1 • NUOVI METODI PER PRODURRE RACCOMANDAZIONI CLINICHE E PER SINTETIZZARE LE PROVE DI EFFICACIA

Silvia Pregno, Alessandro Liberati

INTRODUZIONE

Nell'analisi dei nuovi approcci alla valutazione delle innovazioni in sanità non si può prescindere dal considerare i nuovi strumenti di cui la comunità scientifica ed epidemiologica si è dotata negli ultimi anni per la valutazione dell'appropriatezza e delle nuove tecnologie. Strettamente legato a questo vi è il problema dell'aggiornamento degli operatori sanitari e dell'informazione ai pazienti, connessi con un'effettiva e reale valutazione delle innovazioni e delle ricerche sottostanti (Evans 2007).

Già nel 1995 Mulrow ricordava che ogni anno nella letteratura biomedica venivano pubblicati più di due milioni di articoli in più di 20.000 riviste, letteralmente una montagna di informazioni (Mulrow 1995). Per trovare una soluzione al problema, sono state proposte diverse soluzioni (riviste di pubblicazioni secondarie, revisioni sistematiche, rapporti di health technology assessment, linee-guida e ancora più recentemente i cosiddetti *point of care* product).

Tutto questo ha reso sempre più cogente il problema della "validità e attendibilità" delle sintesi di letteratura in quanto i rischi di riduzionismo e ipersemplificazione da un lato, e di franca "manipolazione interessata" dall'altro, devono essere il più possibile controllati.

Ecco perché negli ultimi anni si è sviluppato un filone di ricerca volto ad approfondire il rapporto tra la necessaria, ma non sufficiente, valutazione del rigore metodologico con cui le informazioni sono ottenute (criteri di inclusione ed esclusione delle fonti primarie, adeguata sintesi e interpretazione dei risultati, ecc.) e della loro effettiva

rilevanza clinico-epidemiologica. Sino a non molto tempo fa, la valutazione avveniva soprattutto mettendo sotto analisi la struttura gerarchica delle prove, basando il giudizio essenzialmente sul tipo di architettura di ricerca (leggi disegno di studio, sperimentale o meno) attraverso il quale erano state ottenute le informazioni.

L'applicazione rigida, e talvolta un po' "acritica" dei tradizionali livelli di evidenza (LdE) è stata così messa progressivamente in discussione, suggerendo la necessità di un approccio più articolato, in cui anche le indicazioni provenienti da studi qualitativi possano essere considerate e incorporate con quelle provenienti da dati quantitativi (Glasziou 2004).

Questa spinta verso una maggiore flessibilità metodologica ha trovato la sua giustificazione nel riconoscimento che non esiste il disegno di studio in assoluto "perfetto e ideale", indipendentemente dal tipo di quesito.

Ciascun quesito di studio (e ricerca) deve essere affrontato con la metodologia e il disegno di studio più adeguato. Infatti, se per molti quesiti terapeutici gli studi controllati randomizzati (RCT) sono la migliore opzione metodologica, per quesiti di tipo non terapeutico (ad esempio, nella valutazione della accuratezza di un test diagnostico o del valore prognostico di un classificazione clinica o di un segno/sintomo) il disegno RCT non solo non è utile, ma la sua applicazione è, di fatto, inappropriata.

A questo si è aggiunta la consapevolezza che, anche in medicina, la conoscenza deve essere concepita come un processo cumulativo e che quindi per la sua valutazione sono necessari, oltre a strumenti di valutazione "puntuale" (ovverosia del singolo studio), anche l'analisi complessiva della sua coerenza, consistenza e rilevanza. Solo questa visione dà infatti sostanza al principio ipotetico-deduttivo enunciato da Karl Popper secondo cui la validità di una teoria è misurata dalla sua dimostrabile capacità di resistere ai tentativi di falsificazione (Popper 1970).

A partire da queste schematiche considerazioni di background sono due i temi che affronteremo in questo capitolo:

- La nuova concettualizzazione proposta dal metodo GRADE per la valutazione della qualità delle prove necessaria nel percorso di formulazione di raccomandazioni di comportamento su temi di rilevanza clinico-epidemiologica.

- Gli approcci necessari al miglioramento della trasparenza nel percorso di valutazione della qualità delle prove che è stata adottata dalla Cochrane Collaboration (CC) come approccio standard alla esplicitazione del percorso di conduzione di una revisione sistematica.

IL METODO GRADE

I diversi schemi di *grading* proposti nel corso degli anni sono derivati per successive modificazioni e miglioramenti da quello messo a punto (sin dal 1979) dalla *Canadian Task Force on the Periodic Health Examination* e successivamente anche dalla *United States Preventive Task Force*.

La maggior parte dei vecchi sistemi di classificazione dei LdE – anche perché sviluppati soprattutto per valutare quesiti relativi all'efficacia degli interventi – considerava come livello più alto quello derivato dalle informazioni provenienti da RCT e come livello più basso le prove basate sull'opinione di esperti in assenza di dati empirici. Esistevano metodi diversi per comunicare il grading delle evidenze e delle raccomandazioni, con messaggi diversi sullo stesso argomento. Ad esempio, la somministrazione degli anticoagulanti orali in pazienti con fibrillazione atriale e valvulopatia mitralica di origine reumatica ha ricevuto diversi livelli di raccomandazioni a seconda dell'organizzazione che ha svolto il lavoro: una raccomandazione di classe I basata su un livello di evidenza B dall'American Heart Association (Fuster 2001); di grado C basata su un livello di evidenza IV dal SIGN (SIGN 2004) e un grado 1C+ (dove il numero indica il rapporto fra i benefici e i rischi e il C+ la qualità metodologica sottostante l'evidenza) dall'American College of Chest Physicians (Albers 2001). Già Sackett nel 1997 metteva in luce l'importanza del quesito clinico rispetto all'enfasi posta sulla metodologia di studio utilizzata, sostenendo che era la domanda clinica a determinare la strategia metodologica appropriata (Sackett 1997).

In passato i clinici e le organizzazioni che elaborano raccomandazioni hanno commesso errori determinati dall'incapacità di tenere in conto la qualità delle prove disponibili (Lacchetti 2002). Ad esempio, per oltre un decennio, si è raccomandato ai clinici di incoraggiare le donne in postmenopausa all'uso della terapia ormonale sostitutiva (ACP 1992). Molti medici hanno applicato questa raccomandazione, pensando di ridurre il rischio cardiovascolare nelle loro pazienti. Se si fosse applicato sin da subito un metodo rigoroso per la valutazione

della qualità delle prove, si sarebbe visto che, poiché i risultati derivavano da studi osservazionali inconsistenti fra loro, le prove legate alla riduzione del rischio cardiovascolare sarebbero state giudicate di qualità molto bassa (Humphrey 2002).

