Valoración de la calidad de la evidencia y fuerza de las recomendaciones

SISTEMA GRADE

Depósito legal:

ISBN: 978-84-96761-81-0

Índice

*Artículos publicados anteriormente en la revista British Medical Journal. BMJ (Ed Esp) 2008; 2(8):443-464

La semFYC también adopta el sistema GRADE

La Sociedad Española de Medicina Familiar y Comunitaria (semFYC), al igual que un número creciente de organizaciones científicas, ha adoptado el sistema GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) para elaborar guías de práctica clínica y clasificar la calidad de la evidencia y la fuerza de las recomendaciones.

La medicina basada en la evidencia (MBE) reconoce dos principios[1]. El primero, que existe una jerarquía de las evidencias que nos confiere una mayor confianza para hacer mejores decisiones clínicas y nos previene de los sesgos y errores arbitrarios. El segundo, que el conocimiento científico sólo no es suficiente para hacer decisiones clínicas. De hecho, la MBE estipula que cualquier intervención clínica requiere de la integración del conocimiento clínico y de los resultados de la investigación, teniendo en cuenta las circunstancias de los pacientes, sus valores y preferencias[1].

Decidir si una determinada intervención clínica resulta adecuada para un paciente concreto equivale a determinar si existe un grado razonable de certeza de que el balance entre los beneficios, por un lado, y los riesgos, los inconvenientes y los costes, por el otro, de dicha intervención es lo suficientemente favorable como para que merezca la pena aplicarla. Dicha decisión es, o al menos debería ser, el resultado final de una serie de juicios secuenciales que, por su complejidad, requiere que los médicos (y los pacientes) la realicen con ayuda[2].

Las guías de práctica clínica (GPC) son una excelente herramienta donde obtener esta ayuda[3]. Para que las GPC sean útiles se han de formular recomendaciones claras basadas en la mejor evidencia disponible y describir las circunstancias, preferencias y valores que han llevado a los autores a desarrollar las recomendaciones. Para que los médicos clínicos (y los pacientes) confíen en las guías, el procedimiento de explicitación de la calidad de la evidencia y los elementos para determinar la fuerza de las recomendaciones ha de ser trasparente[2].

La Canadian Task Force on Preventive Health Care (CTFPHC) desarrolló, hace ya más de 30 años, el primer sistema de clasificación de la calidad de la evidencia y la fuerza de las recomendaciones. Durante las últimas décadas, diversos grupos elaboradores de recomendaciones y GPC han desarrollado nuevos sistemas con la intención de ayudar a los profesionales sanitarios en la toma de decisiones clínicas. Actualmente se contabilizan más de cien sistemas y, aún siendo indudable la contribución de muchos de ellos, la múltiple proliferación de sistemas y la numeración, símbolos y términos utilizados, también ha generado confusión[4].

Mercè Marzo Castillejo
Secretaria del Comité Científico de la semFYC.
Institut Català de la Salut
Rafael Rotaeche del Campo
Coordinador del grupo MBE de la semFYC.
Centro de Salud de Alza, Osakidetza, San Sebastián
Josep Basora Gallisa
Vicepresidente de la Junta Permanente de la semFYC.
Institut Català de la Salut

Correspondencia:
Mercè Marzo Castillejo
semFYC
C/ Portaferrissa 8, pral.
08002 Barcelona
Tel. 93.317.03.33
Correo electrónico:
mmarzoc@gencat.cat

Desde el año 2000, un grupo internacional integrado en su mayoría por expertos en metodología y por clínicos, muchos de ellos procedentes de las organizaciones que establecieron los sistemas de clasificación más conocidos o de organizaciones de notable peso tradicional o actual en la formulación de recomendaciones (US Preventive Service Task Force –USPSTF–, Scottish Intercollegiate Guidelines Network –SIGN–, Oxford Center for Evidence Based Medicine, National Institute for Health and Clinical Excellence –NICE–, han trabajado en la iniciativa GRADE.

El grupo GRADE internacional se propuso: 1) evaluar los diferentes sistemas disponibles; 2) desarrollar un nuevo sistema de clasificación; y 3) diseminar el nuevo sistema a través de la comunidad científica y de sus publicaciones. El desafío ha sido enorme pues todos los sistemas de clasificación tienen sus limitaciones, y muchas de las organizaciones que forman parte del grupo GRADE internacional ya habían gastado recursos significativos para el desarrollo de sus propios sistemas de clasificación[5].

Las primeras conclusiones y propuestas del grupo GRADE fueron publicadas en el año 2004[6]. Los criterios del sistema GRADE son simples y aplicables a una gran variedad de recomendaciones clínicas que abarcan un amplio espectro de decisiones en el manejo de los pacientes. El enfoque del sistema GRADE, para realizar los complejos juicios que subyacen al clasificar la calidad de la evidencia y la fuerza de las recomendaciones, es sistemático y explícito. GRADE es un sistema que ayuda a prevenir errores y a resolver desacuerdos, y facilita la lectura crítica y la comunicación de la información.

Como puede comprobarse a través del ejemplo que sigue, los juicios secuenciales del sistema GRADE guardan similitud con el proceso de toma de decisiones que el clínico sigue en el día a día de la consulta[2]. Disponemos de evidencia de calidad alta, derivada de ensayos clínicos aleatorios bien diseñados y ejecutados, mostrando que los anticoagulantes orales administrados durante más de un año reducen el riesgo de recurrencias en pacientes que han sufrido un episodio de trombosis venosa profunda idiopática[7]. Por otra parte, sabemos que los anticoagulantes orales aumentan el riesgo de sangrado y tienen inconvenientes tales como tener que tomar la medicación y monitorizar el nivel de anticoagulación, además de los costes asociados, sobre todo, a los programas de monitorización[7]. Por ello, la recomendación de anticoagular durante más de un año a todos los pacientes es débil, ya que el balance de los beneficios, por un lado, y de los riesgos, inconvenientes y costes, por el otro, es incierto y los pacien-

tes bien informados pueden inclinarse por opciones diferentes (mantener la anticoagulación oral más de un año o no).

El interés por el sistema GRADE trasciende a los expertos en metodología y elaboradores de GPC, y resulta una herramienta muy interesante para sistematizar el proceso de toma de decisiones en nuestra actividad clínica. Actualmente, numerosas organizaciones han apoyado o están utilizando GRADE como sistema de clasificación en sus recomendaciones y GPC. Estas organizaciones incluyen: la Organización Mundial de la Salud (OMS), la Colaboración Cochrane Internacional, la Agency for Healthcare Research and Quality (AHRQ) de EEUU, el National Institute for Clinical Excellence (NICE) del Reino Unido, BMJ Clinical Evidence del Reino Unido; y diversas sociedades científicas como la American College of Chest Physicians, American Thoracic Society, American College of Physicians Endocrine, Society European Respiratory Society y, también, la semFYC. (La lista completa está disponible en la web del grupo GRADE)[8]. Este amplio apoyo muestra que a nivel internacional el consenso entorno al sistema GRADE es importante.

Además de la semFYC, hay otras sociedades científicas de nuestro entorno, entre ellas: la Sociedad Española de Neumología y Cirugía Torácica (SEPAR) y la Sociedad Española de Rehabilitación y Medicina Física (SERMEF), que también han optado por el sistema GRADE para elaborar algunas de sus guías con GRADE. Asimismo, la Agencia de Calidad del Sistema Nacional de Salud ha publicado un manual metodológico para la elaboración de GPC que incluye un capítulo sobre el sistema GRADE[9]. El manual ha sido elaborado por un grupo de profesionales formado, entre otros, por miembros de las Agencias de Evaluación de Tecnología Sanitaria.

Nuestra sociedad científica, la semFYC, ha valorado las ventajas de apoyar el sistema GRADE para clasificar la calidad de la evidencia y la fuerza de las recomendaciones[10]. Consideramos que el sistema GRADE puede contribuir a mejorar la calidad y transparencia de las GPC producidas en nuestro entorno así como facilitar al clínico la toma de decisiones con sus pacientes. Es así, que a lo largo de estos últimos años, el Comité Científico de semFYC, integrado por profesionales de perfil clínico y metodológico, referentes en sus respectivas responsabilidades dentro de los proyectos semFYC, y los miembros del Grupo MBE hemos ido incorporando la metodología GRADE en algunas de las actividades científicas realizadas, como consensos, recomendaciones, GPC y formación.

Entre las actividades docentes se han impartido cursos GRADE dirigidos a los miembros de los grupos de trabajo de la semFYC y del Programa de Actividades Preventivas y de Promoción de la Salud (PAPPS), en las tres últimas ediciones de la Escuela de Verano de la semFYC, en las primeras jornadas específicas de formación en MBE y GRADE (realizadas con financiación institucional del Instituto Carlos III) y en actividades monográficas semFYC sobre talleres GRADE.

También entre otros propósitos de la semFYC está el colaborar en la difusión y promoción de la traducción de los materiales de GRADE al castellano. En un primer momento se publicó la traducción del primer artículo de GRADE junto a una editorial en la revista de *Atención Primaria*[11]. Ahora nos ha parecido de mucha utilidad la traducción de la nueva serie de artículos que durante el año 2008 se han publicado en el BMJ y que son el objeto de esta monografía.

