Segnalare agli epidemiologi la qualità degli studi: il rischio della cattiva informazione

Communicating the quality of research to the epidemiologists: the risks of misinformation

Danilo Fusco,[1] Fulvia Seccareccia,[2] Anna Patrizia Barone,[1] Paola D'Errigo,[2] Carlo A. Perucci[1]

[1] *Dipartimento di epidemiologia, ASL Roma E, Roma*

[2] *CNESPS, Istituto superiore di sanità*

***Corrispondenza:** Danilo Fusco; e-mail: fusco@asplazio.it*

I temi dell'epidemiologia valutativa, in particolare gli studi osservazionali che hanno l'obiettivo di misurare comparativamente l'efficacia di trattamenti sanitari, rappresentano oggi un ambito importante del lavoro dell'epidemiologia in Italia. In questi studi, con disegni appropriati, gli interventi sanitari (preventivi, diagnostici, terapeutici, riabilitativi) e le loro modalità di produzione (tipo di servizio, soggetti erogatori, volumi di attività, popolazioni) rappresentano le esposizioni di interesse. La stima della misura di associazione tra queste ultime ed esiti diversi (morte, complicanze, ricoveri, disabilità) deve tener conto di possibili confondenti e/o modificatori, ipotizzati e misurati in ciascuno studio. Infatti, a causa della natura osservazionale di questi studi, l'allocazione all'esposizione non è casuale; il ricercatore deve quindi perseguire il miglior equilibrio tra «contrasto dell'esposizione» e «controllo del confondimento». L'effetto del confondimento ha le caratteristiche di un errore sistematico e ha a che fare con la validità della stima; obiettivo del ricercatore è ottenere stime delle misure di associazione il più possibile valide, minimizzare la differenza tra la stima osservata e quella (del tutto teorica) che si sarebbe ottenuta in uno studio randomizzato controllato. Nel campo della valutazione comparativa degli esiti l'insieme delle metodologie di controllo del confondimento e valutazione dell'interazione è spesso chiamato *risk adjustment*; a questo tema recentemente *E&P* ha dedicato una monografia.[1] Riteniamo che il problema della validità delle misure di effetto/associazione debba essere considerato come dominante in questi studi, in fase di disegno, analisi e discussione. In ogni caso, è noto che le misure di associazione sono affette da errore casuale. Questo errore, non avendo caratteristiche di sistematicità, tende a ridursi all'aumentare del numero dei soggetti studiati e il suo effetto, per ogni dimensione campionaria data, può essere quantificato in due modi:

1. affiancando al valore puntuale della stima un intervallo, detto «intervallo di confidenza», a cui si attribuisce una probabilità predeterminata di contenere il valore «vero» della misura considerata;

2. calcolando la probabilità (p-value) che le differenze osservate tra gli esiti nei gruppi posti a confronto siano interamente dovute all'effetto dell'errore casuale, laddove invece non ci sia differenza (ipotesi nulla) tra gli esiti «veri» dei gruppi stessi.

Occorre evitare confusione tra precisione delle stime, quindi loro «significatività statistica», e loro validità. Citiamo a questo proposito Rothman: «*the p-value is supposed to be a probability measure. Furthermore, when the data are very discrepant with the null hypothesis, the p-value is small, and when the data are concordant with the null hypothesis, the p-value is large. …It does not describe the probability that the null hypothesis is correct, however. The null hypothesis may be the most reasonable hypothesis for the data even if the p value is low, or it may be implausible or just incorrect even if the p-value is high*».[2]