Il GRADE Working Group – Grading of Recommendations Assessment, Development and Evaluation – è nato nel 2000 come una collaborazione informale di ricercatori con un interesse a migliorare i sistemi di grading applicati nell'assistenza sanitaria (GRADE Working Group 2004).

Il metodo GRADE può essere utilizzato per produrre raccomandazioni clinico-organizzative (se si utilizzano i due step della valutazione della qualità delle prove e della forza delle raccomandazioni), ma può anche essere utile per la sola valutazione di qualità all'interno di revisioni sistematiche e di rapporti di health technology assessment (Guyatt 2008a).

Il GRADE è un sistema strutturato e trasparente studiato per analizzare e presentare la sintesi della qualità delle prove. Nel corso del suo sviluppo è stato studiato in un ampio ventaglio di quesiti clinici di diversa natura: diagnostici, terapeutici, prognostici (Schünemann 2008). Ora il metodo è usato ampiamente a livello internazionale (www.gradeworkinggroup.org/society) da più di 30 organizzazioni, pur con adattamenti e differenze locali, quali l'Organizzazione Mondiale della Sanità, l'American College of Physicians, l'American Thoracic Society, UpToDate, la Cochrane Collaboration e, in Italia, dal'Agenzia Sociale e Sanitaria della Regione Emilia-Romagna che ha recentemente pubblicato un Dossier sull'utilizzo del metodo per la produzione di raccomandazioni per l'uso di farmaci oncologici (De Palma 2009). La sua diffusione così ampia riflette il rigore metodologico associato alla facilità d'uso.

Il GRADE mette a disposizione di chi sviluppa raccomandazioni o intere linee-guida un sistema strutturato per affrontare e sviluppare tutti i passaggi necessari. Esso infatti richiede che si parta da un approccio strutturato alla formulazione dei quesiti, alla scelta degli esiti (outcome) di interesse, alla graduazione della loro importanza, alla valutazione delle prove e alla loro integrazione con i valori e le preferenze dei pazienti e della società, per giungere alla formulazione delle raccomandazioni. Il metodo GRADE offre quindi numerosi vantaggi che in parte sono condivisi da altri metodi di valutazione, ma che nessuno combina insieme (Atkins 2004).

Il primo passaggio del processo di produzione di una raccomandazione presuppone la formulazione di una domanda chiara e definita basata su quattro componenti (figura 1.1): i pazienti, l'intervento, il confronto e l'esito di interesse (Patients, Intervention, Comparison, Outcomes - PICO) (Oxman 1988). Il GRADE richiede che siano specificati tutti gli esiti che sono importanti per i pazienti fin dall'inizio del processo di formulazione di una raccomandazione e invita anche a differenziare gli esiti critici da quelli importanti. A tal fine viene previsto un percorso strutturato e il più possibile esplicito che passa attraverso l'identificazione – da parte del gruppo di lavoro che deve produrre la raccomandazione – di tutti gli outcome potenzialmente rilevanti e la loro individuale e collegiale valutazione. L'approccio suggerito dagli ideatori del metodo GRADE è quello di sottoporre a valutazione ogni outcome facendo votare ogni partecipante al gruppo di lavoro con una scala compresa tra 1 e 9. La parte alta della scala, da 9 a 7, identifica gli esiti critici nel prendere una decisione clinica; i giudizi da 4 a 6 rappresentano gli esiti che sono importanti ma non critici; quelli da 1 a 3 quelli di limitata o scarsa importanza ai fini decisionali (Guyatt 2008a).

1. Definizione del quesito (Patient, Intervention, Comparison, Outcome)

2. Definizione dell'importanza relativa degli esiti

3. Ricerca delle prove di efficacia

4. Valutazione della qualità delle prove per ciascun esito

I fattori che possono *abbassare* la qualità delle prove:	I fattori che possono *aumentare* la qualità delle prove sono:
• le limitazioni degli studi	• la grande dimensione dell'effetto
• l'inconsistenza dei risultati	• l'assenza di fattori di confondimento che plausibilmente potrebbero ridurre un effetto dimostrato
• la scarsa trasferibilità applicabilità delle prove	
• l'imprecisione della stima degli effetti	• la presenza di un gradiente dose risposta
• bias di pubblicazione	

5. Riassunto delle prove per ciascun esito critico o importante

6. Valutazione della qualità globale delle prove

7. Bilanciamento dei benefici e degli eventi avversi

8. Bilanciamento fra i benefici e i costi

9. Definizione della raccomandazione e della sua forza

Figura 1.1 • Come giungere a una raccomandazione con il metodo GRADE (adattato da GRADE 2004)

La valutazione della qualità delle prove con il metodo GRADE

Il metodo GRADE propone una valutazione della *qualità delle prove* più ampia e articolata di quella proposta da tutti gli altri metodi oggi disponibili (Guyatt 2008b). La principale e maggiore novità introdotta dal metodo consiste nel richiedere uno spostamento da una valutazione *studio-specifica* a una *outcome-specifica*. La motivazione di questo spostamento di attenzione deriva dalla consapevolezza che non basta valutare l'appropriatezza del disegno di studio per determinare la qualità delle prove: ma deve essere presente anche una appropriatezza del disegno di studio rispetto all'*outcome* valutato. Ad esempio *outcome hard* non richiedono l'impiego del doppio cieco mentre *outcome soft*, sì. *Outcome hard* non richiedono il ricorso a forme di cecità parziale del disegno, mentre *outcome soft* sì, e così via.

Quindi, la prima vera innovazione del metodo GRADE per quanto riguarda la qualità delle prove è il passaggio da un approccio study level a uno outcome level. Essa riflette il "grado di fiducia" che si ripone nel fatto che la stima di effetto ottenuta dall'insieme degli studi disponibili sia adeguata a supportare una raccomandazione.

Dal punto di vista operativo, per valutare la qualità delle prove, il metodo GRADE prevede che si inizi dal disegno dello studio (Guyatt 2008b). In tabella 1.1 è descritto il percorso di valutazione delle prove per ogni esito considerato.

Per le raccomandazioni che si occupano di strategie alternative di gestione clinica – a differenza di quelle che si occupano di prognosi o accuratezza dei test diagnostici – gli RCT forniscono, in generale, prove più forti rispetto agli studi osservazionali. Studi osservazionali rigorosamente condotti forniscono a loro volta una prova più forte rispetto alle serie di casi non controllate. Gli RCT senza importanti limitazioni costituiscono quindi una prova di alta qualità, mentre studi osservazionali senza specifici punti di forza o limitazioni importanti costituiscono una prova di bassa qualità.

Il metodo GRADE richiede una valutazione separata della qualità delle prove per ciascuno degli esiti importanti per i pazienti e identifica cinque fattori che possono abbassare la qualità delle prove, applicabili sia agli studi randomizzati che a quelli osservazionali e tre fattori che possono alzarla (figura 1.1).

