QCN
Gruppo di Studio per la
Qualità delle Cure in Neonatologia

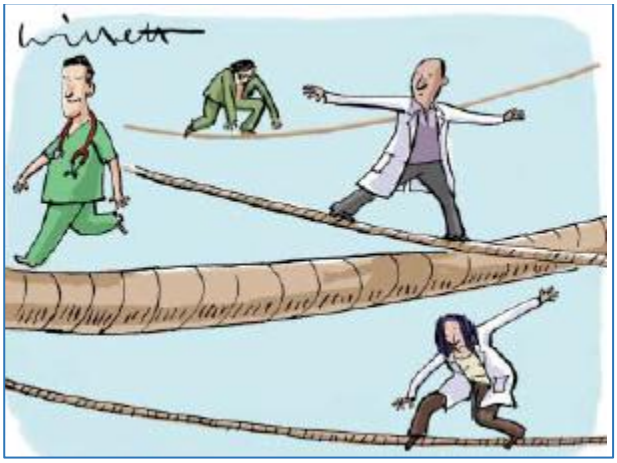
GRADE – AGREE

MANUALE DI CONSULTAZIONE

PER FORMULARE RACCOMANDAZIONI

E

VALUTARE LINEE GUIDA



A Cura di

Daniele Merazzi
Roberto Bellù
Alessandra Coscia
Paola Lago
Luigi Gagliardi

In copertina disegno tratto da GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328:1490–4

Ia Edizione Dicembre 2017

In ricordo di Alessandro e Carlo

… l'incertezza è parte integrante del progresso scientifico ed è ineliminabile dalla pratica della medicina.

… le metodologie di valutazione dei trattamenti sono tutt'altro che perfette, errori o effetti non previsti fanno parte del possibile e possono essere ridotti solo puntando ad una ricerca di migliore qualità e ad una maggiore capacità critica anche dei pazienti sui limiti della medicina e sulla necessità continua di monitorarne i possibili effetti avversi.

Alessandro Liberati

… si tratta, di mettere in discussione quello che è stato chiamato «bias ottimistico», che induce a fare comunque qualcosa anche se non di provata efficacia e con scarsissime probabilità che possa essere utile…

Aderire a un modello, perché se ne riconosce la superiorità, significa muoversi all'interno di un paradigma e capire ed essere capiti solo in virtù di un linguaggio comune e condiviso.

Carlo Corchia

GRADE – AGREE

MANUALE DI CONSULTAZIONE

PER FORMULARE RACCOMANDAZIONI

E

VALUTARE LINEE GUIDA

A Cura di

Daniele Merazzi

UOC Neonatologia – Terapia Intensiva Neonatale, Dipartimento Materno-Infantile, Congregazione delle Suore dell'Addolorata, Ospedale Valduce, Como

Roberto Bellù

UOC Neonatologia e Terapia Intensiva Neonatale, Dipartimento Materno-Infantile, Ospedale A. Manzoni, ASST Lecco

Alessandra Coscia

Terapia Intensiva Neonatale Clinica dell'Università, Dipartimento di Ostetricia e Ginecologia, Ospedale S.Anna, Azienda Ospedaliero-Universitaria, Città della Salute e della Scienza di Torino

Paola Lago

Terapia Intensiva Neonatale e Patologia Neonatale, Dipartimento di Salute della Donna e del Bambino, Azienda Ospedaliera-Università di Padova

Luigi Gagliardi

UOC Pediatria, Area Funzionale Complessa Materno-Infantile, Ospedale Versilia, Viareggio (LU), USL Toscana Nord Ovest

INDICE

PREFAZIONE

La medicina moderna, tenuto conto dei principi dell'Evidence Based Medicine (EBM o "medicina basata sulle prove di efficacia") ha prodotto efficaci strumenti come linee guida e raccomandazioni.

La crescente popolarità dei suddetti strumenti ha indotto numerose organizzazioni a sviluppare metodi di classificazione delle qualità delle prove e della forza delle raccomandazioni. Ad oggi sono disponibili oltre una decina di sistemi di classificazione, non sempre coerenti nella formulazione di raccomandazioni ad uso clinico.

Il Direttivo del Gruppo di Studio (GdS) Qualità delle Cure (QCN) in Neonatologia della Società Italiana di Neonatologia (SIN) ha ritenuto di dover identificare nel metodo GRADE (Grading of Recommendations, Assessment, Development and Evaluation) il sistema di riferimento per la valutazione della qualità dell'evidenza della produzione scientifica e della relativa formulazione della forza delle raccomandazioni.

GRADE ha lo scopo di orientare e facilitare metodologicamente i vari GdS della SIN nella produzione di linee guida cliniche e raccomandazioni secondo criteri di riconosciuta validità internazionale.

La recente approvazione della legge Gelli-Bianco sulla responsabilità professionale riconosce alle linee guida un ruolo fondamentale nel contenzioso giuridico. Per tale ragione le società scientifiche devono direttamente (vedi adesione all'albo) occuparsi della qualità delle linee guida prodotte e delle relative raccomandazioni cliniche formulate.

Il Guideline International Network (GIN) ha analizzato recentemente (marzo 2017) la qualità delle linee guida prodotte da 403 società scientifiche italiane. Sono state analizzate 712 linee guida. Di queste solo 75 (10%) sono state dichiarate metodologicamente complessivamente accettabili. Delle 75 ben 42 (56%) sono state prodotte da 2 sole società scientifiche che hanno adottato un manuale metodologico per produrre linee guida a riprova del fatto che l'utilizzo di metodi adeguati produce risultati eccellenti.

Pertanto il GdS QCN suggerisce alla SIN l'impiego e l'implementazione del manuale qui prodotto.

Si ringraziano il Presidente SIN prof. Mauro Stronati e tutto il Direttivo per il sostegno dato all'iniziativa del Gds.

Daniele Merazzi

Segretario Gruppo di Studio Qualità delle Cure in Neonatologia della Società Italiana di Neonatologia triennio 2014-2017

INTRODUZIONE

La definizione di linea guida risale al 1990, quando l'Institute of Medicine le definì come "raccomandazioni di comportamento clinico prodotte con metodi sistematici per assistere medici e pazienti nel decidere le modalità di assistenza più appropriate in specifiche circostanze cliniche" (1). Questa definizione, ribadita nel 2011, rimane completamente valida e sottolinea la valenza clinica e nel contempo metodologica (il metodo sistematico) delle linee guida.

Nel corso di questi ultimi anni l'approccio alla formulazione e all'utilizzo è molto variato, con una crescita esponenziale dei soggetti produttori di linee guida ma anche la diffusione di documenti di dubbio significato clinico e scientifico (2). Dal punto di vista del metodo, diversi autorevoli enti internazionali (SIGN, NICE, AHRQ) e nazionali (GIMBE, PNLG), oltre che numerose società scientifiche, hanno via via promosso ed adottato metodi simili per garantire il massimo rigore metodologico nella formulazione delle raccomandazioni. Questo aspetto risulta particolarmente rilevante anche in funzione dell'utilizzo potenziale delle linee guida quali strumenti essenziali di governo clinico, utilizzati anche per definire standard assistenziali ed indicatori di processo. In quest'ottica è inoltre rilevante il recente riferimento legislativo alle linee guida come uno dei parametri di riferimento nella valutazione della responsabilità professionale (LEGGE GELLI).

Recentemente si è verificata una convergenza metodologica su un approccio ampiamente condiviso che cerca di coniugare nella formulazione delle linee guida la sistematicità ed il rigore metodologico con gli aspetti clinici: il metodo GRADE (Grading of Recommendations, Assessment, Development and Evaluation) (3) proposto nel 2004 si è rapidamente imposto come metodologia di riferimento per gli sviluppatori e gli utilizzatori di linee guida ed è stato adottato da numerosi enti e società scientifiche, tra i quali l'OMS, la Cochrane Collaboration, l'ILCOR, il CDC, l'NHS, il SIGN e molte altre ancora. La convergenza metodologica su questo approccio rende più valido, trasparente e condiviso il processo di formulazione e valutazione delle linee guida, aumentandone quindi l'autorevolezza scientifica.

Alcuni aspetti cruciali del metodo GRADE, che ne hanno determinato il successo, sono così sintetizzabili:

- Considerazione esplicita della relativa importanza dei vari outcome.
- Separazione chiara tra qualità dell'evidenza e forza delle raccomandazioni.
- Criteri espliciti per elevare (upgrade) o ridurre (downgrade) la qualità dell'evidenza.
- Riconoscimento esplicito dei valori e delle preferenze alla base delle raccomandazioni.
- Processo trasparente per passare dalle evidenze alle raccomandazioni.
- Esplicito consiglio a formulare raccomandazioni anche quando c'è poca evidenza.
- Chiara e pragmatica interpretazione di raccomandazioni "forti" e "deboli".

Questi punti, che verranno dettagliati in seguito, rappresentano l'evoluzione, per certi versi anche radicale, dei precedenti metodi che legavano più automaticamente il tipo di studio alla forza delle raccomandazioni, non considerando sufficientemente la qualità degli studi e l'importanza degli outcome e dei valori clinici nel processo di formulazione delle linee guida.

Questo documento presenta sinteticamente gli aspetti metodologici e clinici più rilevanti del metodo GRADE per la formulazione delle linee guida e del metodo AGREE (complementare a GRADE) per la valutazione delle stesse, aggiornando i precedenti documenti prodotti (4) e proponendoli come riferimenti metodologici a supporto delle attività di produzione e valutazione delle linee guida.

Bibliografia

1. Committee to Advise the Public Health Service on Clinical Practice Guidelines IoM. Clinical practice guidelines: directions for a new program. Washington: National Academy Press; 1990
2. Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet 2000;355:103-106
3. Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328:1490–4
4. Bertino E, Corchia C, Gagliardi L, per conto del Comitato Direttivo della Società Italiana di Neonatologia. La produzione di linee guida nell'ambito della S.I.N. Riv Ital Pediatr. 2000;26:371-5.

GERARCHIA E GRADAZIONE DELLE EVIDENZE

La logica della ricerca scientifica e biomedica in particolare si basa sul confronto tra gruppi per cercare di evidenziare ed isolare l'effetto di trattamenti/interventi su specifici gruppi di pazienti. L'impianto degli studi clinici persegue lo scopo di massimizzare la possibilità di evidenziare questi effetti, quando vi sono, isolandoli dall'effetto di altre variabili confondenti. È quindi fondamentale, a questo proposito, identificare un disegno dello studio adatto a rispondere in modo appropriato a specifici quesiti di ricerca clinica.

Per disegno dello studio si intende la metodologia di base con la quale lo studio è stato organizzato, riferendosi principalmente alle categorie degli studi osservazionali e dei trial clinici randomizzati (RCT). Sebbene l'impianto concettuale sia simile per gli studi osservazionali e per gli RCT cioè il confronto degli esiti in due gruppi, omogenei per le altre caratteristiche, sottoposti (RCT) o esposti (coorte) a due "interventi" alternativi, i risultati degli studi osservazionali e degli RCT non sempre sono concordanti. Accanto ad esempi di buona concordanza esistono numerosi e a volte drammatici esempi di mancanza di concordanza tra studi osservazionali ed RCT (1). La mancanza di concordanza può essere legata a molti fattori, ma il principale, in ordine di importanza pratica ed anche concettuale, è il differente controllo del confondimento nei due tipi di studi. Si potrebbe infatti ipotizzare che in assenza di confondimento i risultati di uno studio di coorte e di un RCT siano perfettamente coincidenti, vista la natura del tutto similare dell'impostazione (figura 1): negli studi di coorte F è tipicamente un'esposizione, spesso ritenuta dannosa, mentre negli RCT è tipicamente un trattamento ritenuto benefico.

