe l'arruolamento di un elevatissimo

numero di soggetti, per ottenere un sufficiente numero di eventi. Arruolando a partire dagli eventi, uno studio caso-controllo, invece, consente di arruolare molti meno pazienti.

In uno studio caso-controllo possiamo, quindi, partire dall'outcome e calcolare l'odds per un soggetto che ha sviluppato una certa malattia di essere stato esposto ad un fattore di rischio (A/C) e l'odds che un soggetto che non ha sviluppato la malattia abbia avuto l'esposizione allo stesso fattore (B/D). Il loro rapporto ci dice se l'odds (la possibilità) di essere stato esposto ad un fattore di rischio è maggiore nei soggetti che hanno sviluppato la malattia.

$$\text{OR di esposizione} = \frac{\text{odds di esposizione tra i malati}}{\text{odds di esposizione tra i non malati}} = \frac{\frac{A}{C}}{\frac{B}{D}} = \frac{A}{C} \times \frac{D}{B} = \frac{\text{A} \times \text{D}}{\text{B} \times \text{C}}$$

Possiamo osservare come, dal punto di vista matematico, l'OR di outcome (prima formula) e l'OR di esposizione (seconda formula) siano equivalenti, non entrando nel calcolo, in nessun caso, i marginali della tabella (cioè i totali di riga e colonna).

d. Formule per il calcolo degli intervalli di confidenza al 95%

Di seguito proponiamo le formule per il calcolo degli intervalli di confidenza (IC) al 95% in base all'approssimazione normale. La forma generale di questi intervalli prevede il calcolo dell'errore standard, e per OR, RR e LR siamo in grado, mediante formula, di definire l'errore standard solo della forma logaritmica (Ln, logaritmo naturale). Per questo motivo, per il calcolo degli IC al 95% di OR, RR e LR, si deve prima effettuare la trasformazione in scala logaritmica, poi calcolare gli estremi della differenza dei due logaritmi e poi ritornare alla scala originale di partenza facendo l'esponenziale degli estremi logaritmici. Per questo motivo, nelle formule per il calcolo degli IC menzionati è presente E*xp*, che significa che si deve calcolare l'esponenziale. Come abbiamo detto, in tutti gli esempi riportati sotto assumiamo che la normale sia una buona approssimazione, come solitamente accade per campioni di dimensione non piccola. Nel caso di campioni di relativamente piccole dimensioni, si dovrebbe ricorrere ad altri metodi di calcolo.

Studi di associazione fra esposizione (a fattori di rischio o trattamenti) ed eventi: RCT, studi osservazionali prospettici o caso-controllo

		Evento		
		Sì	**No**	**Totale**
Esposizione	**Sì**	a	b	**a+b**
	No	**c**	d	**c+d**
	Totale	**a+c**	**b+d**	

Tabella A.4 Tabella 2x2 per studi di associazione fra esposizione ed eventi.

Odds Ratio (OR)

$$\mathrm{OR} = \frac{a \times d}{b \times c} \qquad \mathrm{ES}(\mathrm{Ln}(\mathrm{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Limiti IC 95%: $\mathrm{Exp}\left(\mathrm{Ln}(\mathrm{OR}) \pm 1.96 \cdot \mathrm{ES}(\mathrm{Ln}(\mathrm{OR}))\right)$

Rischio Relativo (RR)

$$\mathrm{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \qquad \mathrm{ES}(\mathrm{Ln}(\mathrm{RR})) = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

Limiti IC 95%: $\mathrm{Exp}\left(\mathrm{Ln}(\mathrm{RR}) \pm 1.96 \cdot \mathrm{ES}(\mathrm{Ln}(\mathrm{RR}))\right)$

Riduzione Assoluta di Rischio (ARR)

$$\mathrm{ARR} = \frac{a}{a+b} - \frac{c}{c+d} \qquad \mathrm{ES}(\mathrm{ARR}) = \sqrt{\frac{a}{a+b} + \frac{c}{c+d}}$$

Limiti IC 95%: $\mathrm{ARR} \pm 1.96 \cdot \mathrm{ES}(\mathrm{ARR})$

Studi di accuratezza diagnostica.