Nuestro agradecimiento a los Laboratorios Sanofi Aventis, que siempre están dispuestos a colaborar en proyectos de la semFYC, por hacer posible su difusión y acercar la información a un amplio número de profesionales de semFYC y de otras sociedades científicas y organizaciones.

1. Guyatt GH, Haynes B, Jaeschke R, et al. Introduction: the philosophy of evidence-based medicine. In: Guyatt GH, Rennie D (ed). Users' guides to the medical literature: a manual of evidence based clinical practice. Chicago: AMA Press; 2002. p. 121-40.
2. Marzo Castillejo M, Montaño Barrientos A. El sistema GRADE para la toma de decisiones clínicas y la elaboración de recomendaciones y guías de práctica clínica. Aten Primaria, 2007;39:457-60.
3. Field Mj, Lohr KN. Clinical Practice Guidelines. From Development to Use. Washington: National Academy Press; 1992.
4. The GRADE Working Group. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches. BMC Health Serv Res. 2004;4:38.
5. Swiglo BA, Murad MH, Schünemann HJ, Kunz R, Vigersky RA, Guyatt GH, et al. Acase for clarity, consistency, and helpfulness: state-of-the-art clinical practice guidelines in endocrinology using the grading of recommendations, assessment, development, and evaluation system. J Clin Endocrinol Metab. 2008;93:666-73.
6. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al, GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328:1490.
7. Kearon C, Kahn SR, Agnelli G, Goldhaber S, Raskob GE, Comerota AJ; American College of Chest Physicians. Antithrombotic therapy for venous thromboembolic disease: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest. 2008 Jun;133(6 Suppl):454S-545S.
8. The Grading of Recommendations Assessment, Development and Evaluation (short GRADE) Working Group. Disponible en: URL: http://www.gradeworkinggroup.org/
9. Grupo de trabajo sobre GPC. Elaboración de Guías de Práctica Clínica en el Sistema Nacional de Salud. Manual Metodológico. Madrid: Plan Nacional para el SNS del MSC. Instituto Aragonés de Ciencias de la Salud-I+CS; 2007. Guías de Práctica Clínica en el SNS: I+CS Nº 2006/0I.
10. Marzo Castillejo M, Basora J, Rotaeche R, Mascort J. La trayectoria científica de semFYC. ¿Hacia dónde queremos avanzar? Aten Primaria. 2005;35:447-8.
11. Marzo Castillejo M, Alonso-Coello P, Rotaeche del Campo R. ¿Cómo clasificar la calidad de la evidencia y la fuerza de las recomendaciones? Aten Primaria. 2006;37:5-7.

BMJ (Ed Esp) 2008; 2(8):443-446

VALORACIÓN DE LA CALIDAD DE LA EVIDENCIA Y FUERZA DE LAS RECOMENDACIONES

GRADE: Un consenso emergente sobre la evaluación de la calidad de la evidencia y la fuerza de las recomendaciones

Las directrices valoran de distintas maneras la calidad de la evidencia y la fuerza de las recomendaciones. En este artículo, se analizan las ventajas del sistema GRADE, que están adoptando un número cada vez mayor de organizaciones sanitarias de todo el mundo

Los expertos que elaboran directrices en diversas partes del mundo evalúan de distinta manera la calidad de la evidencia y la fuerza de las recomendaciones. En consecuencia, para quienes utilizan las directrices es más difícil comprender los conceptos que tratan de comunicar los sistemas de gradación. Desde 2006, el *BMJ* ha solicitado en sus «Instrucciones para los autores» (en www.bmj.com) que los investigadores utilicen de preferencia el sistema de gradación de la evidencia Grading of Recommendations Assessment, Development and Evaluation (GRADE) cuando remitan un artículo sobre directrices clínicas. ¿En qué se basa esta decisión?

En este primero de una serie de cinco artículos se explicará por qué muchas organizaciones utilizan sistemas formales para asignar grados a la evidencia y a las recomendaciones, y por qué esto es importante para los médicos, y se abordará también el enfoque GRADE para las recomendaciones. En los siguientes dos artículos se analizará de qué manera el sistema GRADE clasifica la calidad de la evidencia y la fuerza de las recomendaciones. En los últimos dos artículos se abordarán las recomendaciones para las pruebas diagnósticas y el modelo del sistema GRADE para evaluar la repercusión de las intervenciones en el empleo de los recursos.

El sistema GRADE ofrece ventajas con relación a los sistemas de evaluación previos (cuadro 1). Existen otros sistemas que también tienen algunas de estas ventajas, pero ninguno (con excepción del GRADE) las combina todas[1].

Cuadro 1 | Ventajas del sistema GRADE con respecto a otros sistemas

- Ideado por un grupo ampliamente representativo de especialistas internacionales que elaboran directrices.
- Clara separación entre la calidad de la evidencia y la fuerza de las recomendaciones.
- Evaluación explícita de la importancia de los desenlaces de estrategias de tratamiento alternativas.
- Criterios explícitos y exhaustivos para reducir y aumentar el grado de calidad de las evaluaciones de la evidencia.
- Proceso transparente para ir de la evidencia a las recomendaciones.
- Reconocimiento explícito de valores y preferencias.
- Interpretación clara y pragmática de recomendaciones fuertes frente a débiles para médicos, pacientes y autoridades sanitarias.
- Útil para análisis sistemáticos, valoraciones de tecnologías sanitarias y directrices.

Gordon H Guyatt
Professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON L8N 3Z5 (Canadá)
Andrew D Oxman
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo (Noruega)
Gunn E Vist
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo (Noruega)
Regina Kunz
Associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basilea (Suiza)
Yngve Falck-Ytter
Assistant professor, Division of Gastroenterology, Case Medical Center, Case Western Reserve University, Cleveland, OH 44106 (Estados Unidos)
Pablo Alonso-Coello
Researcher, Iberoamerican Cochrane Center, Servicio de Epidemiología Clínica y Salud Pública (Universidad Autónoma de Barcelona), Hospital de Sant Pau, Barcelona 08041 (España)
Holger J Schünemann
Professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, Roma (Italia)
Para el grupo de trabajo de GRADE

Correspondencia:
G H Guyatt, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main Street, West Hamilton, ON, L8N 3Z5 (Canadá); *guyatt@mcmaster.ca*

Éste es el primero de una serie de cinco artículos que explican el sistema GRADE para evaluar la calidad de la evidencia y la fuerza de las recomendaciones

¿QUÉ ES LA «CALIDAD DE LA EVIDENCIA» Y POR QUÉ ES IMPORTANTE?

Al tomar decisiones para la gestión de la asistencia sanitaria, los pacientes y los médicos deben sopesar las ventajas y los inconvenientes de estrategias alternativas.

Las autoridades sanitarias están influenciadas no sólo por los mejores cálculos de las ventajas e inconvenientes esperados, sino también por su confianza en tales cálculos. La caricatura que ilustra la incertidumbre de los expertos en predicción del tiempo muestra la diferencia entre la evaluación de la probabilidad de un desenlace y la certidumbre de tal evaluación (figura). La utilidad de un cálculo de la magnitud de los efectos de la intervención depende de la confianza en dicho cálculo.

A menudo, los médicos expertos y las organizaciones que ofrecen recomendaciones a la comunidad médica han cometido errores como resultado de no tener suficientemente en cuenta la calidad de la evidencia[2]. Durante una década, los organismos aconsejaron a los médicos que recomendasen a las mujeres posmenopáusicas tomar hormonoterapia restitutiva[3]. Muchos médicos de atención primaria aplicaron este consejo en sus consultas pensando que cumplían con su deber.

La idea de que este tratamiento disminuía sustancialmente el riesgo cardiovascular de las mujeres motivó la recomendación. Si se hubiese aplicado entonces un sistema riguroso para evaluar la calidad de la evidencia, se habría demostrado que los datos se derivaban de estudios de observación con resultados poco concluyentes y que la evidencia de la reducción en el riesgo cardiovascular era de muy baja calidad[4]. El reconocimiento de las limitaciones de la evidencia habría moderado las recomendaciones. Posteriormente, en ensayos aleatorizados comparativos se ha demostrado que la hormonoterapia restitutiva no reduce el riesgo cardiovascular y que puede incluso aumentarlo[5,6].

La Agencia del Medicamento (FDA) estadounidense autorizó los fármacos antiarrítmicos encainida y flecainida basándose en su capacidad para reducir las arritmias ventriculares asintomáticas que pueden causar la muerte súbita. En esta decisión no se tuvo en cuenta que, puesto que la reducción de

«Deduzco que hay una probabilidad de lluvias del 40% y una probabilidad del 10% de que sepamos de qué estamos hablando».

la arritmia reflejaba sólo indirectamente el desenlace de muerte súbita, la calidad de la evidencia sobre la utilidad de los fármacos era de baja. Posteriormente, un ensayo aleatorizado comparativo demostró que los dos fármacos aumentan el riesgo de muerte súbita[7]. La atención apropiada a la baja calidad de la evidencia habría salvado millares de vidas.

