GRADE: una nueva propuesta para clasificar la calidad de la evidencia y graduar la fuerza de las recomendaciones

Ivan Solà
Centro Cochrane Iberoamericano
Institut d'Investigació Biomèdica Sant Pau (IIB Sant Pau)

Grupo de Trabajo de Pediatría Basada en la Evidencia
Reunión Grupo de Trabajo – 1 abril 2011

Centro Cochrane Iberoamericano
Iberoamerican Cochrane Centre

¿No existen demasiados sistemas?

Las propuestas disponibles son inconsistentes en la manera de clasificar la calidad de la evidencia y graduar la fuerza de las recomendaciones.

Los usuarios de guías a veces tienen problemas para entender lo que quieren comunicar algunos sistemas.

GRADE intenta ofrecer un sistema **estructurado, transparente y explícito** para formular recomendaciones.

¿Cuáles son las ventajas del sistema GRADE?

– Desarrollado por un **grupo multidisciplinar y representativo** en el campo de las guías de práctica clínica

– Hace una **clasificación explícita de la importancia de las variables** de resultado de interés

– Diferencia claramente la **calidad de la evidencia** de la **fuerza de las recomendaciones**

– Los criterios para disminuir o aumentar la calidad de la evidencia son **explícitos** y exhaustivos

– El paso de la calidad de la evidencia a la fuerza de la recomendación es **transparente**

– Se reconoce explícitamente la importancia de los **valores y preferencias de los pacientes** en la formulación de recomendaciones

– Propone un **interpretación clara y pragmática** de la fuerza de las recomendaciones (solamente en fuertes o débiles)

– Útil tanto para guías como para revisiones sistemáticas, informes de evaluación de tecnologías, …

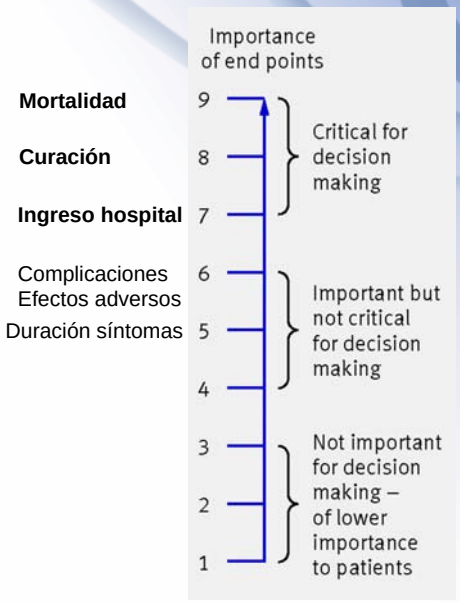
Primeras etapas…

Parte de preguntas clínicas estructuradas (formato PICO)

Las revisiones sistemáticas se usan como fuente de estudios originales, y no como un diseño de estudio al que corresponde un grado de calidad de la evidencia

Cualquier decisión pasa por establecer la importancia relativa de las variables de resultado de interés, diferenciando aquellas clave para la toma de decisiones

¿En pacientes adultos con NAC tratados de manera ambulatoria, qué tratamiento empírico es más adecuado?



¿Qué es la ***calidad de la evidencia*** y por qué es importante?

La calidad de la evidencia se refiere a cuánta ***confianza se puede tener en que la estimación de la magnitud del efecto*** para una intervención es adecuada para apoyar las recomendaciones.

Guyatt et al. BMJ 2008;336:995-8

¿Cómo clasifica la calidad de la evidencia el sistema GRADE?

El sistema GRADE propone una clasificación simple y clara de la calidad de la evidencia en cuatro niveles:

- ***Calidad alta*** — Es difícil que los resultados de nuevos estudios modifiquen la confianza en la estimación del efecto
- ***Calidad moderada*** — la confianza en la estimación del efecto y su magnitud podrían cambiar con nuevos estudios
- ***Calidad baja*** — es probable nuevos estudios modifiquen la confianza en la estimación del efecto y su magnitud
- ***Calidad muy baja*** — cualquier estamación del efecto es muy incierta

Los Grupos de Trabajo deberían realizar la clasificación de la calidad de la evidencia en relación con el contexto clínico en el que se desarrolla la guía

Guyatt et al. BMJ 2008;336:995-8

¿Cómo clasifica la calidad de la evidencia el sistema GRADE? (II)

Se considera que los ensayos clínicos tienen una calidad alta, pero una serie de factores pueden reducir la confianza en sus resultados :

- Limitaciones en el diseño
- Inconsistencia de los resultados
- Evidencia no directa
- Imprecisión
- Publication bias

Guyatt et al. BMJ 2008;336:995-8

Limitaciones en el diseño

La confianza en los resultados de los ensayos clínicos disminuye si se detectan limitaciones claras en su diseño, ya que el riesgo de sesgo será alto.

Las principales fuentes de sesgo en un ensayo clínico son:

- una secuencia de aleatorización mal diseñada,
- ausencia de encubrimiento de la secuencia de aleatorización,
- ausencia de cegamiento (importante en variables subjetivas),
- pérdidas considerables en el seguimiento,
- ausencia de un análisis por intención de tratar,
- ensayos interrumpidos prematuramente por beneficio,
- descripción selectiva de los resultados.

Guyatt et al. BMJ 2008;336:995-8

Empiric antibiotic coverage of atypical pathogens for community acquired pneumonia in hospitalized adults (Review)

Robenshtok E, Shefet D, Gafter-Gvili A, Paul M, Vidal L, Leibovici L

THE COCHRANE COLLABORATION®

Risk of bias in included studies

General

All studies have been fully published. One study was a phase II trial (Feldman 2001), and another was a phase III trial (Lephonte 2004). At least 21 of the 25 studies were sponsored by pharmaceutical companies, all but one manufactured the atypical drug. Eighteen studies reported patient consent and 18 reported approval of a local ethics committee.