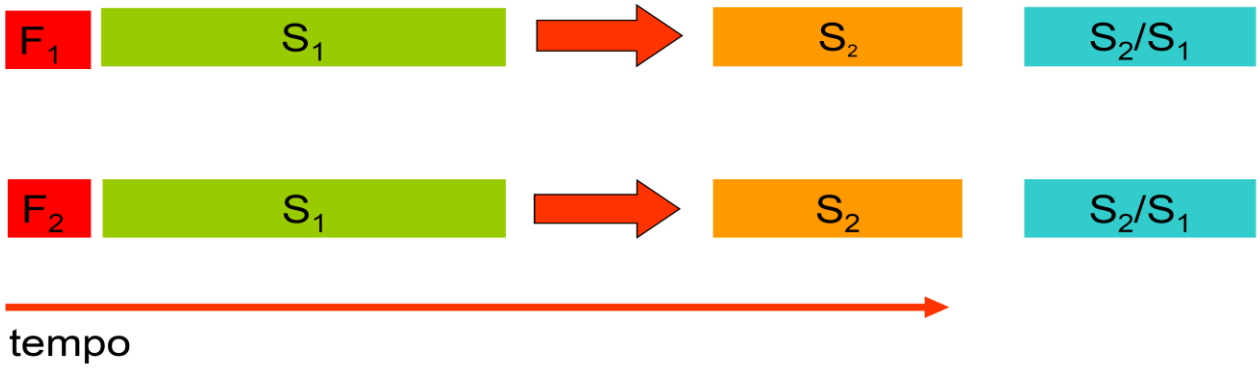


Figura 1: Impianto degli studi di coorte e degli RCT

La grande differenza tra gli studi osservazionali e gli RCT è naturalmente legata al fatto che in questi ultimi il "fattore di esposizione" (trattamento) è attribuito in maniera causale (random) e ciò garantisce in media l'equilibrio dei confondimenti tra i gruppi e l'eliminazione del bias ad essi legato nell'attribuzione del trattamento. Ciò non agisce, ovviamente, su altri aspetti critici dello studio, quali ad esempio la percentuale di perdita al follow-up, o la possibilità di violare la cecità o l'occultamento dell'assegnazione. E' quindi possibile che un RCT "mal condotto" presenti problemi metodologici che ne inficino parzialmente o totalmente la validità, mentre uno studio coorte ben condotto potrebbe presentare, su un analogo tema di ricerca clinica, una validità pari se non superiore a quella di un RCT mal condotto. Per questo motivo la valutazione della qualità degli studi clinici è essenziale in generale ed in particolare nel processo di valutazione del GRADE, che prevede la possibilità di elevare o ridurre il rating di uno studio in base alla sua valutazione qualitativa.

In linea generale, gli RCT partono da una posizione qualitativa più elevata perché se ben condotti sono gli studi che con più probabilità conducono a conclusioni affidabili, mentre gli studi

osservazionali partono da una posizione più bassa, ma questa gerarchia può variare, all'interno della metodologia GRADE, in base alla valutazione qualitativa degli studi e alla valutazione quantitativa delle stime sugli outcome in essi contenuti.

Occorre ricordare comunque come gli RCT non siano sempre realizzabili e come, per determinati outcome (soprattutto quelli rari e per gli eventi avversi) non rappresentino il disegno di studio ideale. Non va nemmeno dimenticato come la maggior parte delle pratiche mediche ed assistenziali si basino ancora su studi osservazionali. E' quindi discutibile un paradigma, forse a volte da qualcuno invocato, che fa coincidere le "prove di efficacia" con gli RCT. Le prove di efficacia, quali prodotti della ricerca medica e scientifica, sono tutti i tipi di studi, che devono essere valutati, in termini di risposte ai quesiti di ricerca proposti, per l'adeguatezza del disegno e per la qualità metodologica con i quali sono stati disegnati e condotti (tabella 1).

Tipo di quesito	**1^ scelta**	**2^ scelta**
Terapia	RCT	Serie storiche
Diagnosi	Studi di coorte	Case report
Prognosi	Studi di coorte	Serie di casi
Eziologia	Studi di coorte	Studi caso-controllo
Procedura clinica o assistenziale	RCT	Serie storiche
Intervento di comunità	Comunity intervention trial	Serie storiche

Tabella 1: Tipo di studio in relazione al quesito clinico

Questi aspetti sono centrali nel metodo GRADE e ne rappresentano una delle novità più sostanziali ed interessanti rispetto ai precedenti metodi di formulazione delle raccomandazioni che stabilivano una corrispondenza diretta e univoca tra tipo di studio e forza delle raccomandazioni (al vertice sempre le revisioni sistematiche con formulazione di raccomandazioni più forti, in basso gli studi osservazionali con formulazione di raccomandazioni più deboli).

Il concetto di gerarchia degli studi viene quindi ripreso dal metodo GRADE, in base alla maggiore o minore probabilità di produrre stime non distorte degli effetti, ma viene riveduta con l'incorporazione della validità metodologica e della qualità dello studio che diventa un fattore importante nel ridefinire la gerarchia delle prove di efficacia identificate per ogni singolo quesito clinico oggetto di raccomandazione (tabella 2).

Qualità delle prove	**Disegno dello studio**
Alta	RCT
Moderata	
Bassa	Studio osservazionale (coorte, caso controllo)
Molto bassa	

Tabella 2: Iniziale gerarchia delle prove secondo GRADE

Bibliografia

1. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342:1887-92

FORMULAZIONE DEL QUESITO CLINICO E VALUTAZIONE DEGLI OUTCOME

Nella formulazione delle linee guida usando l'approccio metodologico proposto dal GRADE è fondamentale, come primo intervento, definire con precisione il *quesito clinico* di maggiore rilevanza.

Questo è realizzato preferenzialmente seguendo la metodologia del PICO o PICOTT (Patient, Intervention, Comparator, Outcomes, Type of question, Type of article) (1), che verrà applicata a tutte la bibliografia rilevante reperita in letteratura. Ogni quesito clinico va quindi declinato per tipologia di paziente o problema a cui ci si riferisce per formulare la raccomandazione in un preciso contesto clinico, per il principale intervento, per il fattore prognostico o esposizione che si vuole considerare, per le più importanti alternative da mettere a confronto e per tipologia di esiti che si vuole considerare come misura dell'impatto dell'intervento. L'acronimo PICOTT considera anche il tipo di quesito (terapia, prognosi, eziologia o prevenzione) e il tipo di studio che potrebbe dare la miglior risposta al quesito (RCT, coorte, serie di casi, etc.) (2) (Tabelle 1a-1b).

P	**Paziente, Populazione o Problema**	**Come descriverei un gruppo di pazienti simile al mio**
I	**Intervento, Fattore prognostico o Esposizione**	**Qual è il principale intervento, fattore prognostico o esposizione che si vuole considerare?**
C	**Confronto o Intervento (se appropriato)**	**Qual è la principale alternativa da mettere a confronto con l'intervento considerato?**
O	**Outcome che si vuole misurare o ottenere**	**Cosa spero di ottenere, misurare, migliorare o modificare?**
T	**Che tipo di quesito è?**	**Diagnosi, Eziologia/Rischio, Terapia, Prognosi, Prevenzione.**
T	**Tipo di studio (RCT, Coorte, Serie di casi, etc.)**	**Che tipo di studio potrebbe dare la miglior risposta al quesito?**

Tabella 1a. Elementi del modello PICO-PICOTT per la formulazione del quesito clinico

Studio	Pazienti	Intervento	Confronto	Outcome	Qualità evidenza

Tabella 1b. Tabella PICO

È necessario quindi differenziare la tipologia del quesito clinico che può essere di tipo diagnostico, eziologico, prognostico, preventivo o terapeutico e la tipologia delle evidenze a supporto, che si ritengono più adeguate a rispondere a quel quesito clinico (studi randomizzati vs studi di coorte vs casistiche etc). Ad esempio per rispondere ad un quesito terapeutico e poter formulare raccomandazioni forti è fondamentale potersi avvalere di studi randomizzati, possibilmente doppio cieco, ben disegnati ed eseguiti, quindi a basso rischio di bias. In assenza di trial randomizzati, gli studi osservazionali ad alta numerosità con evidenza di una forte associazione tra trattamento e outcome, possono essere utili per dare una raccomandazione forte. Viceversa per

un quesito prognostico che stima la probabilità che un paziente vada incontro ad una determinata patologia, possono essere sufficienti studi di coorte o caso-controllo.

La metodologia PICO funziona meglio nei quesiti di carattere terapeutico e diagnostico. In fase di revisione della letteratura è utile riportare ogni studio, che risponda al quesito clinico in oggetto, in una tabella finale allo scopo di consentire poi la sintesi dei risultati.

È fondamentale pertanto, nel definire il quesito clinico, stabilire precisamente la popolazione e l'intervento, che vanno analizzati in sottogruppi se la magnitudine dell'effetto atteso è diverso (es. neonato vs pretermine vs lattante e l'efficacia e sicurezza degli oppioidi vs sedativi). Nel caso l'esito atteso non sia noto, si può improntare un quesito clinico generico, differenziandolo per sottogruppi di pazienti ed interventi e poi in base alla eterogeneità dei risultati trovati, differenziare il quesito clinico in base alle due variabili. Anche la chiarezza con cui viene fatto il confronto in relazione ad uno specifico intervento è importante, al fine di potere ben definire nella raccomandazione finale se tutti gli agenti considerati sono ugualmente efficaci e quindi vanno raccomandati.

Una raccomandazione per essere "ragionevole" richiede che siano considerati tutti gli outcome importanti per il paziente, ma non vanno esclusi gli esiti importanti per altri portatori di interesse o per la salute pubblica. A seconda della prospettiva, l'importanza dell'outcome può cambiare ed è per questo che va definita a priori se l'esito considerato è importante per il paziente, per la salute pubblica o per altri portatori di interesse. Quando le linee guida sono indirizzate ai clinici ed ai loro pazienti da trattare è evidente che la prospettiva deve essere quella del paziente. Gli outcome di maggior importanza riguardano morbilità e mortalità ma altrettanto importanti sono gli effetti avversi riscontrati.

Valutazione degli outcome

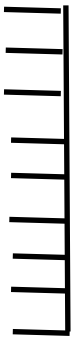
Gli outcome per il paziente sono distinti in critici, importanti ma non critici e di minore importanza. I primi due influenzano in modo significativo la forza della raccomandazione, mentre quelli di minore importanza per il paziente possono essere ininfluenti. In ogni caso l'outcome critico determina la qualità della evidenza complessiva. In altre parole se le evidenze di alta qualità supportano tutti gli outcome individuati nell'ambito del quesito clinico tranne uno, che però è sostenuto da bassi livelli di evidenza e questo outcome è riconosciuto come critico, la complessiva qualità delle evidenze sarà bassa. Se invece il panel di esperti ritiene che quel tipo di outcome è importante ma non critico, allora la qualità complessiva che sosterrà la raccomandazione finale sarà considerata elevata.
Non infrequentemente gli outcome più importanti per il paziente rimangono inesplorati, poiché spesso gli outcome critici sono relativamente infrequenti o si verificano in un lungo periodo di tempo, motivo per cui i ricercatori nel definire il disegno dello studio preferiscono misurare outcome sostitutivi o "surrogati". Per esempio quando si deve valutare l'efficacia di una terapia antiipertensiva, per prendere una decisione clinica, si misurano i valori di pressione arteriosa (outcome indiretto) piuttosto che l'impatto di quell'intervento terapeutico sulla mortalità cardiovascolare, infarto del miocardio e stroke che sono outcome critici ma di difficile misurazione nel breve-medio periodo.