		Malati	**Non Malati**	**Totale**
Index test	**+**	Veri positivi (VP)	Falsi positivi (FP)	**VP + FP**
	-	Falsi negativi (FN)	Veri negativi (VN)	**FN + VN**
	Totale	**VP + FN**	**FP + VN**	

Tabella A.5 Tabella 2x2 per studi di accuratezza di un test diagnostico.

$$\text{sensibilità} = \frac{VP}{VP + FN}$$

$$ES(\text{sensibilità}) = \sqrt{\frac{\text{sensibilità} \cdot (1 - \text{sensibilità})}{VP + FN}}$$

Limiti IC 95%: $\text{sensibilità} \pm 1.96 \cdot ES(\text{sensibilità})$

$$\text{specificità} = \frac{VN}{FP + VN}$$

$$ES(\text{specificità}) = \sqrt{\frac{\text{specificità} \cdot (1 - \text{specificità})}{FP + VN}}$$

Limiti IC 95%: $\text{specificità} \pm 1.96 \cdot ES(\text{specificità})$

$$VPP = \frac{VP}{VP + FP}$$

$$ES(VPP) = \sqrt{\frac{VPP \cdot (1 - VPP)}{VP + FP}}$$

Limiti IC 95%: $VPP \pm 1.96 \cdot ES(VPP)$

$$VPN = \frac{VN}{FN + VN}$$

$$ES(VPN) = \sqrt{\frac{VPN \cdot (1 - VPN)}{FN + VN}}$$

Limiti IC 95%: $VPN \pm 1.96 \cdot ES(VPN)$

$$LR^{+} = \frac{\frac{VP}{VP + FN}}{\frac{FP}{FP + VN}}$$

$$ES(Ln(LR^{+})) = \sqrt{\frac{1}{VP} - \frac{1}{VP + FN} + \frac{1}{FP} - \frac{1}{FP + VN}}$$

Limiti IC 95%: $Exp(Ln(LR^{+}) \pm 1.96 \cdot ES(Ln(LR^{+}))$

$$LR^{-} = \frac{\frac{FN}{VP + FN}}{\frac{VN}{FP + VN}}$$

$$ES(Ln(LR^{-})) = \sqrt{\frac{1}{FN} - \frac{1}{VP + FN} + \frac{1}{VN} - \frac{1}{FP + VN}}$$

Limiti IC 95%: $Exp(Ln(LR^{-}) \pm 1.96 \cdot ES(Ln(LR^{-}))$

e. Formule per il calcolo della soglia di accertamento e della soglia di trattamento

Per calcolare la soglia di accertamento (P_A) e quella di trattamento (P_T) per un determinato tipo di test, occorre conoscere rispettivamente LR^{+} ed LR^{-} del test diagnostico, oltre che il rapporto C/B del trattamento. Infatti, si possono ricavare le soglie applicando le seguenti formule:

$$P_A = \frac{C/B}{C/B + (LR+)}$$

$$P_T = \frac{C/B}{C/B + (LR-)}$$

Queste due formule derivano rispettivamente dalla soluzione di queste due equazioni:

– B x P x SE = C x (1 – P) x (1 – SP) che dice che un test è utile, se positivo, quando $p<p^*$ supera la soglia P_A per la quale il beneficio dei veri positivi bilancia il costo per i falsi positivi.

– $B \times P \times (1 - SE) = C \times (1 - P) \times SP$ che dice che un test è utile, se negativo, quando $p>p^*$ è inferiore alla soglia P_T per la quale il mancato beneficio inflitto ai falsi negativi è bilanciato dal risparmio dei costi per i veri negativi.

Allo stesso risultato, naturalmente, si può arrivare, per prove successive iterate, utilizzando il Nomogramma di Fagan, come il lettore può facilmente, ma faticosamente, provare a fare usandolo all'inverso.

Glossario

Analisi di sensibilità (sensitivity analysis): modalità di analisi che consiste, quando sono possibili diversi scenari, nel provare a eseguire le analisi nel contesto dello scenario migliore e dello scenario peggiore, valutando le eventuali differenze.