Intention-to-treat (ITT) versus per-protocol analysis

We separated the three different study types by considering outcome reporting:

1. Studies performed by intention-to-treat (ITT).
2. Per-protocol studies, in which the number of dropouts was given per study arm.
3. Per-protocol studies, in which the number of dropouts was reported or could be calculated, but not given per study arm.

As mentioned above, the primary outcome in all studies was clinical success. Seven studies reported results by ITT (type 1). Fourteen additional studies reported the number of dropouts per study arm (type 2), permitting re-analysis by ITT assuming failure for all dropouts. Four studies did not refer to dropouts (type 3) and were analyzed by evaluated patients only in the sensitivity analysis. Mortality was not a primary outcome and was usually reported in the safety analysis. Thirteen studies recounted information regarding overall mortality by ITT (type 1), while 11 provided data per-protocol. One study does not specifically mention or rule out deaths (Kobayashi 1984).

Allocation

Adequate allocation concealment was reported in one quarter of the studies (6 out of 25). No sufficient information was available for the other studies. Allocation generation was adequate in 9 out of 25 studies. No information was available for 16 studies. All studies of adequate allocation concealment were also of adequate allocation generation.

Blinding

Ten studies were double blind, one was single blinded and the remaining (14 out of 25) were open label.

Efficacy of Exclusively Oral Antibiotic Therapy in Patients Hospitalized with Nonsevere Community-Acquired Pneumonia: A Retrospective Study and Meta-analysis

Am J Med. 2004;116:385–393.

Theodore K. Marras, MD, MSc, Cherdchai Nopmaneejumruslers, MD, Charles K. N. Chan, MD

APPENDIX. Methodologic Quality Rating Scales for Trials of Oral versus Intravenous Antibiotic Therapy in Community-Acquired Pneumonia

Rating Scale	First Author (Reference)						
	Fredlund (14)	Zuck (16)	Vogel (17)	Bohte (18)	Chan (19)	Castro-Guardiola (20)	Lode (15)
Truly randomized*	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Intention to treat—outcomes of patients who withdrew described and included in analysis†	0.5	0	0.5	0.5	1	1	0.5
Treatment and control groups comparable at entry‡	0.5	0.5	0.5	0.5	1	1	1
Patients, treatment providers, and outcome assessors blinded to treatment group§	0	0	0	0	0	0	0
Care programs, other than trial options, identical‖	0	0	0	0	0	0	0
Inclusion and exclusion criteria clearly defined¶	0.5	0.5	1	0.5	1	1	0.5
Interventions clearly defined and applied via standardized protocol#	1	1	1	1	1	1	1
Outcomes clearly defined**	1	1	1	1	1	1	0.5
Outcomes include clinical success, mortality, and length of stay††	1	0.67	0.67	0.33	1	1	0.67
Follow-up active and of appropriate duration (>3 weeks)‡‡	1	1	0.5	1	0	1	1
Total (maximum score = 10)	6	5.17	5.67	5.33	6.5	7.5	5.67

Inconsistencia

Grandes diferencias en la estimación del efecto entre los estudios que responden a la misma pregunta (variabilidad en los resultados o ***heterogeneidad***), sugieren que existen diferencias reales en estas estimaciones

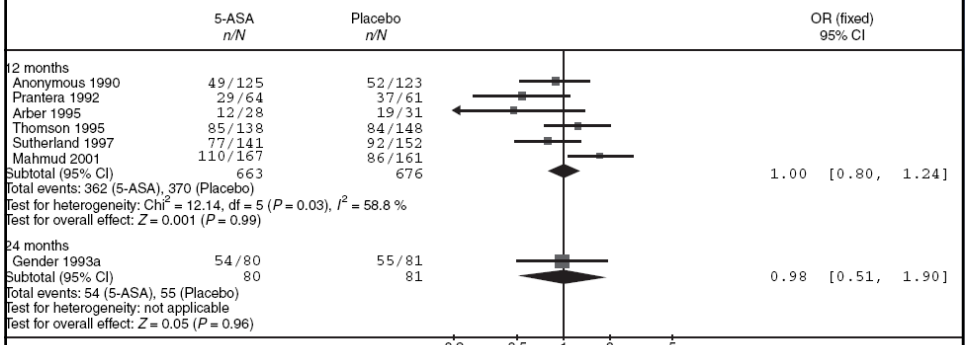
Las diferencias pueden explicarse por:

- población de interés (fármacos con más efecto en poblaciones más graves)
- intervenciones (mayor efecto a mayor dosis)
- variables de resultado (menor efecto cuando el seguimiento es mayor)