Cada vez que se deja de reconocer una evidencia de gran calidad se pueden ocasionar problemas similares. Por ejemplo, los expertos tardaron diez años en recomendar el tratamiento trombolítico para los pacientes con infarto de miocardio, a pesar de la evidencia derivada de ensayos aleatorizados comparativos bien realizados indicativos de que se lograba una reducción en la mortalidad[8].

La atención insuficiente a la calidad de la evidencia conlleva el riesgo de que se establezcan directrices y recomendaciones inadecuadas que pueden llevar a los médicos a poner en práctica medidas que perjudiquen a sus pacientes. Reconocer la calidad de la evidencia ayudará a prevenir estos errores.

¿CÓMO DEBEN ALERTAR A LOS MÉDICOS CON RESPECTO A LA CALIDAD DE LA EVIDENCIA A QUIENES ELABORAN DIRECTRICES?

Los sistemas formales que clasifican la calidad de la evidencia —por ejemplo, de alta a muy baja— son métodos razonables para comunicar la calidad de la evidencia a los médicos. Sin embargo, tienen algunas limitaciones. La calidad de la evidencia es un proceso continuo, y cualquier clasificación definida implica cierto grado de arbitrariedad. No obstante, las ventajas de la simplicidad, la claridad y la intensidad superan estas limitaciones.

¿QUÉ ES LA «FUERZA DE LA RECOMENDACIÓN» Y POR QUÉ ES IMPORTANTE?

Una recomendación de un tratamiento determinado puede ser resultado de ensayos aleatorizados comparativos y rigurosos a gran escala que muestren unas ventajas marcadas y uniformes, con escasos efectos secundarios y mínimas incomodidades y costes. Este es el caso del empleo de un ciclo breve de corticoesteroides orales para las exacerbaciones del asma. Los médicos pueden prescribir estos fármacos a casi todos sus pacientes sin titubeos.

Pero las recomendaciones de tratamiento también pueden originarse a partir de estudios de observación y pueden implicar considerables daños, cargas o costes. Para determinar si es conveniente prescribir un antitrombótico a las mujeres embarazadas con prótesis de válvulas cardíacas es necesario evaluar la magnitud de la reducción en la trombosis de la válvula considerando la incomodidad, los costes y el riesgo de teratogenia del tratamiento. Los médicos que ofrecen estos tratamientos deben ayudar a las pacientes a sopesar cuidadosamente los efectos favorables y adversos de acuerdo con sus valores y sus preferencias.

Por consiguiente, las directrices y las recomendaciones deben indicar: *a*) si la evidencia es de gran calidad y los efectos favorables superan claramente a los adversos, o *b*) si hay un equilibrio cercano o dudoso. Una gradación simple y clara de la recomendación puede transmitir eficazmente esta información clave.

La gradación formal de las recomendaciones tiene limitaciones. Al igual que la calidad de la evidencia, el equilibrio entre los efectos favorables y adversos es un proceso continuo. Por tanto, asignar a recomendaciones concretas categorías como «fuerte» y «débil» implica cierta arbitrariedad. La mayoría de las organizaciones que elaboran directrices han determinado que los méritos de un grado explícito de recomendación superan a sus inconvenientes.

¿QUÉ CARACTERIZA A UN SISTEMA DE GRADACIÓN SATISFACTORIO?

No todos los sistemas de gradación distinguen entre las decisiones relativas a la calidad de la evidencia y la fuerza de las recomendaciones. Los que no lo hacen, crean confusión. La evidencia de gran calidad no implica necesariamente que las recomendaciones sean concluyentes, puesto que también pueden hacerse con una evidencia de baja calidad.

Por ejemplo, los pacientes que experimentan por primera vez trombosis venosa profunda sin un factor desencadenante evidente deben decidir, después de los primeros meses de tratamiento anticoagulante, si continúan tomando warfarina a largo plazo. Los ensayos aleatorizados comparativos de gran calidad muestran que mantener la administración de warfarina disminuye el riesgo de recidivas, pero a costa de aumentar el riesgo de hemorragia y las molestias. Puesto que los pacientes con valores y preferencias variables optarán por decisiones distintas, los grupos de expertos que elaboran las directrices y analizan si los pacientes deben mantener o suspender el tratamiento warfarina están obligados, pese a la gran calidad de la evidencia, a ofrecer una recomendación débil.

Considérese la decisión de administrar ácido acetilsalicílico o paracetamol a niños con varicela. En los estudios de observación se ha encontrado una rela-

ción entre la administración de ácido acetilsalicílico y el síndrome de Reye[9]. Puesto que el ácido acetilsalicílico y el paracetamol tienen efectos analgésicos y antipiréticos similares, la evidencia de baja calidad con respecto a la relación entre el ácido acetilsalicílico y el síndrome de Reye no impide una recomendación clara de paracetamol.

Los sistemas que clasifican la «opinión de los expertos» como una categoría de la evidencia también crean confusión. El criterio es necesario para interpretar toda la evidencia, sea ésta de alta o de baja calidad. Los informes de los expertos sobre su experiencia clínica deberán considerarse explícitamente como evidencia de muy baja calidad, junto con los informes de casos y otras observaciones clínicas no comparadas.

Los sistemas de calificación sencillos con respecto a los criterios sobre la calidad de la evidencia y la fuerza de las recomendaciones facilitan su uso por parte de pacientes, médicos y autoridades sanitarias[1]. Los criterios detallados y explícitos para evaluar la calidad de la evidencia y calificar su fuerza son más claros para quienes aplican las directrices y las recomendaciones.

Aunque muchos sistemas de gradación cumplen, en cierta medida, con estos criterios[1], muchos de ellos son difíciles de utilizar para los médicos que atienden a pacientes. Tratar de comprender una variedad de sistemas no es un empleo eficiente o realista del tiempo de un médico. El sistema GRADE es utilizado por muchos organismos y organizaciones: la Organización Mundial de la Salud, el American College of Physicians, la American Thoracic Society, UpToDate (un recurso electrónico ampliamente utilizado en Norteamérica, www.uptodate.com) y la colaboración Cochrane son algunas de las más de 25 entidades que lo han adoptado. Esta adopción generalizada refleja su éxito como un sistema metodológico de gradación rigurosa fácil de utilizar.

¿CÓMO SE CLASIFICA LA CALIDAD DE LA EVIDENCIA EN EL SISTEMA GRADE?

Para lograr claridad y sencillez, el sistema GRADE clasifica la calidad de la evidencia en uno de cuatro niveles: alta, moderada, baja y muy baja (cuadro 2). Algunas de las organizaciones que lo utilizan han optado por unificar las categorías baja y muy baja. La evidencia basada en ensayos aleatorizados comparativos comienza como evidencia de gran calidad, pero nuestra certidumbre en la evidencia puede disminuir por varias razones, entre las que se incluyen:

Cuadro 2 | Calidad de la evidencia y definiciones

Alta calidad: es muy improbable que las investigaciones adicionales modificarán la certidumbre con respecto al cálculo del efecto.
Calidad moderada: probablemente, las investigaciones adicionales tendrán una repercusión importante en la certidumbre con respecto al cálculo del efecto, y pueden modificarlo.
Baja calidad: muy probablemente, las investigaciones adicionales tendrán una repercusión importante en la certidumbre con respecto al cálculo del efecto, y es posible que lo modifiquen.
Muy baja calidad: cualquier cálculo del efecto es muy dudoso.

Factores que afectan a la fuerza de una recomendación

Factor	Ejemplos de recomendaciones fuertes	Ejemplos de recomendaciones débiles
Calidad de la evidencia	Muchos ensayos aleatorizados de gran calidad han demostrado la utilidad de los corticoesteroides inhalados en pacientes asmáticos	Sólo algunas series de casos han analizado la utilidad de la pleurodesis en el neumotórax
Incertidumbre sobre el equilibrio entre los efectos favorables y adversos	El ácido acetilsalicílico en el infarto del miocardio reduce la mortalidad con un mínimo de toxicidad, molestias y costes	La warfarina en pacientes con bajo riesgo y fibrilación auricular origina una pequeña reducción en la incidencia de accidentes cerebrovasculares, pero un mayor riesgo de hemorragias y molestias importantes
Incertidumbre o variabilidad en los valores y las preferencias	Los pacientes jóvenes con linfoma otorgan invariablemente un mayor valor a los efectos de prolongación de la vida de la quimioterapia que a la toxicidad del tratamiento	Los pacientes ancianos con linfoma pueden no otorgar un mayor valor a los efectos de prolongación de la vida de la quimioterapia que a la toxicidad del tratamiento
Incertidumbre con respecto a si la intervención representa un uso prudente de recursos	El bajo coste del ácido acetilsalicílico para prevenir el accidente cerebrovascular en pacientes con ataques isquémicos transitorios	El coste elevado del clopidogrel y la combinación dipiridamol-ácido acetilsalicílico como profilaxis contra el accidente cerebrovascular en pacientes con ataques isquémicos transitorios

CONCEPTOS BÁSICOS

No considerar la calidad de la evidencia puede conducir a hacer a recomendaciones erróneas; la hormonoterapia restitutiva en las mujeres posmenopáusicas es un ejemplo ilustrativo de ello.
La evidencia de gran calidad que indica que los efectos favorables de una intervención son claramente superiores que sus efectos adversos, o que claramente no lo son, justifica una recomendación fuerte.
La incertidumbre sobre las permutas (porque la evidencia es de baja calidad o los efectos favorables y adversos están muy equilibrados) justifica una recomendación débil.
Las directrices deberían informar a los médicos cuál es la calidad de la evidencia subyacente y si las recomendaciones son fuertes o débiles.
El enfoque de Valoración, Desarrollo y Evaluación de la Gradación de las Recomendaciones (GRADE) es un sistema de evaluación de la calidad de la evidencia y la fuerza de las recomendaciones explícito, exhaustivo, claro y pragmático, que están adoptando un número cada vez mayor de organizaciones de todo el mundo.