Inoltre, così come *deconfounding* e «contrasto dell'esposizione» competono ed è empiricamente impossibile massimizzare contemporaneamente entrambi, lo stesso si può affermare anche per validità e precisione. Per definizione l'assenza di randomizzazione nella «realtà» osservata richiede il controllo del confondimento: idealmente tutti i fattori associati a esposizione ed esito misurati dovrebbero essere inclusi nei modelli di *risk adjusment* per aumentare la validità delle stime di associazione; a parità di numerosità della popolazione allo studio, all'aumentare dei confondenti inclusi, diminuisce la precisione delle stime: all'esigenza di *deconfounding* si oppone la necessità di «parsimonia» dei modelli di aggiustamento. Non possono poi essere considerati confondenti fattori che agiscono da modificatori delle misure di effetto. La *detection* di interazioni dovrebbe avvenire a livelli elevati di sensibilità, cioè utilizzando test di eterogeneità delle misure di associazione tra strati con α sufficientemente elevato, certamente più grande di 0,05. Quando si osserva modificazione delle misura di effetto la stima aggiustata, per quel fattore, è distorta, non valida, vanno prodotte stime strato specifiche, la cui precisione è certamente inferiore. Per assurdo, volendo tener conto di tutti i fattori «veri» che modificano l'effetto di un'esposizione, gli unici studi validi avrebbero per titolo «Su un caso di…». Massima valutazione dell'interazione, precisione nulla.

Quando effettuiamo un confronto tra strutture rispetto agli esiti dell'assistenza sanitaria, è allora necessario stabilire se siamo interessati più alla precisione o alla validità delle stime.

Gli autori dell'articolo pubblicato sul fascicolo di maggio-giugno 2006 di *Epidemiologia & Prevenzione*[3] confrontano i risultati del progetto BPAC,[4,5] che aveva come obiettivo una valutazione comparativa delle cardiochirurgie italiane e che utilizzava un approccio inferenziale di tipo classico-fre-

quentista, con quelli ottenuti mediante un approccio di tipo bayesiano.
Il metodo bayesiano permette di definire intervalli di credibilità all'interno dei quali la probabilità attribuita ai diversi valori non è omogenea e, al contrario di quello classico, l'inferenza è ottenuta direttamente sull'intera distribuzione a posteriori della variabile di interesse, invece che su una sola stima puntuale. Inoltre, mentre nell'approccio frequentista, un intervallo di confidenza a livello 1 - α suggerisce che, se potessimo ripetere lo stesso esperimento sotto le medesime condizioni un numero elevato di volte, il vero valore della variabile di interesse cadrebbe al di fuori di questo intervallo per soltanto α volte, l'approccio bayesiano consente di valutare probabilisticamente i parametri di interesse e di prendere in considerazione l'incertezza sottostante, per cui un intervallo di credibilità al 95% è un intervallo che ha il 95% di probabilità di contenere il parametro in esame.
Le tecniche *Markov chain Monte Carlo* (MCMC) consentono di ottenere intervalli di credibilità intorno al *rank* di ciascuna struttura in studio.[6,7] Come già osservato da Marshall e Spiegelhalter, l'uso delle tecniche MCMC per ottenere distribuzioni posteriori per il *rank* degli ospedali porta a una considerevole incertezza nel *rank* degli stessi, anche quando essi mostrino sostanziali differenze nella performance. La conseguenza di assumere a priori la scambiabilità delle strutture è di ridurre le differenze fra le stesse e rendere i loro *rank* ancora più incerti.[8]
Gli autori dell'articolo[3] riferiscono che il metodo da loro utilizzato consente di evidenziare con ragionevole certezza 5 centri con performance peggiore rispetto ai 7 centri individuati nel progetto BPAC, mentre nessun centro può essere identificato come migliore rispetto agli 8 individuati nello stesso studio. Prevedibile: l'approccio bayesiano tiene in maggior conto la precisione delle stime ed è per questo sicuramente più conservativo. Ma è opportuno rendere più incerta la valutazione delle strutture, con la conseguenza di appiattire le differenze di performance degli ospedali rispetto al valore medio di un esito come la mortalità?
Una della maggiori critiche per le analisi bayesiane riguarda l'utilizzo delle distribuzioni a priori. L'appropriata distribuzione a priori dipende dalla parametrizzazione scelta per il modello; tale scelta rimane ampiamente arbitraria.[9] Alcune distribuzioni a priori consentono di ottenere stime meno conservative, con intervalli di credibilità meno ampi e minore *shrinkage* verso la media.[10]
Anche l'uso di altri metodi di analisi statistica, come per esempio i cosiddetti *multilevel*,[11] considerando altre fonti di variabilità, possono ridurre la precisione e aumentare l'incertezza delle misure di effetto.
Negli studi di valutazione comparativa degli esiti, al di là di qualsiasi considerazione sugli effetti della divulgazione dei risultati, la natura dei confronti eseguiti richiede, nella maggior parte dei casi, che si giunga a conclusioni di tipo qualitativo:

- gli ospedali A, B e C hanno performance peggiori della media nazionale?
- i residenti nella regione K sperimentano, a parità di gravità, esiti peggiori rispetto ai residenti nella regione H?

Mettendo da parte cautele legate a considerazioni su possibili problemi di accuratezza dei dati raccolti o sulla possibile esistenza di fattori confondenti non rilevati, e quindi non controllabili, la possibilità di rispondere a queste domande passa per la scelta di quali rischi vogliamo correre nel dichiarare diversi due ospedali che sono veramente uguali e nel dichiarare uguali due ospedali che sono veramente diversi. Questa scelta dipende sostanzialmente dalla valutazione a priori dei costi e dei benefici associati a risultati veri e falsi positivi e negativi dello studio, quindi dipende dalla definizione delle possibili decisioni, individuali e/o collettive, potenzialmente determinate o condizionate dai risultati dello studio. Ancora una volta l'attenzione si deve spostare sugli usi dei risultati, e non limitarsi agli aspetti metodologici, agli strumenti, la cui scelta dipende largamente dalla natura, finalità, obiettivi, di sanità pubblica degli studi.
Nell'articolo citato[3] si classificano gli ospedali con un numero d'ordine che colloca ciascun ospedale in una posizione di una ipotetica graduatoria. Esplicitamente gli autori affermano che «...più alto è il rank, peggiore è il risultato…», a un numero più elevato corrisponde una più elevata mortalità aggiustata, quindi una performance peggiore. Si tratta di una conclusione potenzialmente fortemente distorta e disinformativa. Il metodo indiretto di aggiustamento utilizzato consente esclusivamente di confrontare ciascun livello di esposizione, cioè ciascun ospedale, con la popolazione di riferimento, in questo caso l'intera popolazione allo studio; non consente di confrontare gli ospedali tra loro, a meno di verificare per ciascun confronto l'omogeneità della distribuzione dei confondenti utilizzati per l'aggiustamento. Il progetto BPAC, infatti, non classificava gli ospedali in graduatoria ma esclusivamente li collocava in tre gruppi: uguali alla media, peggiori o migliori della media, senza mai attribuire a ciascun ospedale una posizione ordinale relativa. Sono proprio gli zelanti autori dell'articolo citato che, per primi, producono *league table*, purtroppo una graduatoria sbagliata e disinformativa. In presenza di forti disomogeneità della distribuzione relativa dei confondenti si possono ottenere risultati paradossali. Per esempio un ospedale che con la standardizzazione indiretta risulta peggiore della media, utilizzando il metodo diretto potrebbe essere migliore di un altro ospedale che invece risultava migliore della media con la standardizzazione indiretta.
Gli autori usano dati pubblicati sul sito www.bpac.iss.it, nei quali, come da protocollo dello studio,[4] il confronto è effettuato utilizzando una standardizzazione indiretta e l'intera popolazione allo studio come riferimento. I risultati di un'analisi sugli stessi dati, effettuata con il metodo diretto, cioè utilizzando un modello di aggiustamento a effetti fissi e pren-