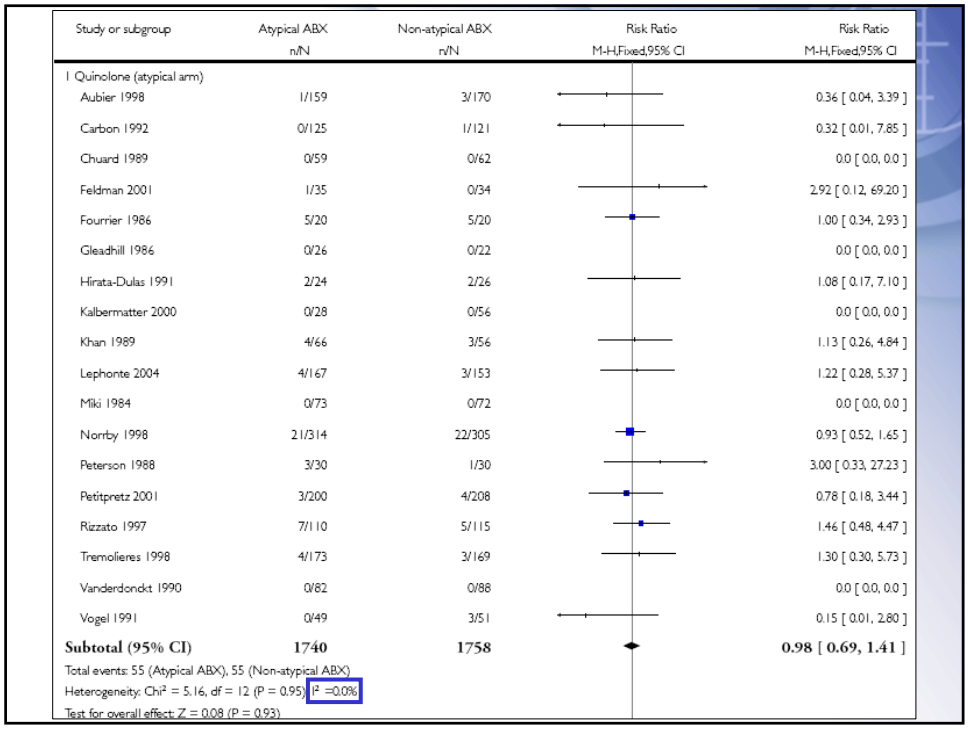
Si no se identifica una explicación para estas diferencias, la confianza en los resultados disminuye

Guyatt et al. BMJ 2008;336:995-8

Oral 5-ASA versus placebo, Outcome = Relapse



Akobeng. Aliment Pharmacol Ther 2008;27:11-18



Evidencia indirecta

Dos razones fundamentales:

Comparaciones indirectas: no se dispone de una comparación de A vs B, pero disponemos de ensayos que comparan A y B vs placebo

Diferencias entre las poblaciones, intervenciones o variables definidas en nuestras preguntas y las disponibles en la literatura relevante.

Guyatt et al. BMJ 2008;336:995-8

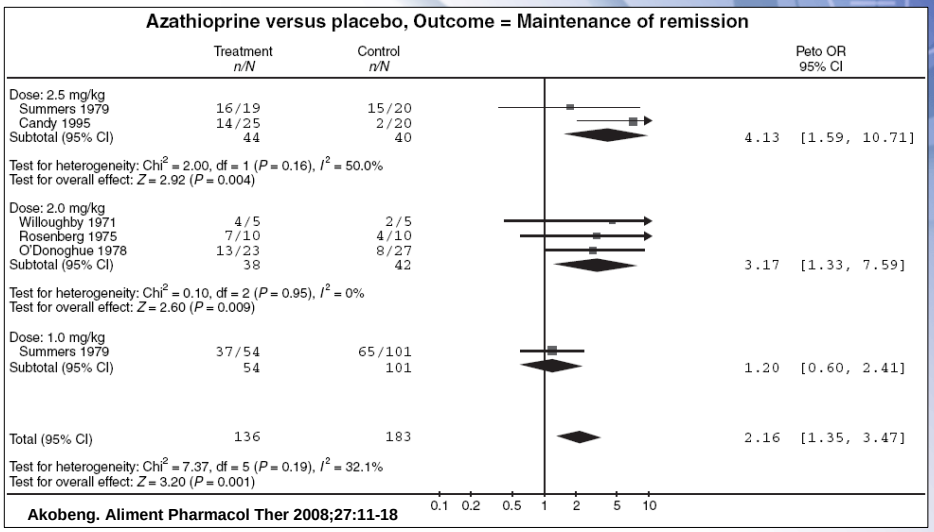
Table 1. Methods and Interventions in Studies of Oral versus Parenteral Therapy in Hospitalized Patients with Community-Acquired Pneumonia*

First Author (Reference)	Inclusion Criteria	Exclusion Criteria	Interventions†	Last Follow-up
Fredlund (14)	Suspected pneumococcal pneumonia	Previous antibiotic use, antibiotic resistance, intensive care unit management, gastrointestinal disturbance	Phenoxymethylpenicillin (6 g, oral) vs. benzylpenicillin (9 g, intravenous)	60 days
Zuck (16)	Age >64 years, ≥1 comorbid disease	Antibiotic resistance	Cefpodoxime proxetil (400 mg, oral) vs. ceftriaxone (1 g, intramuscular)	30 days
Vogel (17)	Age >18 years	Previous antibiotic use, comorbid disease, intensive care unit management	Temafloxacin (1200 mg, oral) vs. cefotaxime (6 g, intravenous)	10 days after last dose
Bohte (18)	Age >18 years	Age >75 years, previous antibiotic use, antibiotic resistance, hospitalized in past week, nursing home residence, intensive care unit management, gastrointestinal disturbance	Azithromycin (1000 mg, then 500 mg, oral) or erythromycin vs. benzylpenicillin (4×10^6 units, intravenous)	21 days after discharge
Chan (19)	Age >14 years	Immunocompromised, intensive care unit management, gastrointestinal disturbance	Co-amoxiclav (1125 mg, oral) vs. co-amoxiclav (3.6 g, intravenous) or cefotaxime (2 g, intravenous)	Discharge
Castro-Guardiola (20)	Age >18 years	Immunodeficiency, aspiration pneumonia, hospitalized in past week, gastrointestinal disturbance	Cefuroxime (1500 mg, oral) or co-amoxiclav (1125 mg, oral) vs. cefonicid (2 g, intravenous), cefuroxime (4.5 g, intravenous), or co-amoxiclav (3 g, intravenous)	30 days
Lode (15)	Age ≥18 years	Previous antibiotic use, hospitalized in past 2 weeks, obstructive or aspiration pneumonia, HIV with CD4 count $<200 \times 10^6$/L, intolerance of oral therapy	Gemifloxacin (320 mg, oral) or ceftriaxone (2 g, intravenous) ± macrolide	28 days after last dose

Imprecisión

Ensayos con ***pocos pacientes*** y ***pocos eventos*** tendrán intervalos de confianza más amplios, y la confianza en sus estimaciones del efecto disminuirá

Guyatt et al. BMJ 2008;336:995-8



Akobeng. Aliment Pharmacol Ther 2008;27:11-18

Analysis 1.5. Comparison 1 Atypical versus non-aytpical, Outcome 5 Mortality per allocation concealment.