Per valutare la relativa importanza degli outcome, devono essere seguiti tre passaggi:

1) Classificazione preliminare degli esiti prima di prendere visione delle evidenze: outcome critico, outcome importante ma non critico, outcome di bassa importanza in relazione al quesito clinico proposto. Questi giudizi si possono basare sull'esperienza degli esperti, dei pazienti o degli altri stakeholders. La Figura 1 presenta una gerarchia di outcome importanti

per il paziente riguardanti ad esempio l'impatto della terapia analgesica in corso di intubazione tracheale nel neonato. Il GRADE suggerisce una scala da 1 a 9, identificando a livello 7-9 gli outcome di importanza critica per prendere delle decisioni cliniche, il livello 4- 6 identifica gli outcome importanti, ma non critici ed il livello 1 -3 identifica gli outcome di limitata importanza.

2) Rivalutazione dell'importanza degli outcome dopo la revisione delle evidenze in letteratura, per assicurare che gli esiti importanti riportati in letteratura siano inclusi, qualora non fossero stati considerati all'inizio e per rivalutare la relativa importanza degli outcome alla luce delle evidenze disponibili.
3) Valutazione del bilancio tra effetti desiderabili ed indesiderabili di un determinato intervento per definire una raccomandazione e la sua forza.



Mortalità	9	Outcome Critico
Lesioni neurologiche maggiori	8	
Outcome neuroevolutivo	7	
Dolore/stress	6	Outcome importante ma non critico
Lesioni delle vie aeree	5	
Bradicardia	4	
Desaturazione	3	
Durata procedura/N° tentativi	2	
Ridotta motilità intestinale	1	Outcome non importante

Figura 1. Gradazione degli outcome per la valutazione dell'effetto della terapia analgesica per l'intubazione tracheale del neonato.

La valutazione preliminare e definitiva di un outcome permette di tenere conto dei risultati emersi dall'analisi della letteratura e renderlo più aderente alle evidenze disponibili.
A questo scopo si utilizzano, per la sintesi delle evidenze, gli strumenti messi a disposizione dal GRADE che consentono di tracciare precisamente il percorso valutativo. Questi strumenti sono l'Evidence Profile (Tabella 2, EP), un modulo che permette di fare la sintesi della qualità della evidenza in base agli outcome considerati, tenendo conto del bilancio costo/beneficio. L'altro strumento è il Summary of Finding (Tabella 3, SoF) che consente di sintetizzare le evidenze per tutti gli outcome considerati.

L'EP (Evidence Profile) comprende una dettagliata valutazione della qualità dell'evidenza ed include ogni specifico giudizio per ciascun fattore che determina la qualità dell'evidenza per ciascun outcome ed assicura che i giudizi espressi nella valutazione della letteratura siano tracciabili e revisionabili.

Valutazione della Qualità							Sintesi delle evidenze					
							N. Pazienti		Effetto		Qualità	Importanza
Studio	Design	Limitazioni	Inconsistenza	Evidenza indiretta	Imprecisione	Altre considerazioni	Caso	Controllo	R. Relativo (RR)	R. Assoluto (RA)		

Tabella 2. Il profilo delle prove (Evidence Profile)

La Tabella SoF include la valutazione complessiva della qualità dell'evidenza per ciascun outcome ma non i dettagli su cui è basata e può essere considerata la sintesi estrema del lavoro di revisione fatta. I due strumenti servono a scopi diversi.

Outcome	Studi/Pazienti	Sintesi delle evidenze	Qualità della evidenza
			⊗⊗⊗⊗
			⊗⊗⊗◯
			⊗⊗◯◯
			⊗◯◯◯

Tabella 3. La sintesi delle prove (SoF, Summary of Evidence)

Nell'ambito dello sviluppo delle linee guida la compilazione sistematica e puntuale della tabella EP assicura che vi sia accordo sul giudizio e tracciabilità alla base della valutazione della qualità della evidenza nell'ambito del gruppo di esperti. Mentre la tabella SoF fornisce le informazioni chiave utili non solo agli sviluppatori ma anche ai fruitori delle linee guida.

Essi sono di supporto determinante una volta che ci si appresta a passare dalla qualità delle evidenze alla forza della raccomandazione

Eseguire con precisione i primi passi del processo GRADE significa garantirne la rigorosità delle conclusioni finali cioè delle raccomandazioni.

Bibliografia

1. Crumeley E, Koufoglannakis D, Stabart K. Teaching EBP, part 1. *Case scenarios and the well-built clinical question.* Bibliotheca Medica Canadiana 2000; 22(2): 80-84.
2. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vust G, Alderson P, Glasziou P, Falck-Ytter Y, Schunemann HJ. GRADE guidelines: 2. Framing the question and deciding on important outcome. J Clin Epidemiol 2011; 64:395-400.

LA VALUTAZIONE DELLA QUALITÀ DELLE PROVE

La valutazione della qualità delle prove, per ogni singolo outcome e complessivamente, è un passaggio fondamentale nella formulazione delle raccomandazioni perché da questo deriva in modo decisivo la "forza" delle raccomandazioni in quanto si determina fino a che punto si può confidare nel fatto che la stima di un beneficio/danno possa essere usata a favore o contro il raccomandare l'uso di quell'intervento.

Un alto livello di qualità delle prove significa un elevato livello di confidenza nei risultati, in quanto si ritiene che sia molto improbabile che ulteriori studi possano cambiare la stima dell'effetto; una qualità molto bassa, all'estremo opposto, significa che le prove sono inaffidabili e che quindi non è possibile fare affidamento sulle stime di effetto disponibili. In mezzo, un livello di qualità moderata significa un discreto grado di confidenza nei risultati, ma che quindi è probabile che ulteriori studi possano confermare o cambiare le stime dell'effetto, mentre una qualità bassa indica risultati poco credibili e che quindi è necessaria ulteriore ricerca per ottenere stime affidabili.

Nel metodo GRADE, ciò può essere rappresentato mediante la seguente tabella:

Il grading della qualità delle evidenze		
Alta	La stima dell'effetto è **molto probabilmente vicina** all'effetto reale.	⊕⊕⊕⊕
Moderata	La stima dell'effetto è **abbastanza affidabile**; l'effetto reale sembra vicino a quello della stima ma potrebbe anche essere diverso.	⊕⊕⊕◯
Bassa	L'**affidabilità** della stima dell'effetto è **scarsa**: l'effetto reale potrebbe essere sostanzialmente diverso dalla stima.	⊕⊕◯◯
Molto bassa	La stima dell'effetto è **inaffidabile**; è verosimile che l'effetto reale sia sostanzialmente diverso dalla stima.	⊕◯◯◯

È interessante notare come il GRADE sottolinei il diverso significato che il termine qualità assume per una revisione sistematica, per la quale corrisponde al limite di confidenza della stima dell'effetto, e per una raccomandazione di linea guida, dove esprime quanto il limite di confidenza è adeguato a supportare una decisione (effetto reale) (1).

L'aspetto veramente innovativo del GRADE è però legato al fatto che la valutazione della qualità delle prove è basata non solo sull'appropriatezza del disegno dei singoli studi, ma considera altri criteri che possono modificare ampiamente la valutazione delle prove. Nei sistemi tradizionali infatti ad un determinato tipo di studio (RCT, revisione sistematica, ad esempio) corrisponde "automaticamente" un elevato grado di qualità e quindi un'elevata forza delle raccomandazioni. Il GRADE rivede in modo sostanziale questo approccio e valuta sia il tipo di studio che altri elementi che ne determinano la qualità, in base a considerazioni di tipo metodologico.

Queste considerazioni si basano sui seguenti criteri:

- Limiti nella **qualità di conduzione** dei singoli studi.
- **Coerenza** dei risultati tra differenti studi.
- **Diretta applicabilità/rilevanza** dei risultati rispetto ai soggetti/pazienti per i quali deve essere formulata la raccomandazione.
- **Precisione** dei risultati.
- **Pubblicazione selettiva** dei dati.
- **Forza dell'associazione** tra intervento e esito *(outcome).*
- Presenza di un **gradiente dose-risposta.**
- **Direzione** degli effetti dei confondenti plausibili.

Il disegno dello studio, nel GRADE, determina la posizione di partenza della valutazione di qualità: uno studio controllato e randomizzato parte da una posizione alta, uno studio osservazionale parte da una posizione bassa, mentre qualsiasi altro tipo di informazione parte da una posizione molto bassa. Queste posizioni però possono variare anche sostanzialmente per un aumento o una riduzione della categoria in base ai criteri sopra citati e così sintetizzabili:

A. **Diminuzione della categoria di attribuzione** (es. da alta a moderata)	1. Limiti gravi (-1 livello) o molto gravi (-2 livelli) nella qualità di conduzione dello studio. 2. Incoerenza nei risultati tra studi diversi sullo stesso quesito (-1 o -2 livelli). 3. Alcune (-1 livello) o importanti (-2 livelli) incertezze circa la diretta trasferibilità dei risultati (*directness*). 4. Imprecisione o dati insufficienti (-1 o -2 livelli). 5. Possibilità di pubblicazione selettiva dei dati (*publication* e *reporting bias*) (-1 o -2 livelli).
B. **Aumento della categoria di attribuzione** (es. da bassa a moderata)	1. Forte associazione intervento-outcome forte, ovvero con rischio relativo > 2 (<0.5), sulla base di prove concordanti provenienti da due o più studi osservazionali, senza alcun fattore di confondimento plausibile (+1 livello). 2. Associazione intervento-outcome molto forte, ovvero con rischio relativo > 5 (<0.2) (+2 livelli). 3. Presenza di un gradiente dose-risposta (+1 livello). 4. Tutti i possibili fattori di confondimento che avrebbero potuto alterare le stime di effetto avrebbero ridotto l'effetto che si osserva (+1 livello).

Un aspetto importante dell'approccio GRADE è quindi rappresentato dal fatto che sia i trial clinici randomizzati (che partono da una qualità presunta alta) che gli studi osservazionali (che partono da una qualità bassa) possono essere rivisti al ribasso se vi è evidenza di un elevato rischio di bias. Questo rischio non va valutato solo per lo studio nel suo complesso, ma outcome per outcome. Ad esempio, il bias di mancanza di cecità può essere molto rilevante nella valutazione di un outcome misurato "soggettivamente" (ad esempio una scala del dolore o della qualità della vita) ma è molto meno rilevante, se non inesistente, per la valutazione di un outcome quale la mortalità. È inoltre da notare che attualmente la maggior parte delle revisioni sistematiche possono avere un'utilità limitata per il fatto che valutano il bias degli studi invece che il bias dei singoli outcome (2).

Analizzando nello specifico i fattori che possono condurre ad una revisione della qualità degli studi, si devono fare alcune osservazioni.

Limiti nella qualità della conduzione dello studio (-1 o -2 livelli di qualità)

Per gli RCT i punti rilevanti da valutare sono:

- Errori o insufficiente attenzione a una **assegnazione in cieco** al braccio di trattamento o di controllo ***(allocation concealment).***
- Mancanza o problemi nella **effettiva realizzazione della cecità,** specie per outcome soggettivi.
- **Perdita al follow up** di una quota importante di pazienti originariamente inclusi negli studi, o
- **Perdite al follow up asimmetriche** nei due gruppi ***(attrition bias).***

- **Esclusione dall'analisi dei soggetti/pazienti persi al follow up** in diversi momenti dello studio ***(violazione intention to treat).***
- **Interruzioni precoci degli studi** a seguito di un eccesso di efficacia o tossicità, secondo modalità non previste dal protocollo.