- Limitaciones del estudio.
- Falta de uniformidad de los resultados.
- Carácter indirecto de la evidencia.
- Imprecisión.
- Sesgo de notificación.

Aunque los estudios de observación (p. ej., los de cohortes y los de casos y testigos) comienzan con una calificación de «baja calidad», pero su gradación ascendente puede estar justificada si la magnitud del efecto del tratamiento es muy considerable (p. ej., artrosis grave y reemplazo de la cadera), si hay datos de una relación dosis-respuesta o si todos los sesgos plausibles reducirían la magnitud de un efecto evidente del tratamiento.

¿CÓMO SE CONSIDERA LA FUERZA DE LA RECOMENDACIÓN EN EL SISTEMA GRADE?

El sistema GRADE ofrece dos grados de recomendaciones: «fuerte» y «débil» (aunque los especialistas que elaboran directrices puedan preferir términos como «condicionales» o «discrecionales» en lugar de «débiles»). Cuando los efectos favorables de una intervención superan claramente a los adversos, o claramente no lo superan, los expertos que elaboran directrices ofrecen recomendaciones fuertes. Por otra parte, cuando las permutas son menos seguras —porque la

evidencia es de baja calidad o indica que los efectos favorables y adversos están muy equilibrados—, las recomendaciones son obligatoriamente débiles.

Además de la calidad de la evidencia, hay otros factores que afectan a la fuerza o la debilidad de las recomendaciones (tabla 1).

Los detalles del grupo de trabajo de GRADE, los colaboradores y los conflictos de interés aparecen en la versión de este artículo publicada en www.bmj.com.

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

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

3 American College of Physicians. Guidelines for counseling postmenopausal women about preventive hormone therapy. *Ann Intern Med* 1992;117:1038-41.

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

5 Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA* 1998;280:605-13.

6 Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321-33.

7 Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. *N Engl J Med* 1991;324:781-8.

8 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 1992;268:240-8.

9 Committee on Infectious Diseases. Aspirin and Reye syndrome. *Pediatrics* 1982;69:810-2.

BMJ (Ed Esp) 2008; 2(8):447-451

ANÁLISIS

VALORACIÓN DE LA CALIDAD DE LA EVIDENCIA Y FUERZA DE LAS RECOMENDACIONES

GRADE: ¿Qué es la «calidad de la evidencia» y por qué es importante para los médicos?

Los responsables de elaborar directrices usan una variedad muy compleja de sistemas para valorar la calidad de la evidencia en la que basan sus recomendaciones. Algunas son superficiales, otras confusas y otras son más perfeccionadas, pero también de mayor complejidad

En 2004, el Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group presentó su propuesta inicial de tratamiento de pacientes[1]. En este segundo artículo de una serie de cinco que prestan atención a la estrategia GRADE para elaborar y presentar las recomendaciones, mostramos cómo GRADE se ha añadido a los sistemas existentes para crear un sistema muy estructurado, transparente e informativo de valoración de la calidad de la evidencia.

LA FORMULACIÓN DE LAS DIRECTRICES DEBE INCLUIR UNA PREGUNTA CLARA

Cualquier pregunta que aborde el tratamiento clínico tiene cuatro componentes: los pacientes, la intervención, la comparación y las variables de interés[2]. Por ejemplo, consideremos lo siguiente: en los pacientes con carcinoma pancreático que se someten a cirugía, ¿cuál es el impacto de una resección modificada que conserva el píloro con respecto a una amplia resección estándar del tumor (variaciones del procedimiento de Whipple) sobre la mortalidad a corto y largo plazo, las transfusiones de sangre, las fugas de bilis, la estancia hospitalaria y los problemas del vaciado gástrico?

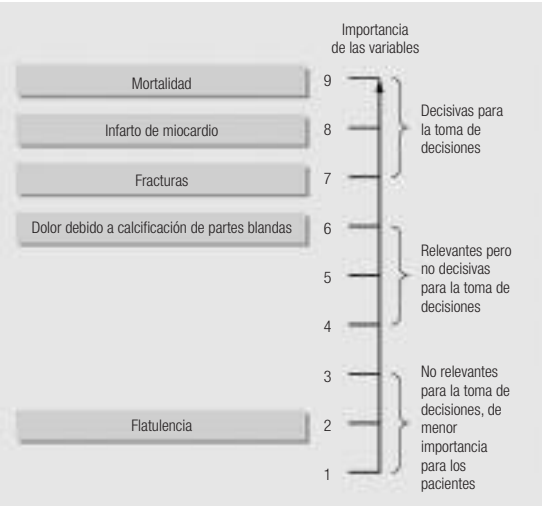


Fig. 1 | Jerarquía de las variables, de acuerdo con la importancia para los pacientes, para la evaluación del efecto de los fármacos que reducen la concentración de fosfato en pacientes con insuficiencia renal e hiperfosfatemia

Gordon H Guyatt
Professor, Department of Epidemiology and Biostatistics, McMaster University, Hamilton ON (Canadá) L8N
Andrew D Oxman
Researcher Norwegian Knowledge Centre the Health Services, PO Box 7004 St Olavs Plass, 0130 Oslo (Noruega)
Gunn E Vist
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004,St Olavs Plass, 0130 Oslo (Noruega)
Regina Kunz
Associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basilea, Hebelstrasse 10, 4031 Basilea (Suiza)
Yngve Falck-Ytter
Assistant professor, Division Gastroenterology, Case Medical Center, Case Western Reserve University, Cleveland OH 44106 (Estados Unidos)
Holger J Schünemann
Professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, Roma (Italia)
Para el grupo de trabajo GRADE

Correspondencia:
G H Guyatt, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main Street, West Hamilton, ON (Canadá) L8N 3Z5 guyatt@mcmaster.ca

Éste es el segundo de una serie de cinco artículos que explican el sistema GRADE para valorar la calidad de la evidencia y la fuerza de las recomendaciones

LOS RESPONSABLES DE ELABORAR DIRECTRICES DEBEN ABORDAR LA IMPORTANCIA DE SUS VARIABLES

Las GRADE obligan a los responsables de establecer directrices a especificar todas las variables relevantes para los pacientes desde el inicio de su desarrollo y a diferenciar las variables decisivas de las importantes que no son críticas[3]. En la figura 1 se presenta una jerarquía de los resultados relevantes para los pacientes con respecto al impacto de los fármacos que reducen la concentración de fosfato en pacientes con insuficiencia renal. La estrategia GRADE sugiere utilizar una escala de nueve puntos para evaluar dicha importancia. El extremo superior de la escala (de 7 a 9) corresponde los resultados de gran importancia para la toma de decisiones. Las valoraciones de 4 a 6 son variables que importantes pero no decisivas, y las de 1 a 3, aspectos de importancia limitada. Los grupos de expertos que elaboran directrices deben procurar seguir este tipo de estrategia explícita.

EVALUAR LA CALIDAD DE LA EVIDENCIA REQUIERE CONSIDERAR EL CONTEXTO

La estrategia GRADE proporciona una definición de la calidad de la evidencia para efectuar recomendaciones. La calidad de la evidencia refleja el grado hasta el cual la confianza en el cálculo de un efecto es suficiente para justificar las recomendaciones. Esta definición tiene dos implicaciones importantes. En primer lugar, los grupos de expertos responsables de establecer directrices deben hacer juicios sobre la calidad de la evidencia relativa al contexto específico en el que se aplica. En segundo lugar, puesto que las revisiones sistemáticas no sirven para hacer recomendaciones, o como mínimo no deberían servir para ello, requieren una definición diferente. Para las revisiones sistemáticas, la calidad de la evidencia refleja el grado de confianza de que el cálculo del efecto es correcto.

EL DISEÑO DEL ESTUDIO ES IMPORTANTE PARA DETERMINAR LA CALIDAD DE LA EVIDENCIA

Al igual que con los sistemas iniciales de clasificación de la calidad de la evidencia[4], la estrategia GRADE empieza con el diseño del estudio. En cuanto a las recomendaciones que abordan estrategias alternativas de tratamiento, con respecto a los problemas de establecer el pronóstico o la precisión de los exámenes diagnósticos, los ensayos aleatorizados suelen proporcionar pruebas más potentes que los estudios obser-

vacionales. Si son rigurosos, éstos proporcionan evidencias más potentes que las series de casos no controlados. En la estrategia GRADE de la calidad de la evidencia, los ensayos aleatorizados sin limitaciones importantes constituyen evidencias de alta calidad, y los estudios observacionales sin especiales puntos fuertes o limitaciones importantes, evidencias baja calidad.