TABELLA 1.1 • ESEMPIO DI UN GRADE EVIDENCE PROFILE

GRADE evidence profile for impact of surgical alternatives for pancreatic cancer from systematic review and meta-analysis of randomised controlled trials in inpatient hospitals of pylorus preserving versus standard Whipple pancreaticoduodenectomy for pancreatic or periampullary cancer

No of studies (No of participants)	Quality assessment					Summary of findings			
	Study limitations	Consistency	Directness	Precision	Publication bias	Relative effect (95% CI)	Best estimate of Whipple group risk	Absolute effect (95% CI)	Quality
Five year mortality:									
3 (229)	Serious limitations (–1)	No important inconsistency	Direct	No important imprecision	Unlikely	0.98 (0.87 to 1.11)	82.5%	20 less/1000; 120 less to 80 more	+++, moderate
In-hospital mortality:									
6 (490)	Serious limitations (–1)	No important inconsistency	Direct	Imprecision (–1)	Unlikely	0.40 (0.14 to 1.13)	4.9%	20 less/1000; (50 less to 10 more)	++, low
Blood transfusions (units):									
5 (320)	Serious limitations (–1)	No important inconsistency	Direct	No important imprecision	Unlikely	—	2.45 units	–0.66 (–1.06 to –0.25); favours pylorus preservation	+++, moderate
Biliary leaks:									
3 (268)	Serious limitations (–1)	No important inconsistency	Direct	Imprecision (–1)	Unlikely	4.77 (0.23 to 97.96)	0	20 more/1000 20 less to 50 more	++, low
Hospital stay (days):									
5 (446)	Serious limitations (–1)	No important inconsistency	Direct	Imprecision (–1)	Unlikely	—	19.17 days	–1.45 (–3.28 to 0.38); favours pylorus preservation	++, low
Delayed gastric emptying:									
5 (442)	Serious limitations (–1)	Unexplained heterogeneity (–1)	Direct	Imprecision (–1)	Unlikely	1.52 (0.74 to 3.14)	25.5%	110 more/1000; 80 less to 290 more	+, very low

Adattato da Guyatt, 2008b

Limitazioni degli studi

La valutazione della qualità delle prove si abbassa se gli studi hanno importanti limitazioni che possono inficiare le stime dell'effetto del trattamento (Guyatt 2008b). Queste limitazioni includono:

- la mancanza del mascheramento della sequenza randomizzata di allocazione al trattamento (allocation concealment);
- mancanza o carenza delle misure adottate per assicurare la "cecità" dei partecipanti allo studio e del personale sanitario e non sanitario rispetto al trattamento assegnato, in particolare se gli esiti sono soggettivi e la loro valutazione è esposta a rischi di distorsione (blinding of participant, personnel and outcome assessor);
- ampie perdite di pazienti al follow-up con incompletezza dei dati disponibili per ciascun esito considerato e/o mancata adesione al principio dell'intention to treat analysis (incomplete outcome data);
- l'interruzione precoce di uno studio a causa degli effetti benefici dell'intervento sperimentale considerato;
- la mancata descrizione di alcuni esiti nei risultati degli studi, tipicamente per quelli in cui non sia stato osservato un effetto dell'intervento sperimentato.

Inconsistenza fra i risultati degli studi esaminati per ciascun esito (inconsistency)

La presenza di ampie differenze nelle stime dell'effetto di un intervento fra gli studi considerati suggerisce la presenza di differenze reali nel sottostante effetto del trattamento. La variabilità può derivare da differenze nelle popolazioni in esame (ad esempio i farmaci possono avere un effetto più ampio in pazienti più gravi), da differenze fra gli interventi considerati (ad esempio effetti più grandi ad alte dosi) o tra gli esiti misurati (ad esempio la diminuzione dell'effetto con il passare del tempo).

Quando l'eterogeneità esiste, ma non si riesce a trovare una spiegazione plausibile a supporto di essa, la qualità delle prove deve essere abbassata.

Tale diminuzione può essere di uno o due livelli, a seconda di quanto si giudichi importante la presenza della stessa sulla fiducia nella stima dell'effetto. Anche in questo caso la motivazione del giudizio sottostante alla riduzione del livello deve essere chiaramente esplicitato in una nota.

La trasferibilità/applicabilità delle prove (indirectness)

Chi produce linee-guida si trova di fronte a due tipi di *indirectness* delle prove. La prima si realizza ad esempio nel caso di confronti indiretti fra due farmaci, utilizzati per la medesima patologia. Benché possano essere disponibili RCT che confrontino direttamente un farmaco A con il placebo e altri che confrontino un farmaco B sempre con il placebo, possono non esserci dati che permettono un confronto tra i due farmaci direttamente. Quando abbiamo un confronto indiretto, la qualità delle prove è considerata più bassa, rispetto a un confronto testa a testa. Il secondo tipo di *indirectness* delle prove include invece differenze fra le popolazioni, interventi, confronti ed esiti di interesse, che rendono poco o per nulla trasferibili le informazioni derivanti dalle prove disponibili alle popolazioni, interventi, confronti o esiti di interesse per la linea-guida (Guyatt 2008b).

Imprecisione nella stima degli effetti

Quando gli studi includono pochi pazienti e pochi eventi e quindi hanno ampi intervalli di confidenza, un panel di una linea-guida può giudicare la qualità delle prove più bassa, proprio perché viene introdotto un livello più o meno elevato di imprecisione nella stima dell'effetto che rende incerti della stessa e del suo reale significato. Anche in questo caso, così come nel caso della inconsistenza e *dell'indirectness*, il livello può essere diminuito di uno o due livelli e la motivazione deve essere sempre chiaramente esplicitata (Guyatt 2008b).

Bias *di pubblicazione*

La qualità delle prove sarà poi ridotta se i ricercatori hanno selettivamente incluso solo gli studi che riportano per gli esiti considerati risultati positivi. Una situazione che tipicamente dovrebbe indurre al dubbio della presenza di questo rischio di distorsione è quella che si verifica quando per un esito gli studi pubblicati sono pochi e tutti finanziati dalle industrie produttrici.

I fattori che possono aumentare la qualità delle prove

Sebbene gli studi osservazionali ben condotti danno comunque luogo a qualità delle prove di livello basso rispetto agli RCT, in alcune circostanze possono produrre prove di qualità moderata o anche alta (Guyatt 2008b). Ciò si verifica quando studi osservazionali ben condotti da un punto di vista metodologico portano ad ampie e consistenti

stime nella dimensione dell'effetto di un trattamento: in questo caso possiamo essere fiduciosi nei loro risultati. Quanto più è ampia la dimensione dell'effetto, tanto più "forte" diventa la prova di efficacia. Ad esempio, una metanalisi di studi osservazionali mostrava che i caschi per i ciclisti riducevano il rischio di traumi cranici in caso di incidenti (OR 0,31 – CI 0,26-0,37) (Thompson 2000). Questo effetto così ampio suggerisce una valutazione di qualità moderata. Anche la presenza di un gradiente dose-risposta induce ad aumentare la qualità delle prove.