Per gli studi osservazionali i principali aspetti sono:

- Mancanza di aggiustamento per i fattori prognostici ***(confondimento).***
- Valutazione degli outcome effettuate in modo differente nei gruppi dello studio ***(detection bias).***
- **Ampie perdite al follow up** o **follow up troppo breve.**
- Negli studi caso-controllo, la probabilità che le informazioni relative all'esposizione siano state raccolte con modalità differenti nei due gruppi ***(information bias).***

"Inconsistency" (Incoerenza, eterogeneità) dei vari studi (-1 livello di qualità) (3):

Questo criterio si applica, per ogni outcome, all'insieme della letteratura disponibile e non al singolo studio. Si manifesta come ampia variabilità delle stime d'effetto (eterogeneità) tra gli studi, senza che vi sia una spiegazione logica (non attribuibile cioè a diversità di popolazione, interventi, esiti o metodo degli studi). Ciò ovviamente aumenta l'incertezza sulla reale entità dell'effetto dell'intervento. La valutazione avviene sui seguenti quattro aspetti:

- La "**stima puntuale**" varia ampiamente tra i diversi studi.
- Gli **intervalli di confidenza** dei diversi studi mostrano una sovrapposizione minima o nessuna sovrapposizione fra loro.
- Il **test statistico per eterogeneità**, che valuta l'ipotesi nulla che tutti gli studi in una metanalisi abbiano la stessa sottostante grandezza di effetto, ha un **basso valore di *p.***
- $\mathbf{I^2}$, che quantifica la proporzione della variazione nelle stime puntuali dovuta alla differenza tra gli studi, **è ampia** *(<40%: basso; 30-60% moderato; 50-90% sostanziale; 75-100% considerevole)*

Nel definire la qualità dell'evidenza per una raccomandazione, l'incoerenza è importante solo quando riduce la fiducia nei risultati in relazione ad una specifica decisione. Nelle figure successive sono riportati alcuni esempi di valutazione dell'inconsistenza.

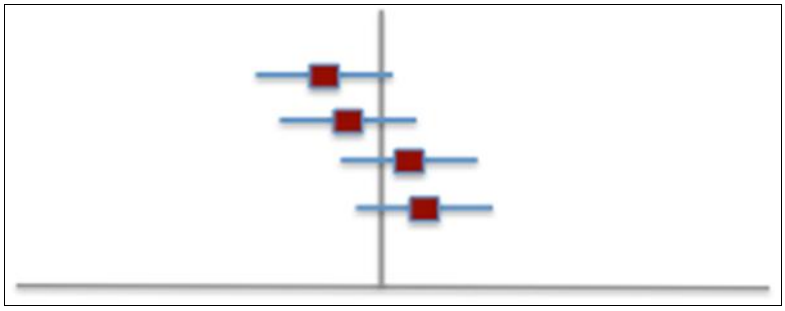


Fig. 1. Differenze nella direzione dell'effetto, ma eterogeneità minima

Fig. 2. Sostanziale eterogeneità, ma di dubbia importanza al fine della raccomandazione

Fig. 3. Sostanziale eterogeneità, di sicura importanza al fine della raccomandazione

In presenza di eterogeneità occorre ovviamente sottolineare la necessità di analizzare le caratteristiche dei diversi sottogruppi per comprendere l'eterogeneità nella composizione degli stessi, al fine, ad esempio di rilevare diversi effetti in gruppi di età o sesso differenti, o in gruppi definiti da altre caratteristiche note dei soggetti.

"Indirectness" (Incertezza sulla diretta applicabilità/rilevanza dei risultati) (-1 o -2 livelli)

La qualità delle prove può essere ridotta se vi è sostanziale differenza per quanto riguarda la popolazione, gli interventi e gli outcome considerati negli studi rispetto a quelli per i quali deve essere formulata la raccomandazione. Altra situazione è quella dove manca il confronto diretto tra gli interventi e ne viene dedotta l'efficacia tramite un confronto indiretto (ad esempio: il confronto A vs B manca e viene dedotto da A vs C e B vs C). (4).

Il criterio dell'incertezza sulla trasferibilità dei risultati (indirectness) si applica all'**insieme della letteratura disponibile** e non al singolo studio.

Quindi, sintetizzando:

- Popolazione, intervento, controllo o esito indiretti: il quesito per il quale si deve fare la raccomandazione si riferisce a una popolazione, intervento, controllo o esito diversi da quelli per cui sono disponibili le prove di efficacia.
- Confronto indiretto: non sono disponibili confronti diretti tra intervento A e intervento B; esistono solo studi che confrontano A con C, e B con C.

A questo proposito è importante ribadire l'importanza del PICO: la definizione compiuta del quesito clinico con le dimensioni PICO è utile infatti anche per valutare la **trasferibilità** più o meno diretta delle prove disponibili. La indirectness dei risultati dipende infatti in larga misura dalla distanza tra gli studi disponibili e il quesito posto dalla raccomandazione che si deve formulare. Se poi si affrontano, all'interno della stessa raccomandazione, quesiti che riguardano sottogruppi, il

giudizio sulla indirectness potrebbe cambiare in funzione di una maggiore o minore focalizzazione del PICO.

Sempre per quanto riguarda l'applicabilità dei risultati degli studi, occorre fare le seguenti precisazioni:

a. Differenze relative alle popolazioni:
    - Popolazioni nelle quali è esclusa la comorbidità.
    - Analisi di sottogruppi di popolazione.
    - Differenze nel rischio di base.
b. Differenze relative agli interventi:
    - Diversi contesti (diverse risorse e competenze).
    - Validità interna (RCT) e esterna (proprio contesto).
    - Implementazione di interventi complessi.
c. Differenze relative agli esiti:
    - Discrepanza tra durata follow-up ipotizzata e reale.
    - Esiti surrogati (ad esempio per quanto riguarda la ventilazione meccanica, l'outcome clinico potrebbe/dovrebbe essere la sopravvivenza o la sopravvivenza senza esiti; un outcome surrogato è lo stato di ossigenazione o la durata della ventilazione).

"Imprecision" (Imprecisione) (-1 livello di qualità) (1)

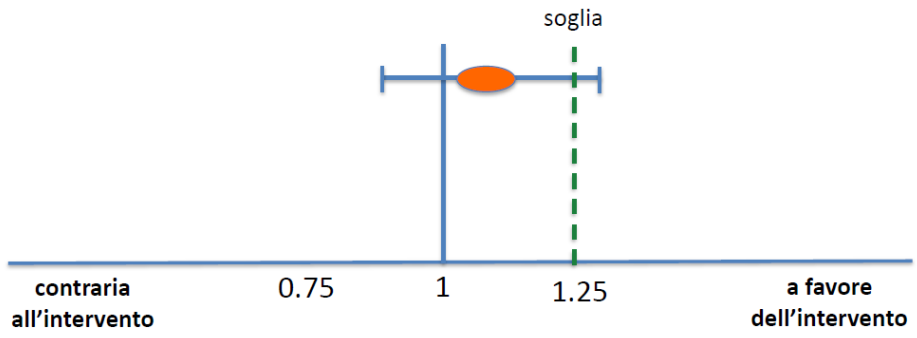
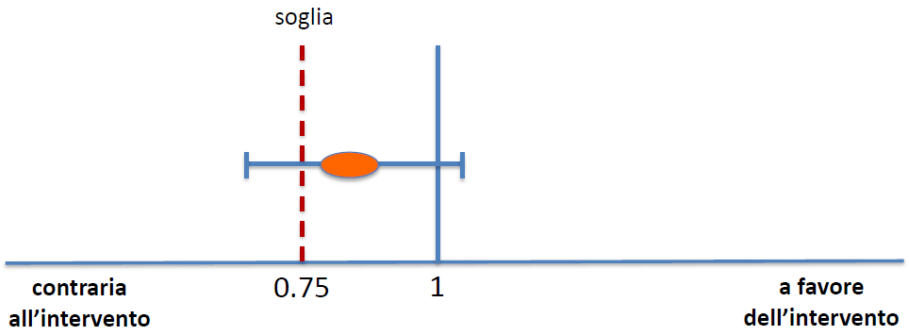
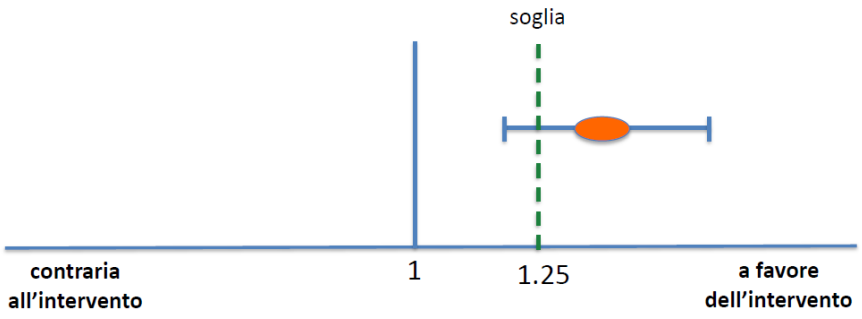
Il criterio principale del GRADE per valutare l'imprecisione è l'attenzione all'intervallo di confidenza (CI) al 95% intorno alla stima aggregata della differenza di effetto tra l'intervento ed il controllo.

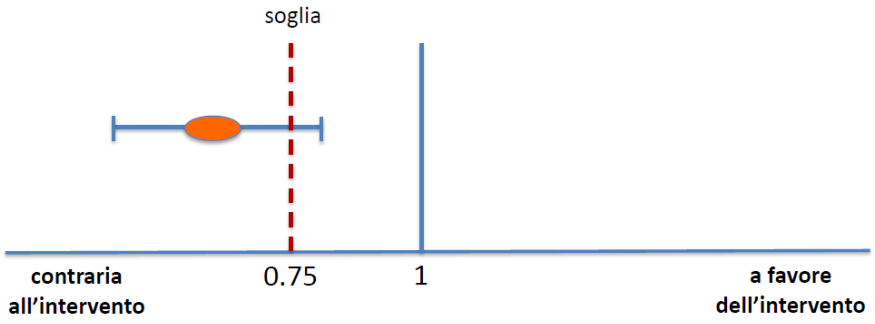
Nell'ottica di una raccomandazione, il panel stabilisce la soglia oltre la quale si può considerare l'effetto clinicamente rilevante *(**"clinical decision threshold"**)*, in termini di beneficio o di danno: la soglia rappresenta il minimo beneficio che si considera rilevante in relazione all'effetto, o il massimo danno che si intende tollerare.

Dal punto di vista pratico, qualora si considerasse il limite inferiore dell'intervallo di confidenza invece che il limite superiore (o viceversa), rispetto alla soglia, e la raccomandazione risultasse diversa, significherebbe che vi è sostanziale imprecisione delle stime e che quindi la qualità delle prove deve essere ridotta.

Da notare che se anche i CI sembrano ragionevolmente stretti, quando gli effetti sono ampi, e sia la dimensione del campione che il numero degli eventi sono modesti, va considerata una riduzione della valutazione della qualità delle prove per imprecisione.