Review: Empiric antibiotic coverage of atypical pathogens for community acquired pneumonia in hospitalized adults

Comparison: 1 Atypical versus non-aytpical

Outcome: 5 Mortality per allocation concealment

Study or subgroup	Atypical ABx n/N	Non-atypical ABx n/N	Risk Ratio M-H,Fixed,95% CI	Risk Ratio M-H,Fixed,95% CI
1 A				
Feldman 2001	1/35	0/34		2.92 [0.12, 69.20]
Genne 1997	2/56	1/56		2.00 [0.19, 21.43]
Khan 1989	4/66	3/56		1.13 [0.26, 4.84]
Norrby 1998	21/314	22/305		0.93 [0.52, 1.65]
Peterson 1988	3/30	1/30		3.00 [0.33, 27.23]
Petitpretz 2001	3/200	4/208		0.78 [0.18, 3.44]
Subtotal (95% CI)	**701**	**689**		**1.06 [0.66, 1.69]**

Total events: 34 (Atypical ABx), 31 (Non-atypical ABx)

Heterogeneity: Chi² = 1.90, df = 5 (P = 0.86); I² =0.0%

Test for overall effect: Z = 0.24 (P = 0.81)

Outcome: 8 Clinical failure per antibiotic (ABx) Treatment

Study or subgroup	Atypical ABx n/N	Non-atypical ABx n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
1 Quinolone (atypical arm)					
Aubier 1998	14/126	22/140		4.1 %	0.71 [0.38, 1.32]
Carbon 1992	19/123	24/120		4.7 %	0.77 [0.45, 1.33]
Chuard 1989	5/59	6/62		1.1 %	0.88 [0.28, 2.72]
Feldman 2001	3/35	3/34		0.6 %	0.97 [0.21, 4.48]
Fournier 1986	6/20	7/20		1.4 %	0.86 [0.35, 2.10]
Gleadhill 1986	5/26	4/22		0.8 %	1.06 [0.32, 3.46]
Hirata-Dulas 1991	12/24	12/26		2.2 %	1.08 [0.61, 1.93]
Kalbermatter 2000	1/28	4/56		0.5 %	0.50 [0.06, 4.27]
Khan 1989	6/66	6/56		1.3 %	0.85 [0.29, 2.48]
Kobayashi 1984	4/41	0/33		0.1 %	7.29 [0.41, 130.64]
Lephonte 2004	38/167	32/153		6.5 %	1.09 [0.72, 1.65]
Miki 1984	11/71	7/68		1.4 %	1.51 [0.62, 3.66]
Norrby 1998	85/314	85/305		16.8 %	0.97 [0.75, 1.25]
Peterson 1988	7/30	9/30		1.8 %	0.78 [0.33, 1.82]
Petitpretz 2001	46/200	44/208		8.4 %	1.09 [0.75, 1.57]
Rizzato 1997	7/110	28/115		5.3 %	0.26 [0.12, 0.57]
Tremolieres 1998	13/138	28/147		5.3 %	0.49 [0.27, 0.92]
Vanderdonckt 1990	21/75	21/81		3.9 %	1.08 [0.64, 1.81]
Vogel 1991	2/38	3/37		0.6 %	0.65 [0.11, 3.67]
Subtotal (95% CI)	**1691**	**1713**		**66.9 %**	**0.89 [0.77, 1.02]**

Total events: 305 (Atypical ABx), 345 (Non-atypical ABx)

Heterogeneity: Chi² = 21.13, df = 18 (P = 0.27); I² =15%

Sesgo de publicación

La calidad de la evidencia disminuye si se sospecha o se comprueba que no se han publicado todos los estudios que se han realizado para contestar una pregunta

Guyatt et al. BMJ 2008;336:995-8

¿Cómo clasifica la calidad de la evidencia el sistema GRADE? (III)

Se considera que los estudios observacionales tienen una calidad de la evidencia baja, pero algunos aspectos pueden incrementar la confianza en sus resultados:

- **Magnitud del efecto grande** (los cascos de bicicleta reducen el riesgo de traumatismo craneal (OR 0.31; 95%CI 0.26 to 0.37))
- **Relación dosis respuesta** (exposición al humo del tabaco y cáncer de pulmón)

Guyatt et al. BMJ 2008;336:995-8

Las variables de resultado clave determinan la calidad de la evidencia para responder a la pregunta

El sistema GRADE propone clasificar la calidad de la evidencia variable por variable a partir de unas tablas resumen

Guyatt et al. BMJ 2008;336:995-8

Table 1 | 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 by Karanicolas et al[11]

	Quality assessment					Summary of findings			
No of studies (No of participants)	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

*Unclear allocation concealment in all studies, patients blinded in only one study, outcome assessors not blinded in any study, >20% loss to follow-up in three studies, not analysed using intention to treat in one study.
†Relative risks (95% confidence intervals) are based on random effect models.
‡Confidence interval includes possible benefit from both surgical approaches.
§I2=72.6%, P=0.006.

Guyatt et al. BMJ 2008;336:995-8

TABLE GRADE evaluation of interventions for autism.