Una volta definita la qualità delle prove per ogni singolo esito i membri del panel – dopo la discussione collegiale – dovranno esprimere un giudizio sulla qualità totale delle prove. Gli autori del GRADE propongono che la qualità dell'evidenza più bassa per tutti gli esiti critici fornisca la base per valutare la qualità della prova generale. Inoltre, gli esiti che sono importanti, ma non critici nella decisione clinica, dovrebbero essere inclusi nei profili delle evidenze e dovrebbero essere considerati quando si fornisce un giudizio fra i benefici e i rischi, ma non dovrebbero essere tenuti in considerazione quando si valuta la qualità della prove generale. La qualità delle prove è un continuum e ogni categorizzazione include un certo grado di arbitrarietà. Tuttavia i vantaggi derivanti dalla semplicità, trasparenza e chiarezza bilanciano questi limiti. In sintesi, considerando e valutando tutti i fattori precedentemente descritti, il sistema GRADE classifica la qualità delle prove in quattro livelli: alto, moderato, basso e molto basso (box1.1).

BOX 1.1 • LA CLASSIFICAZIONE DELLA QUALITÀ DELLE PROVE

Qualità alta	Ulteriori ricerche molto probabilmente non cambieranno la fiducia nella stima dell'effetto.
Qualità moderata	Ulteriori ricerche probabilmente avranno un impatto importante nella fiducia della stima dell'effetto e potranno modificare le stime.
Qualità bassa	Ulteriori ricerche quasi certamente avranno un impatto nella fiducia della stima dell'effetto e molto probabilmente modificheranno le stime.
Qualità molto bassa	Ogni stima dell'effetto è molto incerta.

Come si attribuisce la forza di una raccomandazione

La "forza" di una raccomandazione (e ancor più specificamente la sua graduazione) deve servire ad esprimere (e comunicare) quanto si sia convinti del fatto che "applicare la raccomandazione produrrà più beneficio che danno".

Il GRADE propone di classificare le raccomandazioni in *forti e deboli*, a favore o contro l'uso di un determinato trattamento/intervento. Una raccomandazione *forte* dovrebbe essere attribuita a quelle situazioni nelle quali si ritiene che la maggior parte dei pazienti – una volta adeguatamente informati – sceglierebbe di seguire la raccomandazione (Guyatt 2008c). Una raccomandazione *debole* dovrebbe invece essere usata in tutte quelle situazioni nelle quali si ritiene che le scelte dei pazienti varieranno sulla base dei propri valori e preferenze e che i clinici debbano creare le migliori condizioni per garantire che i pazienti possano scegliere sulla base dei loro valori e preferenze.

La forza della raccomandazione è determinata dal bilanciamento di quattro fattori chiave:

- *Il bilanciamento fra gli effetti desiderati e non desiderati*: tanto più ampia è la differenza fra gli effetti desiderabili e non desiderabili, tanto più è probabile che una raccomandazione sarà forte. Quanto più incerto sarà tale bilanciamento, tanto più alta sarà la probabilità che una raccomandazione sia debole.
- *La qualità delle prove*: tanto più è alta la qualità delle prove, tanto più alta sarà la probabilità di avere una raccomandazione forte.
- *I valori e le preferenze*: tanto più i valori e le preferenze variano o tanto più ampia è l'incertezza in essi, tanto più elevata sarà la probabilità che la raccomandazione sia debole.
- Tanto più alto è il *costo* di un intervento o maggiori le risorse consumate, tanto più sarà bassa la probabilità di avere una raccomandazione forte.

Con questa classificazione il metodo GRADE si è così sforzato di graduare la forza delle raccomandazioni in modo tale da dare strumenti interpretativi e decisionali per pazienti, clinici e decisori sanitari. Nella figura 1.2 sono riassunte le possibili implicazioni per i pazienti, i clinici e i decisori sanitari delle raccomandazioni forti e deboli.

	Le implicazioni di una raccomandazione forte	Le implicazioni di una raccomandazione debole
per i pazienti:	la maggior parte delle persone con quel problema vorrebbero avere l'intervento raccomandato e solo una piccola parte non lo vorrebbe; questo richiede una discussione se l'intervento non è offerto	la maggior parte delle persone con quel problema vorrebbero avere l'intervento, ma molti non lo vorrebbero
per i clinici:	la maggior parte dei pazienti dovrebbe ricevere l'intervento raccomandato	diverse scelte possono essere appropriate per pazienti diversi e bisogna aiutare ogni paziente ad arrivare a una decisione consistente con i propri valori e preferenze
per i decisori sanitari:	la raccomandazione può essere adottata nella maggior parte delle situazioni come una indicazione	il processo decisionale necessita di un importante dibattito e il coinvolgimento di molti stakeholder

Figura 1.2 • Implicazioni di una raccomandazione in funzione della sua forza (adattato da Guyatt 2008c)

I NUOVI STRUMENTI PER UNA CONDUZIONE E PRESENTAZIONE PIÙ TRASPARENTE DEI RISULTATI DELLE REVISIONI SISTEMATICHE

In campo medico, le prime revisioni sistematiche (RS) sono state pubblicate a partire dalla fine degli anni Settanta (Altman 1995). Una RS è un tentativo di rispondere a uno specifico quesito di ricerca attraverso la sintesi, fatta seguendo un percorso metodologico esplicito, di tutte le prove empiriche che rispondono a criteri di eligibilità prespecificati. Ed è proprio l'uso di un percorso metodologico esplicito, volto a minimizzare le possibili distorsioni (bias), a permettere di giungere a risultati affidabili, sulla cui base trarre conclusioni e prendere decisioni.

Le analisi che vengono condotte nell'ambito di una RS possono essere sia narrative che quantitative e un'articolazione generale delle domande che possono trovare soluzione all'interno di una RS è indicata dalle seguenti cinque domande:

- Qual è la *qualità degli studi* che sono disponibili per valutare l'efficacia di un determinato intervento?
- Quale è la *direzione* dell'effetto dell'intervento all'interno degli studi disponibili?
- Qual è la *dimensione* dell'effetto (di quanto migliora o peggiora gli esiti)?
- L'effetto è *omogeneo e coerente* fra gli studi?
- Qual è la *forza delle prove* relative all'effetto misurato?

Nell'ambito di una RS è talvolta possibile eseguire delle sintesi quantitative combinando, con appropriate tecniche statistiche, i risultati di studi diversi. Queste tecniche vanno sotto il nome di metanalisi e sono più o meno appropriate all'interno di una RS a seconda della natura e caratteristiche degli studi primari. Rispondere alla quinta domanda richiede ulteriori giudizi fondati sulla valutazione del disegno di studio del risk of bias, così come della misura statistica dell'incertezza.