Si considerino a tal proposito i seguenti esempi, relativi ad un esito dicotomico (guarito/non guarito, vivo/deceduto, ecc.) per diversi tipi di interventi:

	Il CI al 95% include anche la totale mancanza di effetto (RR 1, effetto non "statisticamente significativo"). La raccomandazione non può essere chiaramente **a favore** dell'intervento, ma il limite superiore dell'intervallo di confidenza include un effetto che, se fosse reale, rappresenterebbe un beneficio che sarebbe comunque vantaggioso.
	Il CI al 95% include anche la totale mancanza di effetto (RR 1, effetto non "statisticamente significativo"). La raccomandazione non può essere chiaramente **contro** l'intervento, ma il limite inferiore dell'intervallo di confidenza include un effetto che, se fosse reale, rappresenterebbe un danno che sarebbe comunque inaccettabile.
	Il CI al 95% non include la mancanza di effetto (RR>1, effetto "statisticamente significativo"). La raccomandazione è **a favore** dell'intervento, ma il limite inferiore dell'intervallo di confidenza della stima dell'effetto attraversa una soglia sotto la quale l'intervento sarebbe meno consigliabile

	Il CI al 95% non include la mancanza di effetto (RR>1, effetto "statisticamente significativo"). La raccomandazione è **contro** l'intervento, ma il limite superiore dell'intervallo di confidenza della stima dell'effetto attraversa una soglia oltre la quale l'intervento sarebbe meno sconsigliabile

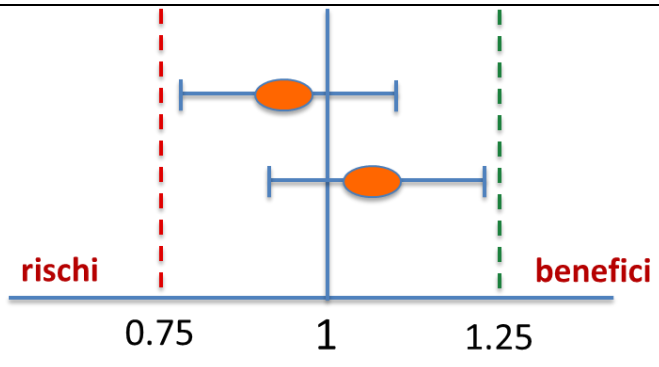
In tutte queste situazioni il rating della qualità deve subire una riduzione (-1 livello) per l'imprecisione delle stime.

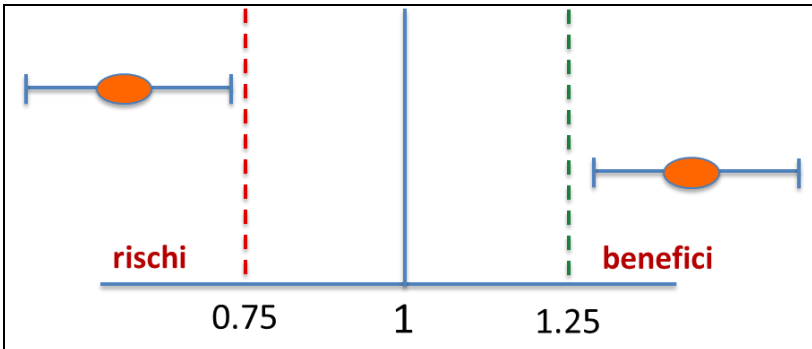
In termini più generali, una stima per un esito dicotomico deve essere considerata imprecisa quando, considerando il limite superiore ed il limite inferiore dei CI al 95%, le decisioni cliniche che ne deriverebbero sarebbero chiaramente diverse.

Per gli esiti continui si può ridurre il livello di qualità delle prove nelle seguenti circostanze:

- Il CI al 95% include la totale mancanza di effetto e il limite superiore o inferiore dell'intervallo di confidenza oltrepassa la differenza minima clinicamente rilevante per i benefici oppure per i danni.
- Nel caso la differenza minima clinicamente rilevante non sia nota (o non sia calcolabile) viene suggerito di abbassare la qualità se il limite superiore o inferiore dell'intervallo di confidenza supera la dimensione dell'effetto di un valore di almeno 0.5, in una direzione o nell'altra.

Il grading **non deve essere ridotto** per imprecisione se:

	L'intervallo di confidenza **include l'effetto nullo e non include** apprezzabili **rischi** o **benefici**

	L'intervallo di confidenza **non include l'effetto nullo e non include** apprezzabili effetti di direzione diversa (**rischi** o **benefici)**

Altra considerazione pratica è legata alla valutazione delle eventuali revisioni sistematiche: se il numero totale dei pazienti inclusi in una revisione sistematica è **inferiore** al numero dei pazienti generato da un convenzionale calcolo della dimensione del campione necessario a definire un'adeguata potenza di un singolo trial ideale, bisogna prendere in considerazione una **riduzione del rating** per **imprecisione. In altre parole** quando gli studi includono pochi pazienti e/o si verificano pochi eventi i risultati devono essere considerati imprecisi a causa di stime con ampi intervalli di confidenza.

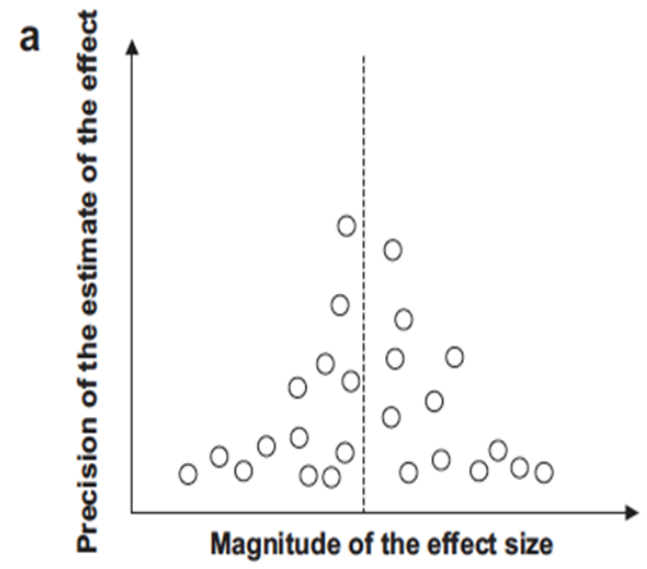
"Publication bias" (Pubblicazione selettiva dei dati) (-1 livello di qualità) (5)

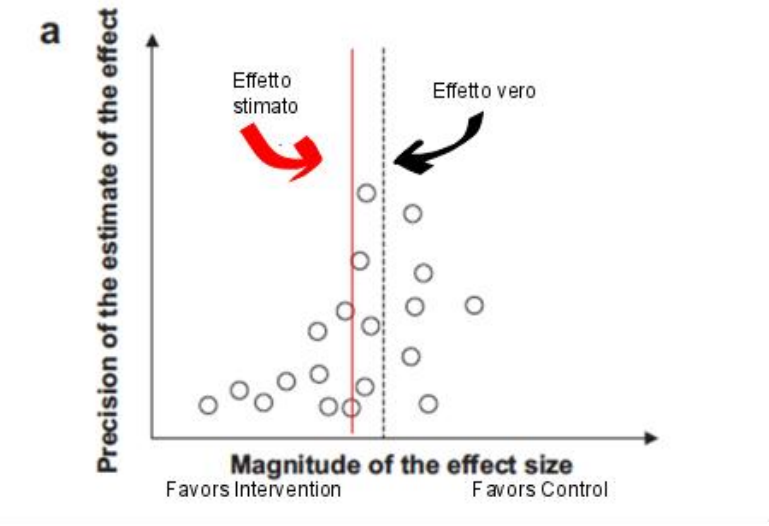
Qualora si verifichi una pubblicazione selettiva di interi studi (***publication bias***) o una pubblicazione selettiva degli outcome (***outcome reporting bias***), gli effetti (positivi o negativi) stimati da uno o più studi possono essere non validi *(sia come sovrastima che come sottostima dell'ipotetico "valore vero").* Esistono metodi statistici per esplorare l'esistenza o meno di tali bias, ma la loro effettiva utilità è oggetto di dibattito anche tra i metodologi. Per quanto riguarda l'approccio GRADE, si raccomanda di considerare e discutere esplicitamente questa possibilità ed esaminare, caso per caso, se in presenza di forte sospetto di publication bias o di outcome reporting bias non sia opportuno procedere a un downgrading della qualità.

L'evidenza empirica dimostra che, in linea generale, studi con effetti statisticamente significativi hanno maggior probabilità di essere pubblicati rispetto a studi con risultati non statisticamente significativi. Inoltre, le revisioni sistematiche eseguite molto precocemente (quando sono disponibili pochi studi) tendono a sovrastimare gli effetti in quanto gli studi "negativi" hanno più difficoltà ad essere pubblicati. Rispetto a ciò, occorre considerare che gli studi "positivi" più precoci, specie se di piccole dimensioni campionarie, devono essere considerati con cautela.

Altra importante considerazione riguarda il fatto che gli studi sponsorizzati dall'industria tendono ad omettere la pubblicazione di risultati "negativi". Si ritiene quindi che gli autori delle revisioni sistematiche dovrebbero, a tal proposito, sospettare il publication bias in presenza di studi piccoli, specie se sponsorizzati dall'industria.

Esiste poi la possibilità di valutare empiricamente i risultati degli studi al fine di identificare il publication bias, ad esempio con i **funnel plot**, che devono però essere interpretati con cautela. A questo proposito si considerino i seguenti esempi, nei quali ogni cerchio rappresenta un RCT:

	Il pattern ideale di distribuzione degli studi assomiglia ad un imbuto (funnel) rovesciato. Gli studi più grandi tendono ad essere più vicini alla linea che rappresenta la stima aggregata dell'effetto. In questo caso, gli effetti degli studi più piccoli sono simmetricamente distribuiti intorno alla stima.

	Questo funnel plot mostra che i piccoli studi non sono simmetricamente distribuiti intorno alla stima aggregata (dominata dagli studi più grandi) né intorno alle stime degli studi più grandi. Gli studi del quadrante in basso a destra mancano (piccoli studi a sfavore dell'intervento). Una possibile spiegazione è il publication bias e una sovrastima dell'effetto rispetto alla realtà.

Il publication bias si genera in diverse fasi della ricerca ed in diversi modi, come riassunto nella seguente tabella:

Fase della ricerca	**Azioni che contribuiscono a produrre bias**
Studi pilota e preliminari	Piccoli studi con risultati negativi tendono a rimanere non pubblicati; l'industria tende a classificarli come di proprietà.
Completamento dei report	Gli autori tendono a considerare non interessanti studi con risultati negativi e non investono tempo e sforzi per pubblicarli.
Selezione delle riviste	Gli autori tendono ad inviare studi con risultati negativi a riviste non indicizzate, in lingue diverse dall'inglese, o a riviste a circolazione limitata.
Considerazioni editoriali	Gli editori tendono a considerare non meritevoli di pubblicazione studi con risultati negativi e rifiutano i manoscritti.
Peer review	I reviewer tendono a concludere che i risultati negativi non sono interessanti e che quindi non meritano di essere pubblicati. La pubblicazione tende comunque ad essere ritardata.

Revisione degli autori e risottomissione degli articoli	Gli autori di studi rifiutati con risultati negativi decidono più frequentemente di non risottomettere il lavoro ad altre riviste.
Pubblicazione dei risultati	Le riviste tendono a ritardare la pubblicazione di studi con risultati negativi.

Aumento della qualità delle prove (6)

Uno degli aspetti più interessanti dell'approccio GRADE è che si può considerare un aumento del livello delle prove, in particolare per gli studi osservazionali ben condotti.

Ciò può avvenire nelle seguenti circostanze:

- In caso di **forte associazione tra intervento ed esito.** Se si osserva un effetto di grandi dimensioni e coerente tra studi diversi, è possibile aumentare il livello di fiducia nella stima dell'effetto. L'aumento si applica solo a studi che non hanno nessun'altra carenza metodologica (nessuna altra diminuzione di livello) (+1 o +2 livelli).
- Presenza di un **gradiente dose-risposta.** Se si osserva un effetto proporzionale alla dose del trattamento è possibile aumentare il livello di fiducia nella stima dell'effetto. L'aumento del livello si applica solo a studi che non hanno nessun'altra carenza metodologica (nessuna altra diminuzione di livello) (+1 livello).
- Se si osserva un effetto **nonostante tutte le possibili distorsioni** (bias) e i potenziali confondenti **vadano nella direzione di diminuire l'entità dell'effetto, è possibile aumentare il livello di fiducia nella stima dell'effetto.** L'aumento del livello si applica solo a studi che non hanno nessun'altra carenza metodologica (nessun'altra diminuzione di livello) (+1 livello).