CINCO LIMITACIONES QUE PUEDEN REDUCIR LA CALIDAD DE LA EVIDENCIA

La estrategia GRADE incluye la separación de las valoraciones de la calidad de la evidencia de cada variable importante para los pacientes e identifica cinco factores que pueden reducirla (v. cuadro)[5]. Estos factores pueden reducir el nivel de calidad de los estudios observacionales y los ensayos aleatorizados controlados.

Limitaciones del estudio

La confianza en las recomendaciones disminuye si los estudios adolecen de importantes limitaciones que puedan sesgar sus cálculos del efecto del tratamiento[6]. Estas limitaciones son la ausencia de ocultación de la asignación; la ausencia de enmascaramiento, en particular si las variables son subjetivas y su evaluación es muy propensa al sesgo; la falta de seguimiento de un número importante de participantes; la falta de cumplimiento de un análisis por intención de tratar; la interrupción del estudio antes de la fecha planificada debido a la detección de un beneficio[7]; o la imposibilidad de describir las variables (de forma característica, aquellas para las que no se observó un efecto).

Por ejemplo, la mayor parte de ensayos aleatorizados que examinan el impacto relativo de la resección tumoral amplia de referencia con respecto a los pro-

Factores que intervienen en la decisión sobre la calidad de la evidencia

Factores que pueden disminuir la calidad de la evidencia:
- Limitaciones del estudio.
- Falta de coherencia de los resultados.
- Carácter indirecto de la evidencia.
- Imprecisiones.
- Sesgo de publicación.
- Factores que pueden aumentar la calidad de la evidencia.
- Gran magnitud del efecto.
- Factores de confusión verosímiles, que reducirían el efecto demostrado.
- Gradiente dosis-respuesta.

Tabla 1 | Perfil de evidencias GRADE del impacto de las alternativas quirúrgicas del cáncer de páncreas a partir de una revisión sistemática y un metaanálisis de ensayos aleatorizados controlados en pacientes hospitalizados para pancreaticoduodenectomía con conservación del píloro, con respecto al procedimiento estándar de Whipple para cáncer de páncreas o perioampular (Karanicolas y cols.[11])

Evaluación de la calidad						Resumen de los hallazgos			
N.º estudios (n.º participantes)	Limitaciones* del estudio	Coherencia	Carácter directo	Precisión	Sesgo de publicación	Efecto relativo (IC del 95 %)	Mejor cálculo en el grupo de riesgo Whipple	Efecto absoluto (IC del 95 %)	Calidad
Mortalidad 5 años:									
3 (229)	Graves (–1)	Incoherencia no importante	Directo	Imprecisión Incoherencia	Improbable	0,98 (de 0,87 a 1,11)	82,5%	20 menos/1.000; de 120 menos a 80 más	+++, moderada
Mortalidad hospitalaria:									
6 (490)	Graves (–1)	Incoherencia no importante	Directo	Imprecisión (–1) ‡	Improbable	0,40 (de 0,14 a 1,13)	4,9%	20 menos/1.000; (de 50 menos a 10 más)	++, baja
Transfusiones de sangre (unidades):									
5 (320)	Graves (–1)	Incoherencia no importante	Directo	Imprecisión	Improbable	--	2,45 unidades	–0,66 (de –1,06 a –0,25); favorece la conservación del píloro	+++, moderada
Fugas biliares:									
3 (268)	Graves (–1)	Incoherencia no importante	Directo	Imprecisión (–1) ‡	Improbable	4,77 (de 0,23 a 97,96)	0	20 más/1.000 20 menos a 50 más	++, baja
Estancia hospitalaria (días):									
5 (446)	Graves (–1)	Incoherencia no importante	Directo	Imprecisión (–1) ‡	Improbable	--	19,17 días	–1,45 (de -3,28 a 0,38); favorece conservación del píloro	++, baja
Retraso del vaciado gástrico:									
5 (442)	Graves (-1)	Heterogeneidad no explicada (–1)§	Directo	Imprecisión (–1)‡	Improbable	1,52 (de 0,74 a 3,14)	25,5%	110 más/1.000; de 80 menos a 290 más	+, muy baja

*Ocultación de la asignación poco clara en todos los estudios, pacientes enmascarados sólo en un estudio, evaluadores de las variables no enmascarados en ningún estudio; pérdidas del seguimiento > 20 % en tres estudios, no analizados usando el principio de la intención de tratar en un estudio.
+ Riesgos relativos (intervalos de confianza del 95 %) basados en modelos de efectos aleatorios.
‡ El intervalo de confianza incluye un posible beneficio de ambas estrategias quirúrgicas.
§12 = 72,6 %, $p = 0,006$.

cedimientos de Whipple modificados para el carcinoma pancreático tuvieron las limitaciones de la falta de ocultación óptima, la ausencia de un posible enmascaramiento de los pacientes y los responsables de adjudicar las variables, y las pérdidas sustanciales del seguimiento. Por tanto, la calidad de la evidencia para cada una de las variables importantes tan sólo fue moderada (tabla 1).

Falta de coherencia de los resultados

Los cálculos del efecto del tratamiento que difieren ampliamente (heterogeneidad o variabilidad en los resultados) entre distintos estudios son indicativos de diferencias reales en el efecto del tratamiento. La variabilidad puede deberse a diferencias en las poblaciones (p. ej., los fármacos pueden producir efectos relativos más amplios en poblaciones con peor salud), las intervenciones (p. ej., mayores efectos con dosis más altas de los fármacos) o los resultados (p. ej., disminución del efecto del tratamiento con el tiempo). Cuando existe heterogeneidad pero los investigadores no identifican una explicación verosímil, la calidad de la evidencia disminuye.

Por ejemplo, los ensayos aleatorizados sobre estrategias alternativas al procedimiento de Whipple depararon cálculos de los efectos sobre el vaciado gástrico que difirieron ampliamente, lo que disminuyó todavía más la calidad de la evidencia (fig. 2).

Carácter indirecto de las evidencias

Los responsables de elaborar directrices se enfrentan a dos tipos de carácter indirecto de las evidencias. El primero se presenta cuando, por ejemplo, se considera el uso de uno de dos fármacos activos. Aunque es posible que no se disponga de comparaciones aleatorizadas de los fármacos, en los ensayos aleatorizados se puede haber comparado ambos fármacos con un placebo, por separado. Estos ensayos permiten hacer comparaciones indirectas de la magnitud del efecto de ambos fármacos. Esta evidencia es de menor calidad que la que habría proporcionado una comparación directa entre ellos.

El segundo tipo de carácter indirecto de las evidencias son las diferencias entre la población, la intervención, la comparación de la intervención y el resultado de interés, y las incluidas en los estudios pertinentes. La tabla 2 presenta ejemplos de cada uno de ellos.

Imprecisiones

Cuando los estudios incluyen un número relativamente reducido de pacientes y pocos acontecimientos (y, por tanto, sus intervalos de confianza son amplios), el grupo de expertos responsable de formular las directrices juzgará que la calidad de la evidencia es menor. Por ejemplo, la mayor parte de variables de los procedimientos alternativos al de Whipple incluyen tanto efectos importantes como ningún efecto en absoluto, y algunos incluyen diferencias importantes en las dos direcciones (tabla 1).

Sesgo de publicación

La calidad de la evidencia disminuirá si los investigadores no publican los estudios (de forma característica, los que no revelan ningún efecto). La situación prototípica que debe suscitar sospecha de sesgo de publicación ocurre cuando la evidencia publicada se limita a un número reducido de ensayos, todos ellos financiados por la industria farmacéutica.