Important outcomes	Global improvement, social function, behavioural function, cognitive function, and adverse effects								
Number of studies (participants)	Outcome	Comparison	Type of evidence	Quality	Consistency	Directness	Effect size	GRADE	Comment
What are the effects of early intensive multidisciplinary intervention programmes in children with autism?									
11 (397) [19]	Cognitive function (communication and IQ scores)	Early intensive behavioural interventions v other therapy	2	−2	−1	0	0	Very low	Quality points deducted for uncertain follow-up and for comparison of means. Consistency point deducted for different comparisons
11 (397) [19]	Behavioural function	Early intensive behavioural interventions v other therapy	2	−2	−1	0	0	Very low	Quality points deducted for uncertain follow-up and for comparison of means. Consistency point deducted for different comparisons
1 (28) [21]	Social function	Child's Talk v existing care	4	−1	0	0	0	Moderate	Quality point deducted for sparse data
1 (29) [22]	Social function	More Than Words v delayed access to programme	2	−1	0	0	0	Very low	Quasi-randomised RCT. Quality point deducted for sparse data
2 (118) [25] [23] [24]	Social function	PECS v other treatment or no treatment	4	−3	0	0	0	Very low	Quality points deducted for sparse data, incomplete reporting, and no subgroup for autism.
1 (22) [26]	Cognitive function	TEACCH v usual care	2	−2	0	0	0	Very low	Quasi-randomised study. Quality points deducted for sparse data and baseline differences
What are the effects of dietary interventions in children with autism?									
1 (20) [29]	Global improvement	Advice to follow gluten and casein free diet v no dietary advice	4	−2	0	0	0	Low	Quality points deducted for sparse data and baseline differences
What are the effects of drug treatments in children with autism?									
1 (66) [31]	Behavioural function	Methylphenidate v placebo	4	−2	0	0	0	Low	Quality points deducted for sparse data and uncertainty about clinical relevance of improvement
3 (208) [32]	Behavioural function	Risperidone v placebo	4	−1	0	0	0	Moderate	Quality point deducted for incomplete reporting
6 (242) [36]	Global improvement	Secretin v placebo	4	0	0	−1	0	Moderate	Directness point deducted for heterogeneous population
What are the effects of non-drug treatments in children with autism?									
No studies found									

Type of evidence: 4 = RCT; 2 = Observational; 1 = Non-analytical/expert opinion. Consistency: similarity of results across studies. Directness: generalisability of population or outcomes.

ClinicalEvidence

¿Qué es la ***fuerza de la recomendación*** y por qué es importante?

Las guías deben indicar si:

(a) la literatura disponible es de calidad y los efectos deseables son mayores que los indeseables, o

(b) existe un balance poco claro entre beneficios y riesgos.

La fuerza de una recomendación es el ***grado en el que podemos confiar que aplicando una recomendación los efectos positivos sobre el paciente serán mayores que los negativos***

Guyatt et al. BMJ 2008;336:1049-51

¿Cómo gradua la fuerza de las recomendaciones el el sistema GRADE?

El sistema GRADE propone una gradación simple de las recomendaciones en:

Fuertes — los efectos deseables de una intervención son claramente mayores que los indeseables, o viceversa

Débiles — cuando el balance entre los beneficios y riesgos es más incierto

Guyatt et al. BMJ 2008;336:1049-51

¿Cómo gradua la fuerza de las recomendaciones el el sistema GRADE? (II)

Hay cuatro factores que determinan si una recomendación es débil o fuerte:

- Calidad de la evidencia
- Incertidumbre sobre el balance beneficio – riesgos
- Incertidumbre en la variabilidad de los valores y preferencias de los pacientes
- Incertidumbre sobre los costes

Guyatt et al. BMJ 2008;336:1049-51

Calidad de la evidencia

Si el grupo de trabajo de la guía no dispone de estimadores de resultado fiables, no debería formular recomendaciones fuertes

Guyatt et al. BMJ 2008;336:1049-51

Balance beneficio – riesgo

Aspecto determinante para graduar la fuerza de las recomendaciones

In patients with atrial fibrillation at low risk of stroke, warfarin can reduce that low risk, but adds inconvenience and increases the risk of bleeding.

Guyatt et al. BMJ 2008;336:1049-51

Valores y preferencias

Algunas preguntas plantean una variabilidad en los valores y preferencias de los pacientes. Se debe considerar el impacto sobre la fuerza de las recomendaciones de esta variabilidad cuando existan diferentes alternativas.

Consider the decision faced by pregnant women with deep venous thrombosis. Warfarin treatment between 6th and 12th week of pregnancy will put infants to minor developmental abnormalities. The alternative treatment, heparin, eliminates the risk for the child, but disadvantages of pain, and cost exist. Clinicians' experience is that women place a high value on preventing fetal complications.

Guyatt et al. BMJ 2008;336:1049-51

Costes

Se debe considerar la variabilidad en términos de disponibilidad de recursos, geográfica o temporales

Higher the costs of an intervention – that is, the more resources consumed – less likely a strong recommendation formulated

Guyatt et al. BMJ 2008;336:1049-51

Para saber más

Serie J Clin Epidemiol

J Clin Epidemiol. 2011 Apr;64(4):383-94.

J Clin Epidemiol. 2011 Apr;64(4):395-400.

J Clin Epidemiol. 2011 Apr;64(4):401-6.

J Clin Epidemiol. 2011 Apr;64(4):407-15.

Serie BMJ

BMJ. 2008 Apr 26;336(7650):924-6.

BMJ. 2008 May 3;336(7651):995-8.

BMJ. 2008 May 10;336(7652):1049-51.

BMJ. 2008 May 17;336(7653):1106-10.

BMJ. 2008 May 24;336(7654):1170-3.