Campione: sottoinsieme di unità statistiche estratte da una popolazione di interesse, allo scopo di studiarne alcune caratteristiche salienti per poi estendere all'intera popolazione i risultati ottenuti sul campione.

Curva ROC (Receiver Operator Characteristics): esamina, per un test con variabile continua, la performance di un test lungo tutto il range dei valori possibili (esempio, diversi cut-off di D-dimero). Un'area sotto la curva di 1 denota un test perfetto, mentre un'area di 0.5 indica un test che ha le stesse probabilità di classificare di un lancio di monetina. Le coordinate di ogni punto della curva sono il tasso di veri positivi pari alla sensibilità (y) ed il tasso di falsi positivi, pari a 1-specificità (x), in corrispondenza di un certo cut-off. L'area sotto la curva ROC esprime il potere diagnostico del test.

Endpoint: la misura dell'esito clinico di uno studio.

Endpoint primario: la misura dell'esito dell'obiettivo principale dello studio; su di esso viene calcolata la numerosità del campione. Il risultato dell'endpoint primario è probante, cioè può essere applicato alla pratica clinica.

Endpoint secondario: è la misura dell'obiettivo (o degli obiettivi) secondario dello studio; a differenza dell'endpoint primario, non è probante.

Endpoint hard: endpoint direttamente rilevante per il paziente (ad esempio mortalità, dialisi, rischio di recidiva infartuale, rischio di ictus).

Endpoint surrogato: endpoint che si ritiene correlato a un endpoint hard, ma di più facile ottenimento (ad esempio, la riduzione del colesterolo correlata alla riduzione del rischio di infarto). Uno studio con un endpoint surrogato non è mai probante (non permette di modificare la pratica clinica in base ai suoi risultati).

Errore di I tipo (rischio di): α indica la probabilità di commettere un errore del I tipo nella verifica di un'ipotesi statistica. Nel caso di verifica di ipotesi relativa all'efficacia di un nuovo trattamento rispetto al vecchio, è la probabilità di trovare che il nuovo trattamento è più efficace del vecchio quando invece non lo è (equivale a un risultato falso positivo in caso di un test diagnostico). Generalmente è fissato al 5%.

Errore di II tipo (rischio di): β indica la probabilità di commettere un errore del II tipo nella verifica di un'ipotesi statistica. Nel caso di verifica di ipotesi relativa all'efficacia di un nuovo trattamento rispetto al vecchio, è la probabilità di trovare che il nuovo trattamento non è più efficace del vecchio quando invece lo è (equivale a un risultato falso negativo in caso di un test diagnostico). Generalmente è fissato al 20% (in alcuni casi anche al 10%).

Errore casuale: errore che si commette nella stima di una certa quantità (ad esempio OR, RR) a causa dell'errore campionario. Si riduce all'aumentare della dimensione del campione.

Errore sistematico o bias: errore che si commette nella stima di una certa quantità a causa di distorsioni introdotte, ad esempio, con disegni di studio non appropriati (bias di selezione) o con modalità di esecuzione non idonee (detection bias, performance bias). Si riduce con disegni di studio appropriati.

Fattore di rischio: è un elemento o un comportamento che può contribuire al manifestarsi o all'insorgere di un effetto patologico. Si parla di fattori di rischio nell'ambito della definizione di una rete causale.

Hazard ratio (HR): misura di associazione che esprime il rapporto fra il rischio istantaneo dell'evento di interesse in un gruppo rispetto ad un altro (esempio: trattati vs non trattati). Utilizzato tipicamente quando si conducono analisi di sopravvivenza (time-to-event analysis, ci interessa non solo se, ma anche quando l'evento si verifica), è solitamente ottenuto da un modello di regressione di Cox. Si interpreta come un rischio relativo.

Incidenza: numero di nuovi casi di malattia osservati in una popolazione nell'unità di tempo.

Index test: è un test che si vuole valutare in uno studio di accuratezza diagnostica e che viene confrontato con il test di riferimento (reference standard) per fare diagnosi in quella particolare condizione.