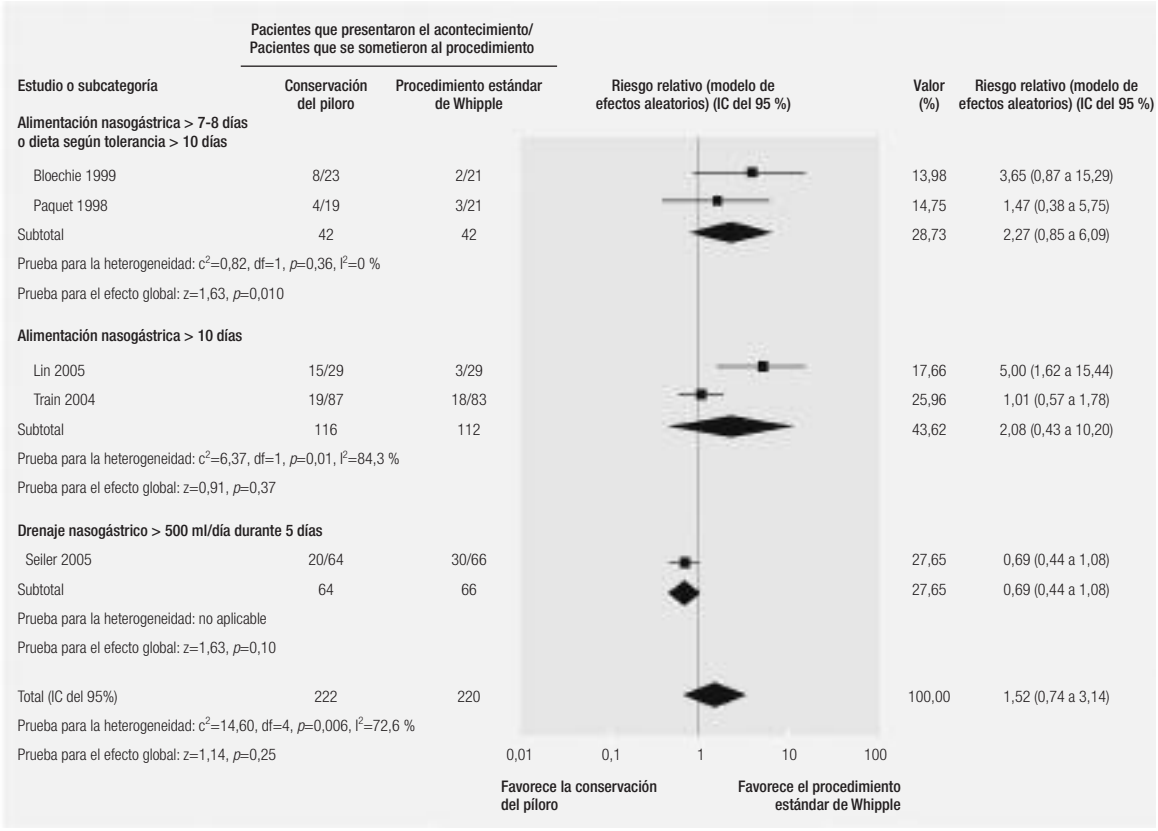


Fig. 2 | Efecto del retraso del vaciado gástrico de la pancreaticoduodenectomía con conservación del píloro con respecto al procedimiento de Whipple estándar para el adenocarcinoma de páncreas

Tabla 2 | La calidad de la evidencia es más débil si las comparaciones en los ensayos son indirectas

Pregunta de interés	Causa del carácter indirecto
Eficacia relativa del alendronato y el risedronato en la osteoporosis	Comparación indirecta: los ensayos aleatorizados comparan el alendronato y el risedronato con un placebo por separado, pero no se han realizado ensayos que comparen ambos fármacos
Oseltamivir como profilaxis de la gripe aviar causada por virus A de la gripe (HN1)	Diferencias en la población: se han realizado ensayos aleatorizados sobre el tratamiento con oseltamivir de la gripe estacional, pero no de la aviar
Cribado mediante sigmoidoscopia para la prevención de la mortalidad por cáncer de colon.	Diferencias en la población;: los ensayos aleatorizados sobre cribado de sangre oculta en heces proporcionan pruebas indirectas, que se basan en la posible eficacia de este la sigmoidoscopia
Elección de fármacos para la esquizofrenia	Diferencias en el comparador: las series de ensayos que comparan los neurolépticos de más reciente aparición con dosis fijas de haloperidol (20 mg) proporcionan pruebas indirectas sobre una posible comparación entre dichos fármacos y las dosis flexibles y más bajas de haloperidol que suelen prescribir los médicos
Rosiglitazona para la prevención de las complicaciones diabéticas en pacientes con riesgo alto de la enfermedad	Diferencias en el resultado: un ensayo aleatorizado demuestra un retraso en el desarrollo de diabetes bioquímica con rosiglitazona, pero no tiene la potencia suficiente para abordar las complicaciones diabéticas

TRES FACTORES QUE PUEDEN AUMENTAR LA CALIDAD DE LA EVIDENCIA

Aunque los estudios observacionales suelen proporcionar evidencias de baja calidad, pese a ser realizados apropiadamente, en circunstancias excepcionales pueden brindar evidencias de calidad moderada o, incluso, alta (v. cuadro)[8].

Cuando los estudios observacionales metodológicamente potentes deparan cálculos amplios o muy amplios y homogéneos de la magnitud del efecto de un tratamiento, podemos confiar en sus resultados. En dichas situaciones, es probable que los estudios observacionales proporcionen una sobrestimación del efecto real, pero poco probable que las deficiencias en el diseño del estudio expliquen todo el beneficio evidente.

Cuanto mayor es la magnitud del efecto, más potente es la evidencia. Por ejemplo, un metaanálisis de estudios observacionales reveló que el uso de cascos para bicicleta redujo en un amplio margen el riesgo de traumatismos craneales en los ciclistas afectados por una colisión (cociente de probabilidades = 0,31, intervalo de confianza del 95 % de entre 0,26 a 0,37)[9]. Este efecto amplio es indicativo de una valoración de la evidencia de calidad moderada. En un metaanálisis de estudios observacionales que evaluó el impacto de la profilaxis con warfarina en la cirugía para valvuloplastia cardiaca, se encontró que el riesgo relativo de tromboembolia con warfarina era de 0,17 (intervalo de confianza del 95 % de 0,13 a 0,24). Este efecto muy amplio es indicativo de una valoración de evidencia de alta calidad.

CONCEPTOS BÁSICOS

La formulación de directrices debe incluir una pregunta clara con la especificación de todos los resultados relevantes para los pacientes.
El sistema GRADE ofrece cuatro niveles de calidad de la evidencia: alto, moderado, bajo y muy bajo.
Los ensayos aleatorizados empiezan como evidencia de calidad elevada y los estudios observacionales, como evidencia de baja calidad.
La calidad puede reducirse como consecuencia de limitaciones en el diseño o la implementación del estudio, las imprecisiones en los cálculos (intervalos de confianza amplios), la variabilidad de los resultados, el carácter indirecto de la evidencia o el sesgo de publicación.
La calidad puede aumentar debido a una magnitud muy amplia del efecto o un gradiente de dosis-respuesta, y también si todos los sesgos verosímiles redujesen el efecto aparente del tratamiento.
Las variables clínicas determinan la calidad de la evidencia global.
Los perfiles de evidencia proporcionan resúmenes transparentes y simples.

La existencia de un gradiente dosis-respuesta o de una situación en la que todos los sesgos verosímiles disminuyeran la magnitud del efecto también aumentaría la calidad de la evidencia.

LAS VARIABLES DECISIVAS DETERMINAN LA VALORACIÓN DE LA CALIDAD DE LA EVIDENCIA ENTRE VARIABLES

Las recomendaciones dependen de la evidencia de diversas variables relevantes para el paciente y de la calidad de la evidencia para cada una de ellas. ¿Cómo debe valorarse la calidad de la evidencia entre variables si su calidad difiere? Esto es, precisamente, lo que se observó en el ejemplo del procedimiento de Whipple, en el que la calidad de evidencia varió desde moderada a muy baja.

La estrategia GRADE sugiere que los responsables de formular las directrices consideren que la calidad de la evidencia entre variables es la asociada a la variable decisiva con la evidencia de calidad más baja. Por tanto, para el ejemplo del procedimiento de Whipple, si los responsables de las recomendaciones hubieran considerado que los problemas de vaciado gástrico eran decisivos, la valoración de la calidad de la evidencia entre variables hubiera sido muy baja. Si el vaciado gástrico fuera importante pero no decisivo, la valoración de la calidad entre variables sería baja (partiendo de los resultados de la mortalidad perioperatoria claramente decisiva), a pesar de la presencia de evidencias de calidad moderada sobre la supervivencia a los 5 años (tabla 1).

LOS PERFILES DE EVIDENCIAS PROPORCIONAN RESÚMENES TRANSPARENTES Y SIMPLES

Los médicos atareados necesitan resúmenes sobre la evidencia que sean concisos, transparentes y fáciles de entender. El proceso GRADE facilita la creación de resúmenes, como el de la tabla 2, que presenta el efecto relativo de la resección estándar con respecto a la más limitada para pacientes con carcinoma pancreático.

Conclusión

El sistema GRADE proporciona una metodología clara y exhaustiva para valorar y resumir la calidad de la evidencia en la que basar las recomendaciones sobre un tratamiento. Aunque siempre se requerirá el juicio clínico en cada paso, esta estrategia sistemática y transparente permite un examen minucioso y una discusión sobre dichos juicios.

AGRADECIMIENTOS

Contribuidores: Todos los autores, incluidos los miembros del GRADE Working Group, contribuyeron al desarrollo de las ideas del manuscrito y lo leyeron y aprobaron. GHG escribió el primer borrador y recopiló los comentarios de los autores y revisores para las versiones posteriores, y es el garante del artículo.

Todos los autores citados a continuación contribuyeron con sus ideas a la estructura y el contenido del artículo, proporcionaron ejemplos, revisaron los borradores del manuscrito y dieron su opinión.

Los miembros del GRADE Working Group son Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Margaret Haugh, Robin Harbour, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Merce Marzo, James Mason, Jacek Mrukowics, Susan Norris, Andrew D Oxman, Vivian Robinson, Holger J Schünemann, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams, y James Woodcock.

Financiación: El estudio no contó con financiación específica.