Da ultimo, alcune considerazioni sulla qualità globale delle prove. Questa va analizzata considerando solo gli outcome essenziali (critici) per la formulazione della raccomandazione relativa al quesito clinico. Qualora la qualità sia diversa fra i singoli outcome essenziali:

- Se i risultati vanno in direzioni opposte *(es. il trattamento oggetto della raccomandazione è migliore in termini di efficacia ma peggiore per quanto riguarda gli effetti indesiderati)*, la qualità globale viene basata sulla valutazione peggiore ossia assumendo come più rappresentativo l'outcome che ha ottenuto la più bassa valutazione di qualità.
- Se i risultati vanno nella stessa direzione per tutti gli outcome (beneficio o danno), viene assunta come qualità globale delle prove la qualità di un singolo outcome essenziale che da solo basterebbe per formulare compiutamente la raccomandazione.

Sintesi dell'approccio GRADE per la valutazione della qualità delle prove (7)

Disegno dello studio	Iniziale qualità delle prove	Più bassa se	Più alta se	Qualità delle prove
RCT	Alta ⟹	**Rischio di bias**	**Effetto ampio**	⊕⊕⊕⊕
		-1 serio -2 molto serio	+1 ampio +2 molto ampio	
		Inconsistenza	**Dose-risposta**	⊕⊕⊕◯
		-1 serio -2 molto serio	+1 evidenza di gradiente	
Studi osservazionali	Bassa ⟹	**Indirectness**	**Possibili confondimenti**	⊕⊕◯◯
		-1 seria -2 molto seria	+1 potrebbero aver ridotto la stima dell'effetto	
		Imprecisione		⊕◯◯◯
		-1 seria -2 molto seria		
		Publication bias		
		-1 probabile -2 molto probabile		

Bibliografia

1. Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines 6. Rating the quality of evidence-imprecision. J Clin Epidemiol. 2011;64:1283-93
2. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). J Clin Epidemiol. 2011; 64:407-15.
3. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. J Clin Epidemiol. 2011;64:1294-302
4. Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol. 2011;64:1303-10
5. Guyatt GH, Oxman AD, Montori V et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64:1277-82
6. Guyatt GH, Oxman AD, Sultan S et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64:1311-6
7. Balshem H, Helfand M, Schünemann HJ et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64:401-6

LA FORMULAZIONE DELLE RACCOMANDAZIONI

L'approccio metodologico proposto dal GRADE definisce anche la modalità di formulazione della raccomandazione da impiegare per rispondere ad un preciso quesito clinico.

La raccomandazione viene espressa, con la relativa forza, in funzione del processo completo previsto dal GRADE ed in relazione all'outcome preso in considerazione. Il GRADE considera l'intero processo della valutazione della qualità delle prove ossia la precisa identificazione del quesito clinico, la scelta degli esiti critici ed importanti, la bontà della stima degli effetti dell'intervento, l'utilizzo delle risorse. Propone infine una tabella di sintesi dei risultati tale da condurre alla stesura della forza della raccomandazione. Negli altri sistemi invece la forza della raccomandazione è stabilita esclusivamente sulla base della qualità degli studi primari. Spesso i clinici considerano come negativa ed eccessivamente rigida tale esclusività.

GRADE prevede che per ciascun obiettivo identificato all'interno di un argomento sia valutata la qualità dell'evidenza dei vari studi per ciascun esito, con particolare attenzione ai più critici, inoltre e soprattutto viene analizzato il rapporto benefici-rischi del determinato intervento prima di poter esprimere la raccomandazione e la sua forza.

La forza di una raccomandazione con GRADE quindi è definita come la misura con cui si può essere certi che le conseguenze desiderabili di un'azione siano maggiori delle conseguenze indesiderabili.

Il GRADE aiuta a definire raccomandazioni forti o deboli, a favore o contro un determinato approccio gestionale nonché a discutere l'interpretazione e la presentazione di queste raccomandazioni.

DIREZIONE della RACCOMANDAZIONE

Indica la direzione in cui si muove la raccomandazione, ossia si parla di raccomandazione **PER** o **A FAVORE** quando gli effetti desiderabili superano gli effetti indesiderabili o di raccomandazione **CONTRO** o **CONTRARIA A** quando gli effetti indesiderabili superano gli effetti desiderabili di una determinata strategia in relazione ad un comparatore o controllo.

Con l'approccio GRADE gli effetti (desiderabili e indesiderabili) di un intervento sperimentale rispetto al controllo sono classificati in "critici" e "importanti ma non critici"

A titolo esemplificativo in tabella sono descritti alcuni degli effetti desiderabili e indesiderabili (1)

Effetti Desiderabili	**Effetti Indesiderabili**
Aumento della longevità	Riduzione delle longevità
Riduzione della morbilità	Complicanze gravi immediate
Riduzione dei sintomi	Effetti collaterali minori a breve termine
	Effetti collaterali rari e gravi a lungo termine
Miglioramento della qualità della vita	Peggioramento della qualità della vita
Riduzione delle risorse usate	Aumento delle risorse impiegate

FORZA della RACCOMANDAZIONE

La forza della raccomandazione, secondo Il metodo GRADE, come per 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.

GRADE stabilisce quattro categorie di raccomandazioni, con una classificazione binaria di raccomandazioni forti o deboli (talvolta identificate come condizionali, discrezionali o qualificate) a favore o contro un determinato trattamento.

Il termine di **raccomandazione forte** è inequivocabilmente comprensibile (gli effetti desiderabili sovrastano gli effetti indesiderabili). Meno intuibile appare l'uso di *raccomandazione debole*. Infatti spesso gli utilizzatori di linee guida attribuiscono una connotazione negativa al termine "debole" come se derivasse da livelli di "evidenza deboli". Pertanto i medesimi utenti potrebbero ignorare le raccomandazioni deboli, o ricondurre la definizione di debole all'incertezza degli sviluppatori nel formulare la raccomandazione adeguata.

GRADE suggerisce tre termini alternativi da utilizzare in caso di *raccomandazione debole*: **condizionale (conditional), discrezionale (discretionary) o qualificato (qualified)** (1)**.**

Le raccomandazioni possono essere **condizionate** dai valori, dalle preferenze del paziente, dalle risorse disponibili o dal contesto in cui l'intervento sarà implementato.

Le raccomandazioni possono essere a **discrezione** del paziente e del medico o **qualificate** da una spiegazione che potrebbe portare a decisioni differenti.

PRESENTAZIONE della RACCOMANDAZIONE

Poiché la raccomandazione in forma passiva potrebbe mancare di chiarezza si preferisce usare la presentazione attiva. Ad esempio:

- Noi raccomandiamo (we recommend …) (forte-strong)
- Noi suggeriamo (we suggest …) (debole-weak)
- I clinici dovrebbero … o non dovrebbero
- Fate o … non fate

Altre forme alternative da suggerire per introdurre le *raccomandazioni deboli* sono:

- I medici potrebbero …
- Si raccomanda condizionalmente …
- Facciamo una raccomandazione qualificata che …

Le raccomandazioni dovrebbero sempre specificare la popolazione a cui si rivolgono (a meno che non sia ovvia) e la strategia di confronto.

In generale poi è preferibile presentare la raccomandazione in favore di un determinato approccio piuttosto che contro. Ad esempio "… si suggerisce l'uso singolo del farmaco A" anziché "… si suggerisce di non aggiungere al farmaco A il farmaco B".

Poiché vi è sempre la possibilità che la terminologia usata sia mal interpretata, GRADE ritiene utile introdurre anche una simbologia grafica e numerica per ovviare a queste incertezze.

Viene suggerito il **simbolo** ↑↑ o il **numero 1** per le **raccomandazioni forti** e il **simbolo** ↑**?** o **il numero 2** per le **raccomandazioni deboli.**

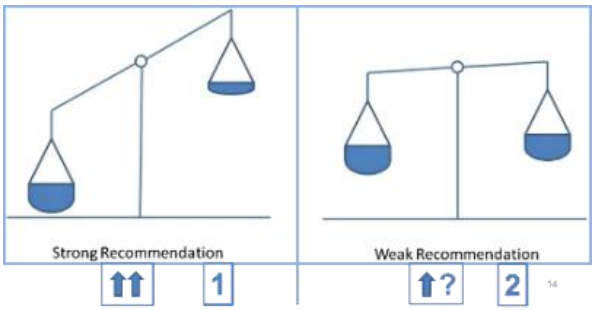


Figura 1. Sintesi di diverse modalità di rappresentazione della forza delle raccomandazioni (1)

I valori e le preferenze sono termini generali che comprendono: prospettive dei pazienti, credenze, aspettative, obiettivi di salute e di vita.

Più precisamente, si riferiscono ai processi impiegati dagli individui per considerare i potenziali benefici, i danni, i costi, le limitazioni e gli inconveniente di una gestione opzionale rispetto un'altra.

Concludendo si può sostenere che una **raccomandazione forte** tiene conto d'implicazioni differenti relativamente alle singole parti coinvolte (pazienti, clinici, decisori sanitari) ma sempre nella direzione del comprovato rapporto di beneficio prodotto dalla sua applicazione (2,3).

Pertanto l'applicazione di una **raccomandazione forte** implica relativamente ai:

- **Pazienti** la considerazione del fatto che la maggior parte delle persone con un determinato problema vorrebbe avere l'intervento raccomandato e solo una piccola parte non lo vorrebbe
- **Clinici** la considerazione del fatto che la maggior parte dei pazienti dovrebbe ricevere l'intervento raccomandato
- **Decisori sanitari** la considerazione del fatto che la raccomandazione può essere adottata nella maggior parte delle situazioni come una indicazione

Di contro una *raccomandazione debole* implica relativamente ai:

- **Pazienti** la considerazione del fatto che la maggior parte delle persone con un determinato problema vorrebbe avere l'intervento, ma molti non lo vorrebbero
- **Clinici** la considerazione del fatto che diverse scelte possono essere appropriate per pazienti diversi e che ogni paziente deve essere supportato nell'optare per scelte conformi ai propri valori e preferenze
- **Decisori sanitari** la considerazione del fatto che il processo decisionale necessita di un dibattito che coinvolga molti stakeholder

SITUAZIONI PARTICOLARI

Esistono inoltre altre 2 situazioni nelle quali non può essere espressa la forza della raccomandazione in modo universale. Si tratta di interventi in corso di ricerca clinica o della scelta di non fornire raccomandazioni.

La raccomandazione con dizione "**Solo-in-ricerca**" può essere espressa nelle situazioni ove:

- Non vi sono prove sufficienti per sostenere l'intervento.
- Ulteriori ricerche potrebbero ridurre l'incertezza degli effetti dell'intervento.
- La ricerca è considerata avere un buon rapporto costi-benefici.

La raccomandazione "Solo-in-ricerca (Only in Research)" può essere accompagnata da una **forte raccomandazione** esplicita a non utilizzare l'intervento sperimentale al di fuori del contesto di ricerca.

Non di rado vi è difficoltà nel fare una raccomandazione a favore o contro una particolare strategia, poiché concludere con "solo-in-ricerca" è inadeguato (1).

Esistono inoltre situazioni in cui non si ritiene possibile formulare raccomandazioni:

- Ove la fiducia nelle stime degli effetti è così bassa da far ritenere una raccomandazione infondata (ad esempio, controllo visivo per lo screening del cancro della pelle USPSTF).
- Quando pur stimando gli effetti moderati o alti, non si è in grado di decidere il senso di una raccomandazione in presenza di compromessi strettamente bilanciati, valori, preferenze e risorse implicate non note o troppo variabili.
- Ove due strategie determinano conseguenze indesiderabili molto diverse.