La Cochrane Collaboration (CC) è la principale organizzazione che, a livello internazionale, produce RS della letteratura scientifica al fine di aiutare le persone a prendere decisioni informate in campo sanitario. Proprio alla luce di questa sua mission principale la CC è costantemente impegnata a contribuire al miglioramento della metodologia delle RS.

In quest'ottica la CC si è dotata di due nuovi strumenti in grado di rendere maggiormente trasparente il processo di produzione dei risultati delle RS e di rendere le proprie revisioni più facilmente utilizzabili ai fini dell'applicazione del metodo GRADE (Higgins 2008). Peraltro, del gruppo promotore del metodo GRADE fanno parte molti autori che hanno un ruolo rilevante proprio all'interno della Cochrane Collaboration (Andrew David Oxman, Gordon Guyatt, Holger Schünemann, tra gli altri).

La valutazione del rischio di distorsione negli studi primari. Le risk of bias tables

Il livello di fiducia che possiamo riporre nelle conclusioni tratte da una revisione sistematica sugli effetti degli interventi considerati dipende dalla validità degli studi inclusi da cui sono tratti i dati e i risultati. La valutazione della validità degli studi inclusi è una componente es-

senziale di una revisione e dovrebbe influenzare l'analisi, l'interpretazione dei risultati e le conclusioni (Higgins 2008).

Si possono tradizionalmente considerare due dimensioni della validità di uno studio. La prima riguarda l'appropriatezza della domanda a cui lo studio vuol dare risposta, spesso descritta come validità esterna, e la sua valutazione dipende dallo scopo per cui lo studio deve essere utilizzato. La validità esterna è strettamente connessa con la generalizzabilità o applicabilità dei risultati di uno studio. La seconda dimensione della validità di uno studio fa riferimento al fatto che esso risponda correttamente al quesito della ricerca, in altri termini se è privo di distorsioni (spesso descritta come validità interna).

Un bias rappresenta un errore sistematico, ossia una deviazione dal vero nei risultati o nelle inferenze da essi derivati. Le distorsioni possono condurre a una sovrastima o sottostima del vero effetto dell'intervento e possono avere dimensioni differenti: alcune sono piccole e quindi di scarsa importanza rispetto all'evento osservato, altre possono essere sostanziali e quindi i risultati possono essere quasi completamente attribuibili alle distorsioni stesse. Inoltre specifiche fonti di bias possono agire in direzioni differenti in studi diversi. In generale è impossibile sapere quanto le distorsioni possono aver inciso sui risultati di un particolare studio, anche se esistono prove empiriche, specialmente per gli studi clinici randomizzati, che alcuni specifici problemi nel disegno, nella conduzione e nell'analisi inducano distorsioni. Dal momento che i risultati di uno studio di fatto possono non essere distorti nonostante la presenza di errori metodologici, risulta più appropriato parlare di rischio di distorsione (risk of bias), piuttosto che di distorsione del risultato (bias).

Differenze nel risk of bias possono aiutare a spiegare la variazione nei risultati degli studi inclusi in una revisione sistematica, ad esempio possono aiutare a spiegare l'eterogeneità dei risultati. Studi più rigorosi hanno maggior probabilità di dar luogo a risultati più vicini al risultato reale. Una metanalisi di risultati di studi con validità differente può dar luogo a conclusioni falsamente positive – concludendo erroneamente che un intervento sia efficace – se gli studi meno rigorosi tendono a sovrastimare l'effetto dell'intervento a causa dei bias in essi contenuti. Ma potrebbero anche giungere allo stesso modo a conclusioni falsamente negative – concludendo erroneamente che l'intervento non è efficace – se le distorsioni contenute negli studi meno rigorosi conducono a una sottostima dell'effetto (Detsky 1992). Risulta

quindi chiaro come sia importante valutare il risk of bias in tutti gli studi inclusi in una revisione, indipendentemente dalla variabilità prevista tanto nei risultati quanto nella validità degli stessi. Infatti i risultati potrebbero essere consistenti fra gli studi, ma tutti gli studi potrebbero essere affetti da distorsioni, e in questo caso le conclusioni della revisione non dovrebbero avere la forza di quelle che si basano su una serie di studi con risultati consistenti fra loro, ma condotti in maniera rigorosa.

Il bias non dovrebbe essere confuso con l'imprecisione. Il primo fa riferimento all'errore sistematico e questo vuol dire che molteplici replicazioni dello stesso studio condurrebbero in media alla stessa risposta errata. L'imprecisione fa riferimento all'errore casuale e questo significa che molteplici replicazioni dello stesso studio produrranno stime degli effetti diverse a causa della variazione di campionamento pur dando luogo in media alla stessa risposta corretta. I risultati di studi più piccoli sono soggetti a una più ampia variazione di campionamento e quindi a una minor precisione e l'imprecisione si riflette negli intervalli di confidenza stimati intorno all'effetto dell'intervento e nel peso dato ai risultati di ciascuno studio nell'ambito della metanalisi. Risultati più precisi danno luogo a un peso maggiore.

Il bias deve essere anche distinto dalla qualità. Le revisioni Cochrane si focalizzano infatti sulla valutazione del risk of bias piuttosto che sulla valutazione della qualità metodologica. E questo per le seguenti ragioni:

- Una considerazione chiave in una revisione Cochrane riguarda quanto siano affidabili i risultati degli studi inclusi ed è la valutazione del risk of bias a rispondere a questa specifica domanda.
- Uno studio potrebbe essere condotto secondo i più alti standard possibili, ma continuare a presentare importanti rischi di distorsione. Ad esempio, in molte situazioni è impraticabile o impossibile predisporre la cecità dei partecipanti o del personale sanitario rispetto al gruppo di intervento nello studio. Sarebbe un giudizio inappropriato descrivere tutti questi studi come di bassa qualità. Ma questo non significa che essi siano privi di distorsioni derivanti dalla consapevolezza dell'intervento assegnato.
- Alcuni indicatori di qualità nella ricerca medica – come l'approvazione del comitato etico, l'aver condotto un calcolo della dimensione del campione o la descrizione di uno studio in linea con i detta-

mi del CONSORT Statement (Moher 2001) – anche se presenti, non implicano in modo diretto l'esclusione del rischio di distorsione.
- Enfatizzare la valutazione del rischio di distorsione (risk of bias) supera l'ambiguità esistente tra la qualità della descrizione degli studi e la qualità della ricerca sottostante, anche se non supera il problema del dover almeno in parte basarsi sui primi per valutare la seconda.