dendo come riferimento un pool di ospedali con la mortalità aggiustata più bassa, sono già stati pubblicati.[5] Si tratta di un confronto con significati e usi ben diversi, che propone a ciascun ospedale di confrontarsi con la miglior possibile performance osservata e consente confronti diretti tra ospedali. In questo caso un *ranking* è possibile. Confronti di esiti tra ospedali effettuati con questo metodo, sono già stati pubblicati[12,13] e sono certamente noti agli autori.
Di tutti questi argomenti gli autori, e verosimilmente i revisori di *E&P*, avrebbero potuto meglio tener conto se avessero avuto la bontà di leggere, e la correttezza di citare, gli articoli scientifici già pubblicati su questo studio e la relativa corrispondenza.[14-17]
L'analisi dello studio BPAC è stata condotta secondo metodi esplicitamente definiti in un rigoroso protocollo (pubblicato),[4,5] definito da un comitato scientifico, di cui uno degli autori (RG) dell'articolo faceva parte. Non risulta dai verbali che RG abbia mai proposto un metodo di analisi come quello utilizzato nell'articolo di *E&P*.
Siamo consapevoli che nessun disegno di studio, tantomeno diversi metodi statistici, hanno la possibilità di misurare una realtà come è veramente, ma possono produrre immagini che possiamo considerare vere in relazione a specifici obiettivi e usi, definendo esplicitamente, in termini di validità e di precisione, limiti e incertezze. Su questi presupposti, politici, manager, epidemiologi, clinici, dovrebbero imparare a decidere, per gli obiettivi di tutela della salute delle persone e della popolazione, misurandosi con l'incertezza, ma scegliendo sulla base delle migliori conoscenze scientifiche disponibili.[1]
Condividiamo le preoccupazione degli autori circa le modalità di comunicazione dei risultati di studi di valutazione comparativa degli esiti; siamo convinti che il Servizio sanitario nazionale, nell'avviarsi a un'attività sistematica di valutazione comparativa di esiti tra popolazioni e soggetti erogatori, saprà sviluppare rigorosi metodi di studio e programmi di comunicazione efficaci e appropriati.
Rigorose e giustificate preoccupazioni sulla comunicazione, avrebbero dovuto indurre gli autori a maggiore accuratezza e precisione nella presentazione dei loro risultati. La figura 2 a pag. 202 di *E&P* inverte sistematicamente l'ordine degli ospedali per mortalità aggiustata a 30 giorni dopo BPAC, attribuendo erroneamente il *rank* 1 all'ospedale con la più elevata mortalità e il 64 a quello con la performance migliore: un deprecabile, ma evitabile, errore sistematico, certamente attribuibile al caso burlone, cui nemmeno il metodo bayesiano può porre rimedio.