ELSEVIER

Journal of Clinical Epidemiology ■ (2010) ■

Journal of Clinical Epidemiology

COMMENTARY

Rating the evidence in comparative effectiveness reviews

Yngve Falck-Ytter[a], Holger Schünemann[b], Gordon Guyatt[b,*]

[a]Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA
[b]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

Accepted 17 January 2010

Authorities acknowledge that systematic reviews provide the optimal basis for collecting and assessing the evidence that bears on patient management recommendations. In his article introducing JCE's series describing the Agency for Healthcare Research and Quality (AHRQ)'s effective health care program, Mark Helfand distinguishes between systematic reviews and ''complex evidence reports'' that address a broader range of questions, including ''definition, diagnosis, management, and follow-up of a disease or condition.'' Aside from definition, all these questions appear to us as an examination of alternative approaches to managing patients. Such issues are best addressed by structured questions and, if brought together in a single document, constitute a series of related systematic reviews or overviews of systematic reviews.

In this commentary, we address the conduct of such systematic reviews. We present a perspective arising from our participation in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group. GRADE, a strategy for rating quality of evidence and grading strength of recommendations [1,2], has been endorsed by a wide variety of prominent organizations (American College of Physicians, World Health Organization, Cochrane Collaboration, National Institute for Clinical Excellence, and UpToDate) and is emerging as the dominant system for rating quality of evidence and strength of recommendations in guideline groups around the world.

In the current US environment, systematic reviews addressing alternative management strategies under the rubric of ''comparative effectiveness reviews'' have gained a high profile. Central to AHRQ's evidence-based practice centers' (EPCs)—and anyone else's—effort to produce such reviews is the question of how to best conduct a thorough assessment of a body of evidence. Crucial to that issue is how one defines what we (and GRADE) call ''quality'' and what, in their article in JCE's series describing EPC's approach to comparative effectiveness reviews, Owen and authors call ''strength of evidence.''

Although the forefathers of grading systems included aspects beyond ''risk of bias'' [3], quality has often been used to describe ''risk of bias.'' Although they do not explicitly define ''strength of evidence,'' Owens et al's characterization is completely consistent with GRADE's definition of quality of evidence from systematic reviews: the extent to which we can be confident in estimates of the magnitude of effect. As Owens and colleagues point out, quality is, therefore, much more than risk of bias, but includes domains, such as precision, consistency, and directness, as well as considerations of publication bias, magnitude of effect, and dose–response relationship.

Owens and colleagues describe the EPC approach as based in large measure on the GRADE working group approach. We agree, and in the remainder of this commentary, we will reflect on the possible differences between the GRADE and EPC approaches, differences that we view as minor.

Both systems use four categories of quality, three of which carry the same labels (high, moderate, and low) and share the same definitions; the final category is characterized by GRADE as ''very low'' and by the EPCs as ''insufficient.'' The term ''insufficient,'' as we understand it, implies insufficient to make a decision. This judgment is in the domain of a guideline panel rather than systematic review authors. Furthermore, the necessity to make decisions even when evidence is low quality—acknowledged explicitly by Owens and colleagues—may apply with equal force when evidence is of very low quality.

GRADE explicitly designates that randomized trials begin as high-quality evidence and may be rated down by limitations in each major area (risk of bias, imprecision, inconsistency, indirectness, and publication bias). Observational studies begin as low-quality evidence and may be rated up by a large magnitude of effect, a dose–response relationship, and an inference that plausible sources of bias could only diminish apparent effects or increase absent effects. This hierarchy of study designs underlies EPC judgments but is not made explicit in the same way.

Owens and colleagues include along with patient-important outcomes, surrogate markers among the major outcomes for their reviews. Although this statement is open

* Corresponding author.
E-mail address: guyatt@mcmaster.ca (G. Guyatt).

doi: 10.1016/j.jclinepi.2010.01.003

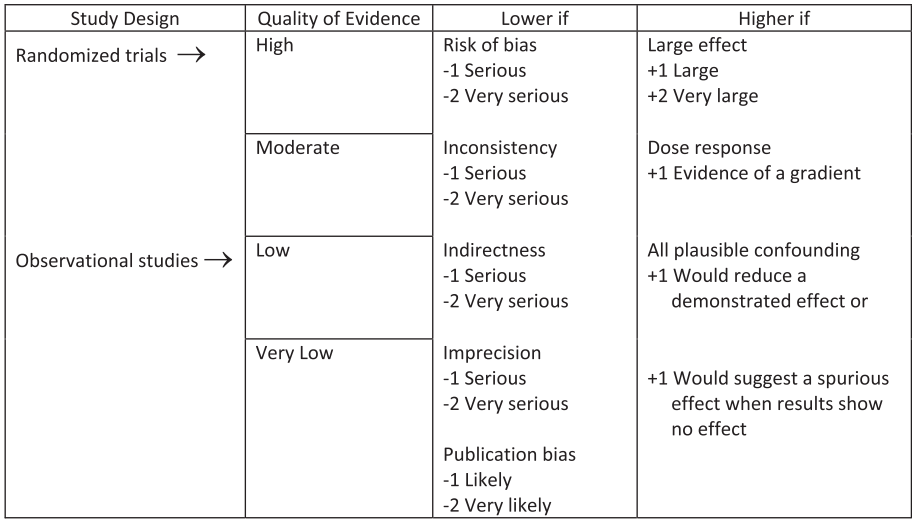


Fig. 1. A summary of Grading of Recommendations Assessment, Development and Evaluation (GRADE)'s approach to rating quality of evidence.

to interpretation, it reflects a possible difference in approach. GRADE advocates focusing exclusively on patient-important outcomes and, when it is necessary to consider surrogates, to view surrogates only as indirect evidence for patient-important outcomes. This implies a need to estimate the magnitude of effect on patient-important outcomes and consider the uncertainty of relying on surrogates to estimate magnitude of effect. This is, however, completely consistent with the example that Owens and colleagues provide of indirectness introduced by the measurement of nutritional variables rather than the patient-important outcome of wound healing.