Il risk of bias presente nei risultati di ciascuno degli studi che contribuiscono alla stima dell'effetto è uno dei molteplici fattori che devono essere presi in considerazione quando si giudichi la qualità di un intero corpo di prove, come verrà delineato nel contesto delle *summary of findings tables*. In quel caso per qualità delle prove si intende quale sia il livello di fiducia che possiamo riporre nel fatto che la stima dell'effetto di un intervento, considerando tutti gli studi esaminati rispetto a uno specifico esito, sia vicina al valore reale.

L'affidabilità dei risultati di uno studio clinico randomizzato (RCT) dipende dalla misura in cui è stato possibile evitare di introdurre distorsioni e un elemento essenziale e imprescindibile di una revisione sistematica è la valutazione del rischio di distorsione presente nei risultati di ciascuno degli studi inclusi. Questa rappresenta l'area di ricerca in cui lo studio del rischio di distorsione è stato maggiormente sviluppato. Per ciascuna delle potenziali fonti di distorsione è importante considerare la loro probabile dimensione e direzione rispetto all'effetto dell'intervento considerato.

Per la valutazione del rischio di distorsione negli RCT, la Cochrane Collaboration scoraggia l'utilizzo di scale o checklist. Moher ha identificato 25 scale e 9 checklist utilizzate per la valutazione della validità interna o della "qualità" degli studi clinici randomizzati (Moher 2001). Queste comprendono da 3 a 57 voci da valutare. Spesso, sono tuttavia inclusi criteri non pertinenti alla valutazione della validità interna, quali il calcolo della potenza dello studio – che è più pertinente alla valutazione della precisione della stima dell'effetto – o la chiara descrizione dei criteri di inclusione ed esclusione – che sono maggiormente pertinenti all'applicabilità dei risultati.

In una revisione Cochrane il processo di valutazione del rischio di distorsione negli studi inclusi si realizza attraverso uno strumento che è stato sviluppato e implementato all'interno del software RevMan5, dedicato alla produzione delle revisioni sistematiche. Tale strumento

è rappresentato dalle *Risk of Bias table* (RoB table, tabella 1.2), che hanno lo scopo di valutare il rischio di distorsione nello studio nell'ambito di sei diversi domini:

- metodo utilizzato per generare la sequenza di randomizzazione (*sequence randomisation*);
- metodo utilizzato per mascherare la sequenza di allocazione al trattamento (*allocation concealment*);
- misure adottate per assicurare la "cecità" dei partecipanti allo studio e del personale sanitario e non, rispetto al trattamento assegnato (*blinding of participants, personnel and outcome assessors*);
- completezza dei dati disponibili per ciascun esito considerato (*incomplete outcome data*);
- descrizione selettiva degli esiti (*selective outcome reporting*);
- altri fonti di distorsione (ad esempio: bias dovuto all'interruzione precoce di uno studio a causa dei benefici) (*other sources of bias*).

Il rischio di distorsione per i domini relativi alla sequenza di randomizzazione, alla sequenza di allocazione al trattamento e alla descrizione selettiva degli esiti può essere valutato considerando il singolo studio nella sua globalità. Per la valutazione dello stesso nell'ambito dei domini relativi alla cecità e completezza dei dati è importante introdurre nella RoB table tante voci quanti sono gli esiti presi in esame perché il giudizio può essere diverso a seconda dell'esito considerato.

La descrizione di ogni dominio nella RoB table si compone di due parti. La prima consta di una descrizione il più possibile esaustiva e dettagliata di quanto è accaduto. Questo dovrebbe avvenire non solo riportando quello che il singolo studio descrive (riportando le frasi del testo da cui le informazioni sono tratte con il relativo riferimento di pagina), ma anche facendo riscorso ad altre informazioni che possono essere derivate dai protocolli, da pubblicazioni di commento, o anche contattando gli autori dello studio primario. Si dovrebbero includere anche altre informazioni che possono influire sul giudizio relativo al rischio di distorsione proveniente da studi condotti dagli stessi autori su quell'argomento.

TABELLA 1.2 • ESEMPIO DI RISK OF BIAS TABLE

Dominio	Giudizio	Descrizione
Sequenza di randomizzazione adeguata?	Sì	*«patients were randomly allocated»* pag. 12 Commento: Probabilmente è stato fatto, in quanto precedenti lavori degli stessi autori descrivono in modo dettagliato l'uso della sequenza di randomizzazione
Mascheramento della sequenza di allocazione al trattamento adeguato?	No	*«using a table of random numbers»* pag. 13 Commento: Probabilmente non fatto
Cecità? (esiti riportati dai pazienti)	Sì	*«double blind, double dummy» «High and low dose tablets or capsules were indistinguishable in all aspects of their outward appearance. For each drug an identically matched placebo was available…»* pag. 13 Commento: Probabilmente fatto
Cecità? (mortalità)	Sì	Ottenuta dalle cartelle cliniche. Gli autori della revisione non ritengono che questo introduca dei bias.

Adattato da Higgins 2008

La seconda parte richiede di assegnare un giudizio relativo al rischio di distorsione per ciascuna specifica voce (considerando le specifiche relative a *blinding, incomplete outcome data e other sources of bias*). Il giudizio può essere formulato attraverso la risposta a quesiti pre-specificati inerenti l'adeguatezza dello studio in relazione al dominio in toto o per ciascuna singola voce considerata (laddove sia previsto). Un giudizio uguale a "Sì" indica un basso rischio di distorsione, "No" indica un alto livello di rischio di distorsione e "Non definibile" indica un livello di rischio di distorsione che o non è chiaro o non è noto.

La sintesi delle valutazioni del rischio di distorsione per ogni esito considerato

Per poter trarre conclusioni sul rischio di distorsione che globalmente pesa sulla stima dei risultati per ogni esito considerato in una revisione è necessario fornire una stima sintetica di quelli rilevati in ciascun studio. Per raggiungere tale risultato la Cochrane Collaboration

esorta gli autori delle revisioni a non usare scale di sintesi, per le motivazioni già citate rispetto alla valutazione del rischio di distorsione in ogni studio (Higgins 2008).

Per prima cosa un autore dovrebbe decidere quali sono i domini più rilevanti per la sua revisione sulla base della tipologia e dell'area indagata dalla stessa. Ad esempio la "cecità dei partecipanti allo studio e del personale sanitario" è particolarmente rilevante quando si valutino esiti "soggettivi" quali il livello di dolore. I livelli raccomandati di sintesi del rischio di distorsione in una revisione sono due:

- sintesi effettuata per ciascun esito all'interno di ciascun studio. Questo è particolarmente auspicabile e raccomandato, in quanto il rischio di distorsione dei risultati può variare nello stesso studio rispetto agli esiti considerati. In questo caso gli autori dovranno valutare il rischio di distorsione per ogni esito considerando sia i domini valutati a livello dello studio in toto (come il metodo di generazione della sequenza di randomizzazione, o il metodo di mascheramento della sequenza stessa), sia i domini valutati specificamente per singolo esito (ad esempio la cecità dei pazienti, degli operatori sanitari e dei valutatori degli esiti).
- sintesi effettuata per ciascun esito fra tutti gli studi. Questo è il livello che dovrebbe essere considerato nel momento in cui si fa una metanalisi o comunque si sintetizzano i risultati di una RS e dovrebbe poi essere incorporato nei giudizi sulla "qualità delle prove" nelle tavole di sintesi dei risultati (summary of findings tables).