Per evitare vuoti decisionali, comunque, nelle condizioni sopra elencate, il panel degli esperti, deputato a fornire raccomandazioni, potrebbe decidere di concludere per una *raccomandazione debole*, accompagnata da qualificazione.

Consideriamo ora le situazioni paradigmatiche ove una **raccomandazione forte** può essere giustificata nonostante la bassa o scarsa fiducia nelle stime degli effetti (4). Si verificano:

- Quando la bassa qualità delle evidenze suggerisce un beneficio in caso di pericolo di vita.
- Quando la bassa qualità dell'evidenze suggerisce un effetto benefico e alte qualità depongono per danni o costi elevati, in tal caso si può formulare una raccomandazione contro l'intervento.
- Quando basse qualità di evidenze sono equivalenti in presenza di due strategie, ma elevate qualità di evidenza di minor danno sono a favore di una delle due strategie.
- Quando l'alta qualità delle evidenze suggerisce equivalenza tra due alternative e vi è bassa qualità per i danni di un dei due trattamenti; si formula in tal caso una raccomandazione contro.
- Quando l'alta qualità delle evidenze suggeriscono modesti benefici e basse o molto basse evidenze suggeriscono possibilità di danni gravi.

In sintesi si può concludere che una **raccomandazione forte** si formula quando i vantaggi dell'adesione all'intervento superano gli effetti indesiderabili e quando la maggior parte dei pazienti, se non tutti, beneficiano dell'intervento.

Una *raccomandazione debole* si formula quando c'è evidenza del beneficio, ma meno certezza circa l'equilibrio dei benefici e dei danni, nonché quando approcci alternativi possono essere più efficaci per alcuni pazienti in alcune circostanze. Pertanto l'applicazione al singolo paziente deve essere ponderata.

Bibliografia

1. Andrews J., Guyatt G., Oxman A.D. et al. GRADE Guidelines 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 66 (2013) 719 -725.
2. Brozek J.L., Akl E.A., Compalati E. et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 3 of 3. The GRADE approach to developing recommendations. Allergy 2011; 66: 588–595.
3. Pregno S., Liberati A. Implicazioni di una raccomandazione in funzione della sua forza (adattato da Guyatt 2008c) L'innovazione nell'assistenza e nuovi strumenti di valutazione
4. Andrews J.C., Holger J. Schunemann et al. GRADE Guidelines 15. Going from evidence to recommendations determinants of reccomendation's direction and strength. J Clin Epidemiol. 66 (2013) 726 -735.

VALUTAZIONE DELLE LINEE GUIDA – L'APPROCCIO AGREE

Le linee guida rappresentano uno strumento di lavoro oramai consolidato, fin dalla prima formale definizione dell'Institute of Medicine risale al risale al 1990, quando furono descritte come "raccomandazioni di comportamento clinico, prodotte con metodi sistematici, per aiutare medici e pazienti nel decidere le modalità di assistenza più appropriate in specifiche circostanze cliniche" (1). A fronte della popolarità e del rapido incremento di produzione di linee guida, apparve da subito evidente il problema della metodologia con la quale venivano redatte (2). Al pari di ogni altra produzione scientifica, infatti, la metodologia utilizzata per la redazione delle linee guida risultava essere il maggior determinante della qualità delle stesse e quindi della loro "efficacia" nel produrre esiti positivi riguardo alla qualità dell'assistenza fornita (3,4).

Si rese subito chiaro, quindi che l'assenza di una descrizione metodologica del processo di formulazione e sviluppo delle raccomandazioni non permetteva di definire quel documento come una linea guida (2). Questa semplice osservazione, che può essere considerata una specie di test di screening della qualità di una linea guida, rimane ancora oggi molto valida e permette, anche ad una prima rapida lettura, una valutazione dei requisiti minimi per poter definire se una linea guida può chiamarsi veramente tale.

Si sviluppò in seguito un vivace dibattito che, dalle prime osservazioni critiche sul metodo (5), condusse alla formulazione di requisiti e quindi al disegno di specifici strumenti di valutazione per la valutazione delle linee guida (6). Tra questi, lo strumento AGREE fu uno dei primi e dei più diffusi a livello internazionale (7). Lo strumento originale era strutturato su 23 elementi chiave raggruppati in 6 domini principali (scopi e obiettivi, coinvolgimento delle figure interessate, rigore dello sviluppo, chiarezza della presentazione, applicabilità, indipendenza editoriale). A distanza di alcuni anni, per far fronte alle mutate esigenze dello scenario clinico e metodologico internazionale, venne prodotta una seconda versione (AGREE II), che manteneva la valutazione con 23 elementi chiave in 6 domini, riorganizzati e ridefiniti alla luce dell'esperienza nel frattempo maturata (8). Nel corso degli anni lo strumento AGREE è stato usato in moltissimi ambiti clinici, compreso quello pediatrico (9). La valutazione critica dei vari strumenti di valutazione ne ha fatto uno dei metodi di riconosciuta validità, e ciò rende conto della sua diffusione (10), che ne fa attualmente lo strumento più utilizzato e validato nel panorama metodologico internazionale (11).

Lo strumento AGREE II infatti risulta il metodo più completo ed affidabile per una valutazione complessiva delle linee guida (12), mentre per scopi più specifici possono essere utilizzati altri metodi, quali il GLIA (13), per la valutazione dell'applicabilità delle linee guida, o l'ADAPTE (14) per la valutazione della qualità dei contenuti clinici. Il consorzio ADAPTE, peraltro, utilizza lo strumento AGREE per la valutazione metodologica delle linee guida e lo integra con altri elementi di valutazione utili per affrontare il tema cruciale dell'adattamento a vari livelli delle linee guida. Recentemente, inoltre, uno sviluppo del metodo AGREE ha portato alla formazione di AGREE-REX, strumento di valutazione della validità e dell'applicabilità di singole raccomandazioni o di gruppi di raccomandazioni di linee guida (15), utilizzabile come complemento di AGREE II.

AGREE II

Lo strumento AGREE II rappresenta quindi un utile e valido metodo di valutazione delle linee guida, ampiamente descritto, anche nella sua evoluzione storica e metodologica, sul sito della fondazione che lo ha sviluppato (http://www.agreetrust.org/).

La versione italiana, a cura del gruppo GIMBE, è disponibile sul sito (http://www.gimbe.org/pubblicazioni/traduzioni/AGREE_IT.pdf). La descrizione seguente vuole illustrarne l'impostazione generale e alcune specifiche relative all'utilizzo.

La valutazione complessiva di una linea guida avviene mediante l'analisi di 23 elementi chiave, o item, organizzati in 6 aree (o domini) principali:

1. **Obiettivi e ambiti di applicazione (item 1-3)**
    1) Gli obiettivi generali della linea guida sono descritti in modo specifico
    2) I quesiti sanitari trattati dalla linea guida sono descritti in modo specifico
    3) La popolazione target (pazienti, cittadini, ecc.) a cui applicare la linea guida sono descritti in modo specifico
2. **Coinvolgimento dei soggetti portatori di interesse (stakeholders) (item 4-6)**
    4) Il gruppo di studio che ha elaborato la linea guida include tutte le categorie professionali rilevanti
    5) Sono stati presi in considerazione il punto di vista e le preferenze della popolazione target (pazienti, cittadini, ecc.)
    6) La linea guida identifica con chiarezza gli utenti target
3. **Rigore metodologico (item 7-14)**
    7) Sono stati usati metodi sistematici per ricercare le evidenze scientifiche
    8) La linea guida descrive con chiarezza i criteri utilizzati per selezionare le evidenze scientifiche
    9) La linea guida descrive con chiarezza i punti di forza ed i limiti delle evidenze scientifiche
    10) La linea guida descrive con chiarezza i metodi usati per formulare le raccomandazioni
    11) Nella formulazione delle raccomandazioni sono stati presi in considerazione i benefici ed i rischi conseguenti alla loro applicazione
    12) Esiste un collegamento esplicito tra le raccomandazioni e le evidenze scientifiche che le supportano
    13) Prima della pubblicazione la linea guida è stata valutata da esperti esterni
    14) È descritta la procedura per l'aggiornamento della linea guida
4. **Chiarezza espositiva (item 15-17)**
    15) Le raccomandazioni sono specifiche e non ambigue
    16) La linea guida descrive con chiarezza le diverse opzioni per gestire la condizione clinica o la problematica sanitaria
    17) Le raccomandazioni principali sono facilmente identificabili
5. **Applicabilità (item 18-21)**
    18) La linea guida fornisce suggerimenti e/o strumenti per facilitare l'applicazione delle raccomandazioni
    19) La linea guida descrive i fattori facilitanti e gli ostacoli per l'applicazione delle raccomandazioni
    20) Sono state considerate le potenziali implicazioni sulle risorse conseguenti all'applicazione delle raccomandazioni
    21) La linea guida riporta i principali indicatori per il suo monitoraggio (*audit)*
6. **Indipendenza editoriale (item 22-23)**
    22) I contenuti della linea guida non sono stati influenzati dagli eventuali sponsor istituzionali o commerciali
    23) Gli eventuali conflitti di interesse dei componenti del gruppo che ha elaborato la linea guida sono stati esplicitamente dichiarati e adeguatamente governati

Vi è poi una valutazione finale della linea guida basata sugli elementi sopra descritti e che deriva dal punteggio complessivo assegnato mediante la valutazione. Lo strumento prevede la valutazione di una linea guida da parte di più esperti (almeno due, ottimale tre), che devono effettuare la valutazione indipendentemente per arrivare poi ad una valutazione concordata. Il manuale, disponibile sia in lingua originale (www.agreetrus.org) che in lingua italiana (www.gimbe.org), fornisce elementi specifici per l'applicazione di ogni singolo item.

La validità del processo di sviluppo di una linea guida è un presupposto per minimizzare il bias di una linea guida, ed è valutabile. Gli strumenti di misura della validità delle linee guida possono essere usati sia per valutare le informazioni esistenti che come ausilio didattico per la formulazione o l'adattamento di linee guida.

In ambito neonatologico il gruppo di studio sulla Qualità delle Cure (QCN) della Società Italiana di Neonatologia ha da diversi anni promosso l'utilizzo di questo strumento attraverso specifici corsi di formazione basati sull'analisi delle evidenze scientifiche e sull'utilizzo pratico dello strumento (16-21).

L'adozione sistematica dello strumento e di una procedura definita e condivisa per la valutazione delle linee guida può essere di fondamentale importanza in un momento in cui le società scientifiche vengono chiamate a produrre, valutare e validare linee guida che risultino utili ai professionisti che vi ricorrono.

AGREE REPORTING CHECKLIST

L'AGREE Reporting Checklist ha lo scopo di assistere gli sviluppatori di linee guida per migliorarne la completezza e la trasparenza nella stesura. La checklist può inoltre fornire indicazioni ai revisori, agli editori di riviste scientifiche e agli utenti delle linee guida sui componenti essenziali di una linea guida. La checklist mantiene la stessa struttura di AGREE II di 6 domini di qualità e dei suoi 23 elementi chiave, fornendo un processo logico e sistematico per la valutazione delle informazioni essenziali.

AGREE-REX

Se l'obiettivo di AGREE II è quello di una valutazione della qualità complessiva di una linea guida, l'obiettivo del progetto AGREE-REX è quello di sviluppare una risorsa utile, affidabile e valida per integrare l'AGREE II nel processo di valutazione della credibilità clinica e attuabilità delle singole raccomandazioni o di gruppi di raccomandazioni di linee guida.