Conflictos de interés: Todos los autores participan en la divulgación del sistema GRADE, y el éxito de éste tiene una influencia positiva en su carrera académica. Los autores citados han compensaciones por dietas y por presentaciones que incluyeron una revisión de la estrategia GRADE para valorar la calidad de la evidencia y clasificar las recomendaciones. GHG es consultor de UpToDate; su actividad consiste en ayudar a la empresa a usar el sistema GRADE. HJS es *documents editor* y experto en metodología de la American Thoracic Society; una de sus funciones en estos cargos es contribuir a implementar el uso del sistema GRADE. HJS recibe la beca *The human factor, mobility and Marie Curie actions scientist reintegration European Commission: IGR 42192—GRADE.*

Procedencia y revisión por expertos: No solicitada; revisión externa por expertos.

1 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
2 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.
3 Schunemann H, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. *Health Res Policy Syst* 2006;5:4-22.
4 Fletcher SW, Spitzer WO. Approach of the Canadian Task Force to the periodic health examination. *Ann Intern Med* 1980;92(2 Pt 1):253-4.
5 Schunemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605-14.
6 Guyatt G, Cook D, Devereaux PJ, Meade M, Straus S. Therapy. In: Guyatt G, Rennie D, eds. *The users' guides to the medical literature: a manual for evidence-based clinical practice.* Chicago: AMA publications, 2002.
7 Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA* 2005;294:2203-9.
8 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;334:349-51.
9 Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database Syst Rev* 2000;(2):CD001855.
10 Cannegieter SC, Rosendaal FR, Briet E. Thromboembolic and bleeding complications in patients with mechanical heart valve prostheses. *Circulation* 1994;89:635-41.
11 Karanicolas PJ, Davies E, Kunz R, Briel M, Koka HP, Payne DM, et al. The pylorus: take it or leave it? Systematic review and meta-analysis of pylorus-preserving versus standard whipple pancreaticoduodenectomy for pancreatic or periampullary cancer. *Ann Surg Oncol* 2007;14:1825-34.

BMJ (Ed Esp) 2008; 2(8):452-454

VALORACIÓN DE LA CALIDAD DE LA EVIDENCIA Y LA FUERZA DE LAS RECOMENDACIONES

GRADE: De la evidencia a las recomendaciones

El sistema GRADE clasifica las recomendaciones de las directrices como fuertes o débiles. En el presente artículo, se examina el significado de estas descripciones y sus implicaciones para pacientes, médicos y responsables de establecer normativas

Éste es el tercero de una serie de cinco artículos que describe la estrategia Grading of Recommendations Assessment, Development and Evaluation (GRADE) para desarrollar y presentar recomendaciones de tratamiento para los pacientes. En el presente artículo abordamos cómo la estrategia GRADE sugiere a los médicos que interpreten la fuerza de una recomendación.

¿CUÁL ES EL SIGNIFICADO DE LA FUERZA DE UNA RECOMENDACIÓN?

La fuerza de una recomendación refleja el grado hasta el que podemos confiar en que los efectos deseados de una intervención sean superiores a los adversos. Los efectos deseados incluyen la disminución de la morbilidad y la mortalidad, la mejora de la calidad de vida, la reducción de la carga del tratamiento (como tener que tomar medicación o la incomodidad de las pruebas de laboratorio) y la disminución de los gastos en recursos. Las consecuencias indeseables incluyen los efectos adversos que producen un impacto perjudicial sobre la morbilidad, la mortalidad o la calidad de vida o un mayor uso de recursos.

Los sistemas de clasificación previos han usado hasta nueve categorías de fuerzas de las recomendaciones[1]. El sistema GRADE sólo tiene dos categorías; aunque, en este artículo, las caracterizaremos como fuertes y débiles, los grupos de expertos que elaboran directrices pueden seleccionar diferentes términos para caracterizar las dos categorías de fuerza. Cuando utilicen el sistema GRADE, pueden hacer recomendaciones firmes si confían en que los efectos deseables del cumplimiento de la recomendación son superiores a los indeseables. Las recomendaciones débiles indican que los efectos deseables del cumplimiento de una recomendación son, probablemente, mayores que los indeseables, pero el equipo de expertos tiene menos seguridad.

LAS RECOMENDACIONES FUERTES Y DÉBILES PROPORCIONAN UNA GUÍA ESPECÍFICA

La clasificación binaria del sistema GRADE de la fuerza de las recomendaciones proporciona una dirección clara para los pacientes, los médicos y los responsables de elaborar normativas. Las implicaciones de una recomendación fuerte son:

- Para los pacientes: en su situación, la mayoría desearía que se implementasen las acciones recomendadas y sólo una pequeña proporción no estaría de acuerdo con ello; los pacientes deben expresar a su médico su deseo de hablar de ello si no se les ofrece la intervención.
- Para los médicos: la mayor parte de los pacientes debe recibir los procedimientos recomendados.
- Para los responsables de elaborar las normas: en la mayoría de las situaciones, la recomendación puede adoptarse como norma.

Las implicaciones de una recomendación débil son:

- Para los pacientes: en su situación, la mayoría desearía que se implementasen las acciones recomendadas, pero algunos las rechazarían.
- Para los médicos: deben reconocer que cada paciente requiere una elección distinta y que han de ayudar al paciente a tomar una decisión sobre el tratamiento teniendo en cuenta sus valores y sus preferencias.
- Para los responsables de elaborar normativas: estas requerirán un debate detallado y la participación de la mayor parte de los interesados.

A medida que los médicos comprenden mejor la variabilidad de los valores y las preferencias de los pacientes, prestan más atención a las ayudas estructuradas para la toma de decisiones que facilitan este proceso[2]. Ante una recomendación fuerte, no es necesario el uso de una ayuda para la decisión: casi todos los pacientes informados efectuarán la misma elección. Una recomendación débil indica que una ayuda para la decisión podría ser útil.

Los directivos de los sistemas sanitarios están cada vez más interesados en garantizar la calidad de la asistencia. Las directrices nos ayudan a diferenciar las estrategias que constituyen la calidad de la asistencia de otras que son facultativas. El sistema GRADE proporciona guías claras sobre estas opciones: las opciones de tratamiento asociadas con recomendaciones fuertes (pero no con las débiles) son buenas candidatas a los criterios de calidad. Cuando una recomendación es débil, abordar con el paciente y su familia las ventajas relativas de las estrategias alternativas de tratamiento puede convertirse en un criterio de calidad.

CUATRO FACTORES CLAVE DETERMINAN LA FUERZA DE UNA RECOMENDACIÓN

El primer determinante de la fuerza de una recomendación es el equilibrio entre las consecuencias deseables e indeseables de las estrategias alternativas de tra-

Gordon H Guyatt
Professor, Department of Epidemiology and Biostatistics, McMaster University, Hamilton ON L8N (Canadá)
Andrew D Oxman
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004 St Olavs Plass, 0130 Oslo (Noruega)
Regina Kunz
Associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basilea, Hebelstrasse 10, 4031 Basilea (Suiza)
Yngve Falck-Ytter
Assistant professor, Division of Gastroenterology, Case Medical Center, Case Western Reserve University, Cleveland OH 44106 (Estados Unidos)
Gunn E Vist
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo (Noruega)
Alessandro Liberati
Associate professor, Universidad de Módena
y Reggio Emilia y Agenzia Sanitaria Regionale, Bolonia (Italia)
Holger J Schünemann
Professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, Roma (Italia)
Para el grupo de trabajo GRADE

Correspondencia:
G H Guyatt, *guyatt@mcmaster.ca*

Éste es el tercero de una serie de cinco artículos que explican el sistema GRADE para valorar la calidad de la evidencia y la fuerza de las recomendaciones

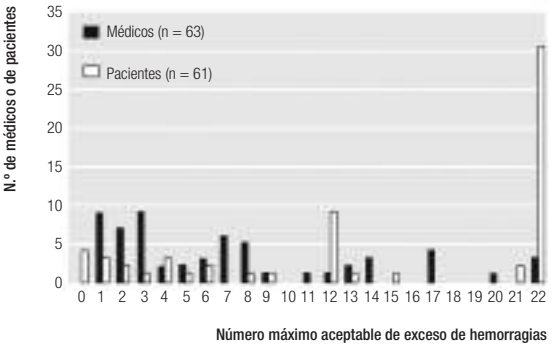


Fig. 1 | Variación de los umbrales de hemorragia gastrointestinal grave considerada aceptable por médicos y pacientes para la prevención de ocho ictus en 100 pacientes

tamiento (tabla 1). Por ejemplo, se considerará el uso de esteroides prenatales en las mujeres que van a dar a luz prematuramente. La administración de esteroides a la futura madre reduce el riesgo de síndrome del distrés respiratorio del recién nacido con efectos adversos, incomodidades y costes mínimos. Las ventajas de la administración de estos preparados son mucho mayores que sus inconvenientes, lo que indica lo apropiado de una recomendación fuerte.

Cuando las ventajas y los inconvenientes están equilibrados, la recomendación debe ser débil. Por ejemplo, consideremos a los pacientes con fibrilación auricular y un bajo riesgo de ictus. La warfarina puede reducir aún más este riesgo bajo, pero aumenta las inconveniencias y entraña un mayor riesgo de hemorragia. En estas circunstancias, es probable que la elección apropiada difiera para cada paciente.