La sintesi globale del rischio di bias ossia quella relativa a tutti gli esiti considerati e agli studi inclusi per ciascuno di essi, è da riservarsi alle linee-guida. Ciò per due motivi. Primo, perché tale sintesi richiede l'applicazione di un giudizio di valore su quali esiti siano critici o meno per assumere una decisione (ad esempio se trattare o meno con un farmaco piuttosto che un altro una categoria di pazienti) e spesso i dati che provengono dagli studi inclusi in una RS non sono sufficienti a fornire informazioni su esiti che possono essere critici, quale gli eventi avversi di un intervento. Secondo, perché il giudizio relativo alla criticità degli esiti può variare in situazioni differenti sulla base dei diversi valori dei soggetti o sulla base della variazione di specifici fattori a livello locale, quali il rischio di base di sviluppare un certo esito da parte di individui cui si potrebbe applicare un determinato intervento.

Le summary of findings tables

Le "tavole di sintesi dei risultati" (summary of findings tables – SoF) presentano i principali risultati di una revisione in un formato tabulare, semplice e trasparente. Esse mettono a disposizione le informazioni "chiave" riguardo alla qualità delle prove, la dimensione dell'effetto degli interventi esaminati e una sintesi di tutti i dati disponibili per ciascuno degli esiti importanti in una revisione (Higgins 2008). A tal fine la pianificazione delle SoF deve avvenire fin dal momento della stesura del protocollo, con la selezione degli esiti che vi saranno inclusi.

La selezione degli esiti è un momento importante e sensibile nella pianificazione della revisione – spesso non formalizzata nel passato nelle revisioni Cochrane (Pregno 2008) – e fondamentale per assicurare un set ottimale di informazioni utili a chi dovrà assumere decisioni cliniche o in sanità pubblica. Solo definendo chiaramente il quesito della revisione ed elencando e prioritarizzando tutti gli esiti importanti per i pazienti fin dalla stesura del protocollo si otterrà tale scopo.

In una SoF devono essere descritti in capo alla tavola (tabella 1.3): il quesito a cui vuole rispondere, la popolazione di interesse, l'ambito assistenziale in cui gli studi sono stati condotti, gli interventi considerati (l'intervento sperimentale) e, in ultimo, il gruppo di controllo rispetto all'intervento sperimentale.

Dal punto di vista del formato, la presentazione standard di una SoF include i seguenti sei elementi in un formato fisso:

- una lista di tutti gli esiti importanti sia quelli favorevoli sia quelli sfavorevoli;
- una misura relativa al rischio di base per ciascun esito, che può essere espresso come una frequenza, come una media o in altro modo appropriato e rappresenta il rischio nel gruppo di controllo oppure in una popolazione per cui il rischio di quell'esito sia noto;
- una stima relativa e assoluta della dimensione dell'effetto dell'intervento (se siano entrambi appropriati) o altrimenti uno solo di essi;
- il numero di studi da cui sono state tratte le informazioni per quell'esito e il numero di partecipanti a tali studi;
- un giudizio sulla qualità delle prove per ciascun esito (che può variare di esito in esito);
- uno spazio per i commenti.

TABELLA 1.3 • ESEMPIO DI UNA SUMMARY OF FINDINGS TABLE

Compression stockings compared with no compression stockings for people taking long flights

Patients or population: Anyone taking a long flight (lasting more than 6 hours)
Settings: International air travel
Intervention: Compression stockings
Comparison: Without stockings

Outcomes	**Illustrative comparative risks** (95% CI)		Relative effect (95% CI)	Number of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk	Corresponding risk				
	Without stockings	**With stockings**				
Symptomatic deep vein thrombosis (DVT)	See comment	See comment	Not estimable	2821 (9 studies)	See comment	0 participants developed symptomatic DVT in these studies
Symptom-less deep vein thrombosis	**Low risk population**		**RR 0.10** (0.04 to 0.26)	2637 (9 studies)	++++ **High**	
	10 per 1000	**1 per 1000** (0 to 3)				
	High risk population					
	30 per 1000	**3 per 1000** (1 to 8)				
Superficial vein thrombosis	**13 per 1000**	**6 per 1000** (2 to 15)	**RR 0.45** (0.18 to 1.13)	1804 (8 studies)	+++0 **Moderate**	
Oedema Post-flight values measured on a scale from 0, no oedema, to 10, maximum oedema	The mean oedema score ranged across control groups from **6 to 9**	The mean oedema score in the intervention groups was on average **4.7 lower** (95% CI – 4.9 to 4.5)		1246 (6 studies)	++00 **Low**	
Pulmonary embolus	See comment	See comment	Not estimable	2821 (9 studies)	See comment	0 participants developed pulmonary embolus in these studies
Death	See comment	See comment	Not estimable	2821 (9 studies)	See comment	0 participants died in these studies
Adverse effects	See comment	See comment	Not estimable	1182 (4 studies)	See comment	The tolerability of the stockings was described as very good with no complaints of side effects in 4 studies

Adattato da Higgins 2008

Quando in una revisione siano presenti oltre all'analisi principale anche altre analisi per ciascuna di queste si potrà produrre una SoF separata. Una delle motivazioni principali è legata al diverso rischio delle popolazioni in esame nell'ambito delle analisi secondarie.

Tecnicamente per produrre una SoF si può utilizzare il software GRADE Profiler, in grado di raccogliere le informazioni direttamente dal software RevMan5, attualmente in uso nella Cochrane Collaboration per la produzione delle revisioni sistematiche.

Uno degli elementi più significativi è che gli autori devono includere nella SoF tutti gli esiti importanti sia in presenza che in assenza dei relativi dati. Un altro elemento importante è che in presenza di più esiti simili (ad esempio misurati in tempi differenti oppure più eventi avversi minori), questi vengano combinati fra di loro sulla base del loro significato e importanza per rispondere ai quesiti della revisione in modo da evitare informazioni ridondanti e non essenziali. Ogni motivazione della combinazione degli esiti deve essere esplicitata e chiaramente descritta in una nota della SoF, così da rendere la sintesi più informativa per chi deve prendere le decisioni. Sempre nell'ottica di presentare solo le informazioni essenziali, all'interno di una SoF non si devono includere più di sette esiti clinici.