Intention-to-treat: modalità di analisi dei dati di un trial clinico basata sul gruppo di randomizzazione. I pazienti sono considerati, ai fini delle analisi, appartenenti al gruppo a cui sono stati randomizzati, anche se nel corso dello studio passano all'altro gruppo o interrompono tutti i trattamenti.

Intervallo di confidenza (solitamente al 95%): insieme di valori che, in base al campione osservato, riteniamo plausibili per il parametro (ad esempio una media, RR, OR, HR, e tutte le altre misure viste) di nostro interesse. Siamo ragionevolmente certi (confidenti al 95%) che il vero valore del parametro che stiamo stimando sia uno fra quelli appartenenti all'intervallo di confidenza. Concetto di stima intervallare.

Likelihood ratio (o rapporto di verosimiglianza) negativo (LR-): rapporto tra la proporzione di pazienti negativi al test tra i malati e i negativi al test tra i non malati. Più il suo valore è basso e più il test è informativo: la sua negatività riduce la probabilità di malattia. Se LR- è uguale a 1 il test è inutile (la probabilità a priori di malattia è identica alla probabilità di malattia dopo il test).

Likelihood ratio (o rapporto di verosimiglianza) positivo (LR+): rapporto tra la proporzione di pazienti positivi al test tra i malati e i positivi al test tra i non malati. Più il suo valore è alto e più il test è informativo: la sua positività aumenta la probabilità di malattia. Se LR+ è uguale a 1 il test è inutile (la probabilità a priori di malattia è identica alla probabilità di malattia dopo il test).

Meta-analisi: metodologia statistica utilizzata nelle revisioni sistematiche finalizzata a combinare i risultati di più studi su uno stesso argomento (ad esempio, un particolare trattamento, un certo test diagnostico) fornendo così una stima sintetica.

Modello di regressione: approccio statistico che permette di valutare l'associazione fra una (o più) variabile indipendente (causa) e una variabile dipendente (effetto). La variabile dipendente è solitamente uno degli endpoint dello studio. Si possono avere modelli univariati (una sola variabile dipendente, una sola variabile indipendente) oppure multivariati (i puristi utilizzano il termine di regressione multipla, una sola variabile dipendente, due o più variabili indipendenti).

Modello di regressione logistica: modello di regressione in cui la variabile dipendente è di tipo qualitativo (evento: decesso, stroke, ospedalizzazione…). Permette quindi di valutare l'associazione fra una (o più) variabile indipendente (causa) ed il rischio di evento di interesse (effetto). I risultati sono generalmente espressi mediante odds ratio.

Modello di regressione di Cox: modello di regressione in cui si parte da una variabile dipendente di tipo qualitativo (evento: decesso, stroke, ospedalizzazione…), ma in cui si tiene conto anche del tempo a cui si verificano gli eventi (time-to-event analysis). In sostanza, permette, quindi, di valutare l'associazione fra una (o più) variabile indipendente (causa) ed il tempo al verificarsi dell'evento di interesse (effetto). I risultati sono generalmente espressi mediante hazard ratio.

Number Needed to Treat (NNT): numero di pazienti da trattare per ottenere un beneficio col trattamento.

Number Needed to Harm (NNH): numero di pazienti da trattare per avere un effetto collaterale col trattamento.

Odds ratio (OR): misura di associazione fra due variabili usata per stimare il rischio relativo in casi particolari (ad esempio, negli studi caso controllo). È una buona stima del rischio relativo in caso di eventi rari. È utilizzata anche in studi prospettici quando, per fini di analisi, è opportuno utilizzare modelli di regressione logistica.

Per-protocol: modalità di analisi dei dati di un trial clinico che misura l'efficacia di un trattamento considerando solo i pazienti che hanno rispettato il protocollo, essendo realmente esposti ai trattamenti pianificati. Si contrappone all'analisi per intention-to-treat.

Potenza di uno studio: probabilità di vedere una differenza tra due trattamenti quando questa esiste; è tanto maggiore quanto è più grande il campione. Generalmente è fissata all'80% (in alcuni casi anche 90%). È il complemento a uno dell'errore di secondo tipo (1-β).

Prevalenza: frequenza di soggetti della popolazione che ha una certa condizione in un dato momento.