Bibliografia

1. Arcà M, Fusco D, Barone AP, Perucci CA. Introduzione ai metodi di risk adjustment nella valutazione comparativa dell'outcome. *Epidemiol Prev* 2006; 30 (4-5) suppl (1): 1-48.
2. Rothman KJ. *Epidemiology. An introduction*. New York, Oxford University Press, 2002.
3. Di Tanna GL, Cisbani L, Grilli R. Segnalare ai cittadini la qualità degli ospedali: il rischio della cattiva informazione. *Epidemiol Prev* 2006; 30(3): 199-204.
4. Seccareccia F, Capriani P, Diemoz S, Taioli E, Tosti ME, Greco D; Gruppo di Ricerca Italiano Progetto BPAC. Cross-sectional study of cardiac surgery centers within the «CABG Project» (short-term outcome in patients undergoing coronary artery bypass graft surgery in Italian cardiac surgery centers). *Ital Heart J Suppl* 2003; 4(1): 32-38. Italian.
5. Seccareccia F, D'Errigo P, Rosato S, Tosti ME, Manno V, Badoni G, Greco D e il Gruppo di Ricerca Progetto BPAC. Studio degli esiti a breve termine degli interventi di By-Pass Aorto-Coronarici (BPAC) nelle cardiochirurgie italiane. *ISTISAN* 2005; 05/33: 1-57.
6. Gelman A, Rubin DB. Markov chain Monte Carlo methods in biostatistics. *Stat Methods Med Res* 1996; 5: 339-55.
7. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov chain Monte Carlo methods in practice*. New York, Chapman and Hall, 1996.
8. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technol Assess* 2000; 4(38): 1-130.
9. Clayton D, Hills M. *Statistical models in epidemiology*. New York, Oxford University Press, 1993.
10. Howley PP, Gibberd R. Using hierarchical models to analyse clinical indicators: a comparison of the gamma-Poisson and beta-binomial models. *Int J Qual Health Care* 2003; 15(4): 319-29.
11. Brown H, Prescott R. *Applied Mixed Models in Medicine*. Chichester, John Wiley and Son, 2006.
12. Agabiti N, Ancona C, Forastiere F, Arca M, Perucci CA. Evaluating outcomes of hospital care following coronary artery bypass surgeryin Rome, Italy. *Eur J Cardiothorac Surg* 2003; 23(4): 599-606; discussion 607-8.
13. Fantini MP, Stivanello E, Frammartino B et al. Risk adjustment for inter-hospital comparison of primary cesarean section rates: need, validity and parsimony. *BMC Health Serv Res* 2006; 6: 100.
14. Seccareccia F, Perucci CA, D'Errigo P et al.; research group of the Italian CABG Outcome Study. The Italian CABG Outcome Study: short-term outcomes in patients with coronary artery bypass graft surgery. *Eur J Cardiothorac Surg* 2006; 29(1): 56-62; discussion 62-64.
15. Seccareccia F, Perucci CA, Fusco D, D'Errigo P. Reply to Biondi-Zoccai et al. *Eur J Cardiothorac Surg* 2006; 29: 856.
16. Seccareccia F, Perucci CA, Fusco D, D'Errigo P. Reply to Hekmat et al. *Eur J Cardiothorac Surg* 2006; 29: 857-58.
17. Seccareccia F, Perucci CA, D'Errigo P, Fusco D. Concerning the Editorial comment by Dr Menicanti. *Eur J Cardiothorac Surg* 2006; 29(5): 858-59; author reply 859-60.

Risposta degli autori
The authors' reply

Il nostro articolo non aveva lo scopo di criticare la metodologia utilizzata nel contesto dello studio sugli esiti del bypass aorto-coronarico nelle cardiochirurgie italiane coordinato dall'Istituto superiore di sanità.
Semplicemente affrontava il modo in cui i mass media riportano informazioni comparative sulla qualità dei servizi sanitari. A questo scopo venivano considerati due «casi modello», uno dei quali era rappresentato dal modo in cui i risultati di quello studio sono stati utilizzati dai mezzi di comunicazione di massa, derivandone, come era prevedibile, classifiche di «buoni e cattivi» centri cardiochirurgici.
Sotto questo profilo, l'estesa parte introduttiva dei commenti di Fusco et al. con la descrizione delle metodolo-

gie statistiche di risk adjustment e di comparazione (diretta o indiretta) dei centri è un sempre utile ripasso dei fondamentali, ma non ci pare molto pertinente. Lo stesso può dirsi della rivendicazione della validità delle analisi condotte in quello studio, validità di cui semplicemente non abbiamo discusso in alcun modo, dandola per scontata al punto da aver usato i risultati dello studio ISS come parametro di riferimento per mettere in discussione (nel secondo «caso modello» considerato) la rappresentazione della qualità dei servizi fatta attraverso un indice di reputazione da parte di una rivista non scientifica. In questo contesto quindi, anche il riferimento al fatto che uno di noi (RG) ha fatto parte del Comitato scientifico dello studio ISS sul bypass è assolutamente irrilevante.
Anche se hanno perso di vista la «foresta» del senso del nostro articolo (ma certamente possiamo essere stati poco chiari noi…), Fusco et al. hanno invece (e molto opportunamente) individuato l'«albero» di un refuso di stampa in una delle figure, colpevolmente sfuggitoci in fase di revisione di bozze.
Siamo loro grati per aver portato alla attenzione nostra e dei lettori di E&P questo errore, deprecabile come loro stessi dicono, ma non infrequente nella letteratura scientifica nazionale e internazionale e comunque non sufficiente a giustificare il tono inutilmente sgradevole dei loro commenti.