CONCLUSIONI

La qualità delle informazioni e la modalità di presentazione dei risultati sono due elementi importanti per cercare di risolvere i problema dell'aggiornamento degli operatori sanitari. L'utilizzo delle informazioni derivanti da studi affetti da errori sistematici può provocare gravi effetti sulla salute umana, quali malattie evitabili e addirittura morti premature. C'è quindi il rischio di concludere che alcuni trattamenti siano utili, quando non lo sono, e viceversa, se non viene posta un'adeguata attenzione al ruolo del rischio di distorsione che può affliggere i risultati degli studi, al ruolo del caso e quindi, più in generale alla qualità e validità della ricerca su cui fondiamo gran parte dell'operato clinico e in sanità pubblica.

Gli ultimi cinque anni possono essere definiti come un momento di transizione verso una maggior esplicitazione sia della qualità delle prove sia dei risultati delle sintesi della letteratura, nei diversi strumenti per il miglioramento dell'appropriatezza clinica e di sanità pubblica.

Il metodo GRADE sta rappresentando una delle più importanti novità nel campo della valutazione dei servizi sanitari, con ricadute sia dal punto di vista degli operatori sanitari sia dei pazienti. In generale, gli ideatori di questo metodo hanno cercato di lavorare sulle rigidità dell'evidence-based medicine, basata, secondo una interpretazione eccessivamente forzata, solo sugli studi clinici randomizzati. L'esplicitazione della domanda clinica e la scelta degli esiti importanti sono state poste così al centro del percorso di valutazione degli interventi: tutto questo è stato associato a un'esplicita valutazione delle prove con un riflesso positivo sia sulla trasparenza delle informazioni sia sulla trasferibilità delle stesse nei diversi ambiti applicativi.

Questa nuova filosofia è stata guardata con molto interesse in brevissimo tempo a livello internazionale dalle principali organizzazioni che effettuano attività di valutazione degli interventi sanitari. L'adozione da parte della Cochrane Collaboration di una metodologia che migliora la trasparenza e la capacità informativa delle RS e ne facilita l'uso ai fini di produzione delle raccomandazioni, secondo la metodologia GRADE, ha comportato un grande cambiamento anche all'interno della CC stessa.

L'innovazione concettuale e pratica introdotta con il GRADE permette, in generale, una più complessiva valutazione della qualità delle prove e una maggiore fruibilità delle informazioni, in particolare per gli esiti considerati importanti per e dai pazienti. Così come descritto da Evans esiste infatti molta differenza nella valutazione degli esiti da parte dei medici e da parte dei pazienti (Evans 2007). In generale, anche questi ultimi dovrebbero avere un maggior ruolo nella valutazione degli interventi sanitari, come ad esempio per gli effetti indesiderati dei trattamenti. L'uso degli strumenti proposti dal metodo GRADE potrebbe aiutarli ad avere sempre maggior peso sia nelle decisioni legate agli interventi da introdurre nella pratica clinica sia nella scelta delle ricerche da finanziare.

BIBLIOGRAFIA

ACP (1992). American College of Physicians Guidelines for counseling postmenopausal women about preventive hormone therapy. Ann Intern Med 117:1038-41.

Albers G, Dalen JE, Laupacis A et al. (2001). Antithrombotic therapy in atrial fibrillation. Chest 119:194S-206S.

Altman DG, Chalmers I, Egger MS et al. (1995). Systematic reviews in health care meta-analysis in context. London: BMJ Publishing Group.

Atkins D, Eccles M, Flottorp S et al. (2004). Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches. The GRADE Working Group. BMC Health Serv Res 4(1): 38.

Carlsen, B Norheim FO (2008). "What lies beneath it all?". An interview study of GPs' attitudes to the use of guidelines. BMC Health Services Research 8: 218.

De Palma R, Liberati A, Papini D et al. (2009). La produzione di raccomandazioni cliniche con il metodo GRADE. L'esperienza sui farmaci oncologici. Dossier n° 172. Disponibile online (http://asr.regione.emilia-romagna.it/wcm/asr/collana_dossier/doss172.htm).

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. Journal of Clinical Epidemiology 45: 255-65.

Evans I, Thornton H, Chalmers I (2007). Come sapere se una cura funziona. Roma: Il Pensiero Scientifico Editore.

Fuster V, Rydén LE, Asinger RW et al. (2001). ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation. J Am Coll Cardiol 38:1266.

Glasziou P, Vandenbroucke J, Chalmers I (2004). Assessing the quality of research. BMJ 328: 39-41.

GRADE Working Group (2004). Grading quality of evidence and strenght of recommendations. BMJ 328: 1490-8.

Guyatt GH, Oxman AD, Kunz R et al. for the GRADE Working Group (2008a). GRADE: going from evidence to recommendations. BMJ 336:1049-51.

Guyatt GH, Oxman AD, Kunz R et al. for the GRADE Working Group (2008b). GRADE: what is "quality of evidence" and why is it important to clinicians? BMJ 336: 995-8.

Guyatt GH, Oxman AD, Vist GE et al. for the GRADE Working Group (2008c). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336: 924-6.

Higgins JPT, Green S eds (2008). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 [updated February 2008]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Humphrey LL, Chan BK, Sox HC (2002). Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Ann Intern Med 137: 273-84.

Institute of Medicine (1990). Clinical Practice Guidelines: Directions for a New Program. Field MJ, Lohr KN, eds. Washington, DC: National Academy Press.

Lacchetti C, Guyatt G (2002). Surprising results of randomized trials. In: Guyatt G, Drummond R, eds. Users' guides to the medical literature: a manual of evidence-based clinical practice. Chicago, IL: AMA Press.

Moher D, Schulz KF, Altman DG (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 357: 1191-4.

Mulrow CD (1995). Rationale for systematic reviews In: Chalmers I, Altman CD, eds. Systematic reviews. London: BMJ Books.

Oxman AD, Guyatt GH (1988). Guidelines for reading literature reviews. CMAJ 138: 697-703.

Popper KR (1970). Logica della scoperta scientifica. Il carattere autocorrettivo della scienza. Torino: Einaudi Editore.

Pregno S, Oxman AD (2008). Implementation of "risk of bias (RoB) tables and Summary of findings" (SoF) tables in Cochrane reviews: a pilot study. German Journal of Evidence and Quality in Health Care Suppl VI: 60.

Programma Nazionale Linee Guida (2002). Come produrre, diffondere e aggiornare raccomandazioni per la pratica clinica. Manuale metodologico – Roma.

Sackett DL, Wennberg JE (1997). Choosing the best research design for each question. BMJ 315: 1636.

Schünemann HJ, Oxman AD, Broze J et al. (2008). Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 336: 1106-10.

SIGN (2004). A guideline developers' handbook. SIGN Publication No. 50 Published February 2001 Last updated May 2004.

Thompson DC, Rivara FP, Thompson R (2000). Helmets for preventing head and facial injuries in bicyclists. Cochrane Database Syst Rev 2:CD001855.