Owens and colleagues define consistency as ''the degree to which reported effect sizes from included studies appear to have the same direction of effect.'' In a comparison of intervention A and B, small, apparent effects that favor A may be completely consistent (ie, easily explained by chance) with small apparent effect that favors B. Thus, in considering consistency, GRADE focuses primarily on the magnitude of differences in estimates of effect and the associated precision of those estimates.

Another apparent difference we view as purely semantic. When the patients under consideration differ in important ways from those studied or the interventions under consideration differ from those studied, GRADE classifies the evidence as indirect. The EPCs classify such considerations under the rubric ''applicability.'' The underlying conceptual issues are, however, identical.

Determining the final quality of evidence for each important outcome requires careful consideration of all domains. Owens and colleagues acknowledge the strengths, in terms of transparency, of explicit decisions to rate up or down the quality of evidence and thus provide EPCs with the option of using GRADE's approach in this regard (Fig. 1). They correctly point out the lack of evidence to choose between this approach and what they call a ''qualitative approach''—which makes the same judgments without specific attribution to individual domains. Owens' and colleagues' wise counsel to EPCs to make the rationale for their decisions about final ratings of evidence quality explicit suggests that in the end the approaches differ very little.

We reiterate that these differences are of little importance. This is good news for the EPCs target audiences: they can interpret EPC's strength of evidence ratings in the same way as the quality of evidence ratings used by the over 40 systematic review, guideline, and health technology assessment agencies worldwide that have adopted GRADE.

References

[1] Guyatt GH, Oxman AD, Vist G, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. for the GRADE Working Group. Rating quality of evidence and strength of recommendations GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.

[2] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ; GRADE Working Group. Rating quality of evidence and strength of recommendations: what is ''quality of evidence'' and why is it important to clinicians? BMJ 2008;336:995–8.

[3] Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J 1979;121(9):1193–254.

PREGUNTA – APARTADO: ¿Está recomendado el uso de la risperidona para la modificación del comportamiento en personas con un trastorno del espectro autista?

Tabla descriptiva de los estudios evaluados en el desarrollo de la síntesis de la evidencia

DISEÑO	PACIENTES	INTERVENCIÓN(*)	VARIABLES	RIESGO DE SESGO COMENTARIOS
ID: Jesner 2007 **Diseño**: Revisión Cochrane con MA **Objetivo**: Evaluar la eficacia y seguridad de la risperidona en personas con un trastorno del espectro autista **Año de búsqueda**: Abril 2006 **Tipo de estudios incluidos**: ECA doble ciego **Núm. estudios incluidos**: 3	**Criterios inclusión**: ECA de risperidona vs placebo en pacientes con un diagnóstico de trastorno del espectro autista, con almenos una medida de resultado estandarizada **Núm. pacientes**: 179 **Edad**: 5 años a adultos **Sexo**: ¾ partes de los participantes eran hombres **Variables autismo**: Diagnóstico DSM en todos los estudios	**Intervención**: Risperidona, en diferentes dosis y planes de incremento **Control**: Placebo **Comparaciones**: Risperidona vs Placebo	**Principal**: Comportamiento Valoración Global **Secundaria:** Efectos adversos	AMSTAR 1: ok AMSTAR 2: NO **AMSTAR 3: ok** **AMSTAR 4: ok** AMSTAR 5: ok **AMSTAR 6: ok** **AMSTAR 7: ok** **AMSTAR 8: ok** AMSTAR 9: ok AMSTAR 10:NO AMSTAR 11:N/A **Comentarios**: La revisión requiere una actualización.
Ref Bibliogràfica: Jesner OS, Aref-AdibM, Coren E. Risperidone for autism spectrum disorder. Cochrane Database of Systematic Reviews 2007, Issue 1. Art. No.: CD005040. DOI: 10.1002/14651858.CD005040.pub2				

I.2. Tabla GRADE de resumen de los resultados disponibles y calidad de la literatura científica

PREGUNTA – APARTADO: ¿Está recomendado el uso de la risperidona para la modificación del comportamiento en personas con un trastorno del espectro autista?

Bibliography: Jesner OS, Aref-AdibM, Coren E. Risperidone for autism spectrum disorder. Cochrane Database of Systematic Reviews 2007, Issue 1. Art. No.: CD005040. DOI: 10.1002/14651858.CD005040.pub2.

Quality assessment							Summary of findings					Importance
							No of patients		Effect		Quality	
No of studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Other considerations	ABC (Aberrant Behavior Checklist)	Control	Relative (95% CI)	Absolute		
Irritability (Better indicated by lower values)												
2	randomised trials	no serious limitations	serious[1]	no serious indirectness	no serious imprecision	none	88	90	-	MD 8.09 lower (12.99 to 3.19 lower)	+++O MODERATE	CRITICAL
Hyperactivity (Better indicated by lower values)												
2	randomised trials	no serious limitations	no serious inconsistency[2]	no serious indirectness	serious[3]	none	88	90	-	MD 8.98 lower (12.01 to 5.94 lower)	+++O MODERATE	IMPORTANT
Weight gain (Better indicated by lower values)												
2	randomised trials	serious[4]	no serious inconsistency	no serious indirectness	no serious imprecision	none	89	90	-	MD 1.78 higher (1.15 to 2.41 higher)	+++O MODERATE	CRITICAL

[1] Heterogeneidad considerable (I2=78%)
[2] Heterogeneidad leve (I2=20%)
[3] Intervalos de confianza amplios, posiblemente debido a estudios con resultados poco precisos
[4] Estudios sin mayores limitaciones en el diseño, pero con un tiempo de seguimiento muy breve para la medida adecuada de la variable de resultado.