Probabilità post-test: probabilità a posteriori (cioè dopo l'esecuzione di un test diagnostico) che quel paziente abbia una certa malattia.

Probabilità pre-test: probabilità a priori (ad esempio, prima di eseguire un certo test) che quel paziente abbia una certa malattia.

Randomizzazione: assegnazione casuale di un trattamento nell'ambito di un trial clinico, finalizzata a ottenere due gruppi di pazienti simili per tutti i fattori prognostici noti e non noti, eliminando i bias di selezione (selection bias) nell'assegnazione dei trattamenti.

Reference standard (o, più impropriamente, gold standard): è un test (o procedura) diagnostico che classifica i pazienti come malati o non malati e rappresenta il test migliore per la diagnosi della malattia in studio (test di riferimento).

Revisione narrativa: revisione della letteratura senza criteri prestabiliti. Ad esempio, un esperto scrive un articolo sulla terapia dello scompenso cardiaco.

Revisione sistematica: rassegna sistematica di tutti i lavori pubblicati aventi per oggetto la valutazione di efficacia di una data terapia per una certa patologia oppure la valutazione di accuratezza diagnostica di un particolare test. Differisce dalla revisione narrativa in quanto non riflette il parere di singoli esperti, ma rappresenta un processo finalizzato a ricerca trasparente e con criteri prestabiliti di studi su un certo argomento per avere migliore sintesi possibile delle informazioni.

Riduzione di Rischio Assoluto (**ARR**, absolute risk reduction): differenza rischio assoluto tra due gruppi (ad esempio, non trattati e trattati).

Rischio assoluto (RA): rapporto fra il numero di soggetti che hanno manifestato l'evento (ad esempio, il decesso) ed il numero totale dei soggetti considerati.

Rischio relativo (RR): rapporto fra due rischi assoluti (ad esempio, nei trattati rispetto ai non trattati), esprime quanto è più probabile che un evento avverso si manifesti nei trattati rispetto ai non trattati.

Sensibilità: probabilità di test positivo nei malati.

Soglia decisionale di indifferenza: probabilità di malattia al di sopra di cui i benefici di trattare il paziente per la patologia in questione superano i rischi dati dagli effetti collaterali del trattamento.

Specificità: probabilità di test negativo nei non malati.

Statistica k (di Cohen): statistica utilizzata per calcolare la concordanza tra operatori escludendo l'effetto del caso.

Test SnNOut: test ad a elevata sensibilità, utile, quando fornisce esito negativo, per escludere una patologia.

Test SpPIn: test a elevata specificità, utile, quando fornisce esito positivo, per confermare la presenza di una malattia.

Valore predittivo negativo (VPN): probabilità che il soggetto non sia malato se il test è negativo.

Valore predittivo positivo (VPP): probabilità che il soggetto sia malato se il test è positivo.

Introduzione all'approccio critico alla decisione clinica

Giovanni Casazza e Giorgio Costantino

Questo volume vuole essere uno strumento democratico e rivoluzionario, rivolto a chi pensa sia importante fare medicina in modo critico e che gli unici Maestri siano i pazienti. La pubblicazione, leggera ma rigorosa, fornisce un'introduzione alla valutazione critica della letteratura scientifica utile nella pratica clinica quotidiana. Partendo dal quesito clinico si affrontano tutti gli aspetti utili nella decisione clinica: la ricerca bibliografica, l'interpretazione e l'applicazione dei risultati degli studi al singolo paziente, gli elementi di base del processo cognitivo decisionale. Ogni argomento inizia con un caso clinico reale, in modo da facilitare la comprensione e la rilevanza pratica degli argomenti trattati. La presenza di un glossario, di un'appendice per gli argomenti più tecnici e una ricca bibliografia completano il volume.

In copertina: Lighthouse storm, bgianfilippo, immagine disponibile sul sito pxhere.com, CC0 – public domain, https://pxhere.com/it/photo/1655016

ISBN 979-12-55101-02-4 (print)
ISBN 979-12-55101-03-1 (PDF)
ISBN 979-12-55101-04-8 (EPUB)
DOI 10.54103/milanoup.164