Gian Luca Di Tanna,
Luca Cisbani, Roberto Grilli
Agenzia sanitaria regionale
Emilia-Romagna

l'opinione

Il rischio di nessuna informazione
The risks of no information

Roberto Grilli e collaboratori dichiarano di voler affrontare il modo in cui i mass media riportano informazioni comparative sugli esisti degli interventi sanitari, ma nel loro articolo non lo descrivono.
Esaminando i quattro esempi da loro citati in bibliografia a proposito dello studio BPAC è facile rilevare che i primi due (*Il Messaggero* e *Il Sole24Ore*) si sono limitati a pubblicare un estratto del comunicato stampa dell'Istituto superiore di sanità (ISS) il giorno successivo alla sua diffusione. L'unica elaborazione giornalistica è quella comparsa sul *Corriere della Sera* il giorno dopo e sul *Corriere Salute* la settimana successiva, con la firma di Roberto Satolli, cioè di chi scrive queste righe. In altre parole, pur essendo la comunicazione dell'ISS una novità assoluta per il nostro Paese, non è stata percepita come notizia, né dai telegiornali né dai quotidiani nazionali né dai settimanali d'attualità. Il caso del *Corriere* infatti è la rondine che non fa primavera: vedere il sottoscritto come unico rappresentante dei media è un po' troppo *bias* in questo caso, come è ben noto ai lettori di E&P, di cui sono collaboratore e sono stato per anni anche editore. E come è noto anche a Roberto Grilli, con cui da quasi dieci anni discuto in ogni sede sull'opportunità di divulgare i dati di esito.
Quindi la vera conclusione su cui riflettere è che la modalità prevalente con cui i media reagiscono alla disponibilità di questi dati è il silenzio e l'indifferenza. Tanto che lo stesso *Panorama* pochi mesi dopo non ha neppure pensato di incrociare i propri risultati di reputazione delle cardiochirurgie con quelli di mortalità dell'ISS, per verificarne l'attendibilità.
Cosa che invece hanno fatto molto opportunamente Roberto Grilli e collaboratori, con ciò assumendo esplicitamente i risultati dell'ISS come *gold standard* di riferimento.
Se i dati di mortalità sono dunque di buona qualità, in che modo la rappresentazione che ne ha dato l'unico media che se ne è occupato è distorta?
Sembra di capire che l'errore sia stato quello di identificare semplicisticamente alcuni centri come «migliori» e altri come «peggiori» della media, mentre le analisi «più sofisticate» condotte dagli autori dimostrerebbero che ciò non è possibile.
Poiché è evidente l'incongruenza di aspettarsi dai media analisi «più sofisticate» di quelle condotte dall'ISS, è probabile che in questo caso si stia parlando a nuora perché suocera intenda. Ma in ogni modo la tesi appare poco convincente. Gli autori infatti, nello sforzo di trovare un ragionevole compromesso tra validità scientifica e comprensibilità, suggeriscono come possibile soluzione i *funnel plot*. Nel grafico da loro elaborato, purtroppo, sono ancora facilmente identificabili otto centri migliori con una mortalità aggiustata al di sotto dell'imbuto e sette peggiori al di sopra.
Insomma, per quanto sofisticate siano le analisi e poco comprensibili le rappresentazioni, è difficile nascondere il fatto che i risultati dell'ISS individuano un piccolo gruppo di ospedali con mortalità sotto la media e un altro piccolo gruppo sopra la media. Questi ultimi in particolare meriterebbero non solo un approfondimento, ma soprattutto un intervento urgente da parte delle istituzioni che ne hanno responsabilità.

Roberto Satolli
Zadig, Milano