I.3. Modelo de síntesis crítico de la literatura científica.

PREGUNTA – APARTADO: ¿Está recomendado el uso de la risperidona para la modificación del comportamiento en personas con un trastorno del espectro autista?

[Descripción de la literatura disponible] Se han identificado 2 revisiones sistemáticas recientes que analizan exclusivamente la efectividad y seguridad de la risperidona en pacientes con TEA (Jesner 2007 i Chavez 2006). La revisión de Jesner 2007 es una revisión Cochrane en la que se realiza un metanálisis. La revisión de Chavez 2006 se centra en estudios en niños. Existen publicadas 2 revisiones que abordan diferentes tratamientos farmacológicos en general, aportando también información sobre la risperidona (Broadstock 2007 i Parikh 2008). La revisión de Broadstock 2007 se centra en adolescentes y adultos. La revisión de Parikh 2008 solamente incluye estudios en niños y adolescentes que anañizan almenos una medida de resultado relacionada con la agresión.	
[Justificación del uso de la literatura] Se ha seleccionado la revisión de Jesner 2007 por ser la de mayor calidad metodológica y porque la fecha final de la búsqueda bibliográfica es de abril de 2006, siendo más reciente que las revisones de Chavez 2006 y Broadstock 2007. La revisión de Parikh 2008 no aporta información sobre la fecha final de la búsqueda bibliográfica, se limita a MEDLINE y publicaciones en inglés, y en cualquier caso, no encuentra ningún ECA que no haya sido identificado en la revisión de Jesner 2007.	
[Descripción de los estudios valorados en el desarrollo de la recomendación] La revisión de Jesner 2007 incluye 3 ECA (McCracken 2002, McDougle 1998, Shea 2004) que analizan la eficacia y seguridad de la risperidona, comparada con placebo, en individuos con diagnóstico de TEA u otros trastornos del desarrollo generalizados basados en los criterios de *Diagnostic and Statistical Manual of Mental Disorders (DSM) IV*. El número de participantes en los estudios fue pequeño (31, 79 y 101 participantes respectivamente) sumando un total de 211 participantes (152 hombres y 59 mujeres). Uno de los estudios incluye sólo adultos, otro niños (de 5 a 12 años de edad) y el tercer estudio incluye niños y adolescentes (de 5 a 17 años de edad). El tiempo de seguimiento es breve en los tres estudios (de 8 a 12 semanas).	
[Descripción de los resultados por variable de resultado de interés] Comparada con placebo la risperidona muestra una mejoría en algunos aspectos conductuales (irritabilidad, aislamiento social, hiperactividad, estereotipia e impresión clínica global) y empeoramiento en el habla inapropiada. Al medirlas con la *Aberrant Behavior Checklist* la risperidona mostró una reducción signaficativa los niveles de irratibilidad (2 ECA, 179 participantes; DM: - 8.09, 95% CI - 12.99 a - 3.19; P <0.0001) hiperactividad (2 ECA, 179 participantes; DM: - 8.98, 95% CI - 12:01 a - 5.94; P <0.0001)	Calidad moderada
El principal efecto secundario registrado a corto plazo es el aumento de peso en los participantes tratados con risperidona (2 ECA, 179 participantes; DM: 1.78, 95% CI 1.15 a 02:41; P <0.0001).	Calidad moderada

Factores moduladores de la fuerza de las recomendaciones

Factor	Comentari
Balance entre beneficios y efectos indeseables	Algunos efectos adversos como el aumento de aproximadamente 1.7 kg en los niños que recibían risperidona, podrían limitar su uso.
Calidad de la evidencia	Moderada. Aunque los estudios no tenian grandes limitaciones en su diseño, incluyeron pocos pacientes y tuvieron una corta duración (aportando problemas de imprecisión).
Valores y preferencias	No se dispone de estudios que hayan valorado este aspecto.
Costes y uso recursos	No se dispone de datos sobre costes.

Valores y preferencias

Referencia:	**Resumen del estudio::**

Costes

Referencia:	**Resumen del estudio::**

Decisión

Recomanación	**□ fuerte,**	**□ a favor**	**de una intervención.**
	□ débil,	**□ en contra**	

Tipo de recomendación

- **Recomendación fuerte:** el grupo elaborador confía en que los potenciales efectos beneficiosos derivados de llevar a cabo la recomendación son mayores que los potenciales efectos adversos.
- **Recomendación débil:** el grupo elaborador concluye, aunque no tiene la certeza, que los potenciales efectos beneficiosos derivados de llevar a cabo la recomendación son mayores que los potenciales efectos adversos.

Formulación de la recomendación

Fuerza de la recomanación:	**Texto de la recomendación:**

Refrencias bibliográficas

Jesner OS, Aref-AdibM, Coren E. Risperidone for autism spectrum disorder. Cochrane Database of Systematic Reviews 2007, Issue 1. Art. No.: CD005040. DOI: 10.1002/14651858.CD005040.pub2.

Research Units on Pediatric Psychopharmacology Autism Network. McCracken JT, McGough J, Shah B, Cronin P, Hong D, Aman MG, Arnold E, Lindsay R, Nash P, Holloway J, McDougle CJ, Posy D, Swiezy N, Kohn A, Scahill L, Martin A, Koenig K, Volkmar F, Carroll D, Lancor A, Tierney E, Ghuman J, Gonzalez NM, Grados M, Vitiello B, Ritz L, Davies M, Robinson J, McMahon D. Risperidone in children with autism and serious behavioral problems. The New England Journal of Medicine 2002; 347(5):314–321.

Shea S, Turgay A, Carroll A, Schulz M, Orlik H, Smith I, Dunbar F. Risperidone in the treatment of disruptive behavioral symptoms in children with autistic and other pervasive developmental disorders. Pediatrics 2004;114(5):644–641.