Il progetto, ancora in corso ed in fase di validazione, comprende 11 item di valutazione che affrontano 4 concetti chiave (domini) relativi alla credibilità clinica e attuabilità delle raccomandazioni, giustificazione delle prove, applicabilità clinica, giustificazione dei valori dei pazienti, degli operatori e delle parti in causa, applicabilità delle raccomandazioni:

Dominio	Item
1. Giustificazione delle prove	1. Evidenze
2. Applicabilità clinica	2. Rilevanza clinica 3. Rilevanza per i pazienti/per le popolazioni 4. Rilevanza per l'implementazione
3. Giustificazione dei valori	5. Valori degli sviluppatori della linea guida 6. Valori degli utilizzatori target 7. Valori dei pazienti/delle popolazioni 8. Valori delle politiche sanitarie/assistenziali 9. Allineamento dei valori
4. Applicabilità	10. Applicabilità locale 11. Risorse, capacità, strumenti

In base agli elementi sopra riportati sono state sviluppate tre versioni dello strumento, adatte a diversi tipi di utilizzatori:

AGREE-REX Valutazione: sviluppato per valutare la credibilità clinica e la possibilità di implementazione delle raccomandazioni esistenti. Per ogni item gli utilizzatori valutano come è documentata la caratteristica e se la caratteristica è considerata della formulazione delle raccomandazioni.

AGREE-REX Applicazione: Disegnato per determinare se le raccomandazioni sono appropriate per l'adattamento, l'adozione o l'implementazione in un determinato contesto. Per ogni item oltre ai due concetti sopra riportati vi è un terzo concetto relativo all'appropriatezza della documentazione e delle considerazioni delle raccomandazioni per il contesto specifico dell'utilizzatore.

AGREE-REX Sviluppo e reporting: Disegnato per informare gli sviluppatori su cosa considerare nello sviluppo di raccomandazioni e su cosa riportare in un documento di linee guida. Questo strumento può essere utile per una strategia di controllo interno, per manuali di procedure, per corsi metodologici di valutazione o per valutazioni editoriali da parte di riviste.

Nell'intento degli sviluppatori di questi strumenti, un primo passo è la valutazione complessiva della qualità di una linea guida mediante lo strumento AGREE II. Se questa valutazione da esito positivo, il successivo passo è la valutazione di singole raccomandazioni o gruppi di raccomandazioni al fine di poterle adottare, modificare o implementare direttamente.

Gli strumenti AGREE II e AGREE Reporting Checklist sono utili per valutare nel loro complesso il processo di sviluppo delle linee guida ed il rigore metodologico di sviluppo e reporting (22).

AGREE-REX si concentra invece sulla valutazione dello sviluppo e della formulazione di raccomandazioni clinicamente credibili ed utilizzabili. Gli strumenti sono complementari l'uno all'altro e possono essere usati contemporaneamente. Una strategia efficiente può comportare una prima valutazione tramite AGREE II o AGREE Reporting Checklist al fine di valutare la qualità metodologica globale di una linea guida; se la linea guida soddisfa gli standard metodologici minimi, si può quindi utilizzare l'AGREE-REX per valutare la credibilità clinica e la implementazione delle singole raccomandazioni o di gruppi di raccomandazioni della linea guida.

Bibliografia

1. Committee to Advise the Public Health Service on Clinical Practice Guidelines IoM. Clinical practice guidelines: direction s for a new program. Washington: National Academy Press; 1990.
2. Feder G, Grimshaw J et al. Using clinical guidelines. BMJ 1999;318:28
3. Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet 2000;355:103-106
4. Shaneyfelt et al. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA 1999;281:1900
5. Cluzeau et al. Development and application of a generic methodology to assess the quality of clinical guidelines. International Journal for Quality in Health Care 1999;11:21
6. Graham ID, Calder LA et al. A comparison of clinical practice guideline appraisal instruments. International Journal for Technology Assessment in Health Care 2000;16:1024
7. Cluzeau F, Burgers J., Brouwers M., Grol R, Makela M, Littlejohns P, Grimshaw J, Hunt C.. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Quality & Safety in Healthcare 2003;12;18-23
8. Brouwers M, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, Fervers B, Graham ID, Grimshaw J, Hanna S, Littlejohns P, Makarski J, Zitzelsberger L for the AGREE Next Steps Consortium. AGREE II: Advancing guideline development, reporting and evaluation in healthcare. *Can Med Assoc J. 2010;182:E839-842*
9. Boluyt N, Lincke CR, Offringa M. Quality of evidence-based pediatric guidelines. Pediatrics. 2005;115:1378-91
10. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care. 2005;17:235-42
11. Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L.The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. J Evid Based Med. 2015;8:2-10
12. Siering U, Eikermann M, Hausner E, Hoffmann-Eßer W, Neugebauer EA. Appraisal tools for clinical practice guidelines: a systematic review. PLoS One. 2013 9;8:e82915
13. Kashyap N, Dixon J, Michel G, Brandt C, Shiffman RN (2011). Glia: Guideline Implementability Appraisal V. Available: http://gem.med.yale.edu/glia/doc/GLIA_v2.pdf Accessed 28 January 2017
14. ADAPTE Collaboration (2010) Guideline adaption: a resource toolkit; version 2.0. Available: http://www.g-i-n.net/document-store/adapte-resource-toolkit-guideline-adaptation-version-2. Accessed 28 January 2017
15. http://www.agreetrust.org/agree-research-projects/agree-rex-recommendation-excellence/ . Accesso 10/2/2017
16. Bellù R, Zanini R. Sono valide le linee guida che usiamo in Reparto? Riv Ital Pediatr 2001; 27:645-650
17. Bellù R, Zanini R. Linee guida in ambito neonatologico: valutazione critica di alcuni documenti per la pratica clinica. Prospettive in Pediatria 2005; 35:265-270
18. Bellù R, Longhi R, Santucci S, Zanini R. Il progetto "Linee guida" del GSAQ e del QCN. Ospedale & Territorio 2003; 5:43
19. Bellù R, Massironi E, Del Prete A, Meschi V, Bettinelli A. Applicazione di raccomandazioni per la pratica clinica in ambito neonatologico: l'identificazione ed il trattamento della sepsi precoce. Prospettive in Pediatria 2007; 37:159-163
20. "Caratteristiche e significato delle linee guida". Corso di formazione "Linee guida e qualità delle cure in pediatria". Roma 26-28 giugno 2003
21. Manuale metodologico sulle linee guida. Workshop "La valutazione delle linee guida in ambito pediatrico e neonatologico", Varenna, 30-31 marzo 2006
22. Brouwers MC, Kerkvliet K, Spithoff K, AGREE Next Steps Consortium. The AGREE Reporting Checklist: a tool to improve reporting of clinical practice guidelines. BMJ 2016; 352:i1152.

CONCLUSIONI

Dopo molti anni dalla comparsa delle linee guida sulla scena scientifica internazionale, il metodo GRADE rappresenta un significativo passo in avanti nel coniugare rigore metodologico e rilevanza clinica delle raccomandazioni; su questo metodo si è registrata la convergenza dei principali enti internazionali e nazionali nonché di numerose società scientifiche interessate alla produzione e alla valutazione delle linee guida; la complementarietà del metodo AGREE II rende completo il pannello degli strumenti clinici e metodologici indispensabili ai soggetti interessati alle attività di produzione, valutazione ed adozione di linee guida.

I seguenti riferimenti esterni possono contribuire a rendere disponibile il materiale necessario per queste attività:

https://gradepro.org/	GRADEpro e GRADEpro Guideline Development Tool: soluzioni web complete e facili da usare per stesura, la sintesi e la presentazione di raccomandazioni e linee guida. Comprendono anche il manuale GRADE.
http://www.agreetrust.org/	Materiale completo AGREE (aggiornato alla versione II) per la valutazione di linee guida e raccomandazioni
http://www.gimbe.org/pagine/569/it/agree-ii	Grigia AGREE II e AGREE Reporting Checklist, traduzione italiana

Il Gruppo di Studio Qualità delle Cure della Società Italiana di Neonatologia propone questi strumenti a tutta la comunità scientifica neonatologica italiana, proponendosi come riferimento e supporto metodologico nelle attività inerenti le linee guida.

APPENDICE

Il grading della qualità delle evidenze		
Alta	La stima dell'effetto è **molto probabilmente vicina** all'effetto reale	⊕⊕⊕⊕
Moderata	La stima dell'effetto è **abbastanza affidabile**; l'effetto reale sembra vicino a quello della stima ma potrebbe anche essere diverso	⊕⊕⊕◯
Bassa	L'**affidabilità** della stima dell'effetto è **scarsa**: l'effetto reale potrebbe essere sostanzialmente diverso dalla stima	⊕⊕◯◯
Molto bassa	La stima dell'effetto è **inaffidabile**; è verosimile che l'effetto reale sia sostanzialmente diverso dalla stima	⊕◯◯◯

Tab. 1 GRADE: Grading della qualità delle evidenze.

Diminuzione della categoria di attribuzione (es. da alta a moderata)	1. Limiti gravi (-1 livello) o molto gravi (-2 livelli) nella qualità di conduzione dello studio 2. Incoerenza nei risultati tra studi diversi sullo stesso quesito (-1 o -2 livelli) 3. Alcune (-1 livello) o importanti (-2 livelli) incertezze circa la diretta trasferibilità dei risultati (*directness*) 4. Imprecisione o dati insufficienti (-1 o -2 livelli) 5. Possibilità di pubblicazione selettiva dei dati (*publication* e *reporting bias*) (-1 o -2 livelli)
Aumento della categoria di attribuzione (es. da bassa a moderata)	1. Forte associazione intervento-outcome forte, ovvero con rischio relativo > 2 (<0.5), sulla base di prove concordanti provenienti da due o più studi osservazionali, senza alcun fattore di confondimento plausibile (+1 livello) 2. Associazione intervento-outcome molto forte, ovvero con rischio relativo > 5 (<0.2) (+2 livelli) 3. Presenza di un gradiente dose-risposta (+1 livello) 4. Tutti i possibili fattori di confondimento che avrebbero potuto alterare le stime di effetto avrebbero ridotto l'effetto che si osserva (+1 livello)

Tab. 2 GRADE: Criteri che modificano (diminuzione o incremento) la qualità delle evidenze.

Forza di raccomandazione = grado di sicurezza che gli effetti desiderabili di un intervento superino gli effetti indesiderabili

Raccomandazione forte = si è sicuri che aderendo alla raccomandazione gli effetti desiderabili superino quelli indesiderabili
Raccomandazione debole = gli effetti desiderabili probabilmente superano quelli indesiderabili

Fattori chiave:
1. Bilancio tra conseguenze desiderabile e non desiderabili delle strategie alternative
2. Qualità dell'evidenza
3. Incertezza circa i valori e le preferenze che darebbero i vari soggetti alle varie alternative diminuisce la forza della raccomandazione
4. Costi

Tab. 3 GRADE: Significato della Raccomandazione e della relativa forza.

Disegno dello studio	Iniziale qualità delle prove	Più bassa se	Più alta se	Qualità delle prove
RCT	Alta ⟹	**Rischio di bias**	**Effetto ampio**	⊕⊕⊕⊕
		-1 serio -2 molto serio	+1 ampio +2 molto ampio	
		Inconsistenza	**Dose-risposta**	⊕⊕⊕◯
		-1 serio -2 molto serio	+1 evidenza di gradiente	
Studi osservazionali	Bassa ⟹	**Indirectness**	**Possibili confondimenti**	⊕⊕◯◯
		-1 seria -2 molto seria	+1 potrebbero aver ridotto la stima dell'effetto	
		Imprecisione		⊕◯◯◯
		-1 seria -2 molto seria		
		Publication bias		
		-1 probabile -2 molto probabile		

Tab. 4 GRADE: Sintesi dell'approccio GRADE per la valutazione della qualità delle prove**.**