El segundo factor que determina de la fuerza de una recomendación es la calidad de la evidencia. Si no estamos seguros de la magnitud de los beneficios y los riesgos de una intervención, hacer una recomendación fuerte a favor o en contra de unos procedimientos resulta problemático. Por ejemplo, las medias de compresión graduada producen un amplio efecto aparente en la reducción de la trombosis venosa profunda en los individuos que hacen viajes largos en avión. Sin embargo, los ensayos aleatorizados en los que se basa el efecto adolecieron de problemas metodológicos, ya que las técnicas para determinar la trombosis venosa profunda no eran reproducibles y no hubo ocultación. A pesar del amplio beneficio aparente, el uso de medias de compresión sólo merece una recomendación débil[3].

El tercer determinante de la fuerza de las recomendaciones es la incertidumbre o la variabilidad relativas a los valores o las preferencias. Dado que las estrategias alternativas de tratamiento siempre tendrán ventajas e inconvenientes y, por tanto, se compensarán, para determinar la fuerza de cualquier recomendación es importante el modo en que el equipo de expertos valora los beneficios, los riesgos y las inconveniencias de los tratamientos.

Consideremos, por ejemplo, la prevención de los ictus en los pacientes con fibrilación auricular. En comparación con la ausencia de tratamiento antitrombótico, la warfarina reduce el riesgo de ictus en aproximadamente el 65 %, pero aumenta el de hemorragia gastrointestinal grave. Devereaux y sus colaboradores preguntaron a 63 médicos y 61 pacientes cuántas hemorragias gastrointestinales graves tolerarían en 100 pacientes y estarían dispuestos a prescribir o tomar el fármaco para prevenir ocho ictus (cuatro menores y cuatro mayores) en 100 pacientes[4]. En la figura 1 se muestran los resultados de este estudio. Mientras que las respuestas de los médicos fueron muy diversas, la mayoría de pacientes otorgaron una elevada importancia a evitar un ictus y estuvieron dispuestos a aceptar un riesgo de hemorragia del 22 % para reducir su probabilidad de experimentar un ictus en un 8 %. Sin embargo, incluso entre los pacientes, la diversidad de valores y preferencias fue evidente: muy pocos pacientes estuvieron dispuestos a aceptar un ligero riesgo de hemorragia. Estos datos indican que una recomendación fuerte de la administración de warfarina sólo estaría justificada en los pacientes en riesgo elevado de ictus.

Compárese estos datos con la decisión a la que los médicos se enfrentan en el caso de una mujer embarazada con trombosis venosa profunda. El tratamiento con warfarina a las 6-12 semanas de embarazo entraña un riesgo para el feto de anomalías del desarrollo relativamente menores. La alternativa (la heparina) elimina el riesgo para el feto, pero este beneficio se obtiene a expensas de dolor, las incomodidades y el coste superior. La experiencia del médico es que una mayoría abrumadora de mujeres otorga un valor elevado a la prevención de las complicaciones fetales. Por tanto, a pesar de sus ventajas, la recomendación fuerte de sustituir la warfarina por heparina está justificada.

El determinante final de la fuerza de una recomendación es su coste, que varía mucho más con el tiempo y entre áreas geográficas que el resto de resultados. Los costes de los fármacos suelen caer en picado cuando las patentes caducan, y los del mismo fármaco difieren ampliamente en distintas regiones. Además, las implicaciones de los recursos varían ampliamente. Por ejemplo, la prescripción anual de un fármaco de coste elevado serviría para pagar el salario de una enfermera en los Estados Unidos, pero el de 30 enfermeras en China.

Por tanto, aunque unos costes mayores reduzcan la probabilidad de que pueda hacerse una recomendación fuerte de una intervención, el contexto de la recomendación es decisivo. Por esta razón, en la consideración de la distribución de los recursos, los expertos que elaboran directrices deben especificar el ámbito al que debe aplicarse cada recomendación.

ES POSIBLE QUE LAS RECOMENDACIONES FUERTES NO SEAN IMPORTANTES DESDE TODAS LAS PERSPECTIVAS

Si las consecuencias de la elección son relativamente poco importantes, algunos pacientes no se preocuparán por las recomendaciones, incluso si son fuertes. Esto es más probable si deben tomar muchos fármacos nuevos o se les sugieren muchos cambios de hábitos.

Tabla 1 | Factores determinantes de la fuerza de una recomendación

Factor	Comentario
Equilibrio entre efectos deseables y adversos	Cuanto mayor es la diferencia entre los efectos deseables e indeseables, mayor es la probabilidad de que esté justificada una recomendación fuerte
Calidad de la evidencia	Cuanto mayor es la calidad de la evidencia, mayor es la probabilidad de que se justifique una recomendación fuerte
Valores y preferencias	Cuanto más varían los valores y preferencias, o mayor es la incertidumbre con respecto a ellos, mayor es la probabilidad de que esté justificada una recomendación débil
Costes (asignación de recursos)	Cuanto mayores son los costes de una intervención (es decir, mayores los recursos consumidos), menor es la probabilidad de que esté justificada una recomendación fuerte

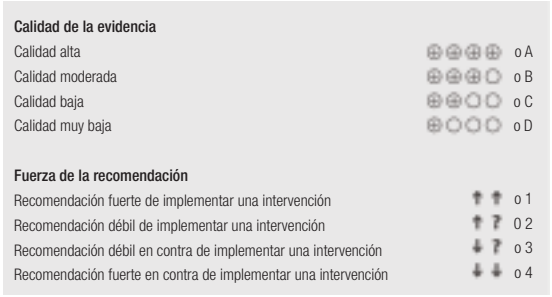


Fig. 2 | Representación de la calidad de la evidencia y la fuerza de las recomendaciones

PUNTOS CLAVE

La fuerza de las recomendaciones refleja el grado hasta el cual podemos confiar en que los efectos deseables de una intervención son superiores a los indeseables.
El sistema GRADE clasifica las recomendaciones como fuertes o débiles.
Una recomendación fuerte significa que la mayor parte de los pacientes informados elegiría el tratamiento recomendado y que los médicos pueden estructurar sus interacciones con los pacientes en consecuencia.
Una recomendación débil significa que las elecciones de los pacientes variarán de acuerdo con sus valores y preferencias, y que los médicos deben garantizar que la asistencia coincide con los valores y preferencias del paciente.
La fuerza de la recomendación está determinada por el equilibrio entre las consecuencias deseables e indeseables de las estrategias alternativas de tratamiento, la calidad de la evidencia, la variabilidad en los valores y preferencias y el uso de recursos.

Cuando establecen prioridades, los gobiernos y los responsables sanitarios también deben considerar otros factores distintos de la fuerza de una recomendación, como la prevalencia del problema de salud, las consideraciones de la equidad y loa posibilidad de mejorar la calidad de la asistencia, factores que pueden mejorar la influencia de una intervención sobre la salud de la población.

LAS RECOMENDACIONES DE USAR LAS INTERVENCIONES EN UN CONTEXTO DE INVESTIGACIÓN PUEDEN SER APROPIADAS

En ocasiones, los equipos de expertos que elaboran directrices deben decidir si recomiendan intervenciones prometedoras asociadas con efectos adversos o costes considerables y sin pruebas suficientes de beneficios que justifiquen su utilización. Pueden ser reacios a cerrar la puerta a una intervención de estas características o a proporcionar inapropiadamente una recomendación débil para su utilización. Su temor se hará realidad si las recomendaciones apropiadas en contra del uso de la intervención en la práctica clínica hacen que no se lleve a cabo una investigación adicional.

PUEDE SER CONVENIENTE PRESENTAR DE MODOS DIVERSOS LA CALIDAD DE LAS EVIDENCIAS Y LA FUERZA DE LAS RECOMENDACIONES

La mayoría de los equipos de expertos que elaboran directrices han empleado letras y cifras para resumir sus recomendaciones, pero los han usado de modo diferente, y esto podría inducir a confusión[5]. Las representaciones simbólicas de la calidad de la evidencia y de la fuerza de las recomendaciones son interesantes porque carecen de estos inconvenientes. Por otra parte, las organizaciones pueden tener buenas razones para seleccionar letras y cifras. Los médicos parecen sentirse muy cómodos con ello y son especialmente apropiados para la comunicación verbal.

El sistema GRADE ofrece representaciones simbólicas útiles y, para las organizaciones que desean usar cifras y letras, una representación de elección de cifras/letras adecuada para evaluar la calidad de la evidencia y los grados de la recomendación (fig. 2)[5].

AGRADECIMIENTOS

Los miembros del GRADE Working Group son Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Margaret Haugh, Robin Harbour, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Merce Marzo, James Mason, Jacek Mrukowics, Susan Norris, Andrew D Oxman, Vivian Robinson, Holger J Schünemann, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams, y James Woodcock.

Contribuidores: Todos los autores citados y otros miembros del GRADE Working Group contribuyeron al desarrollo de las ideas del manuscrito y lo leyeron y aprobaron. GHG escribió el primer borrador y recopiló los comentarios de los autores y revisores para las versiones posteriores, y es el garante del artículo.

Todos los autores citados contribuyeron a las ideas sobre la estructura y el contenido, proporcionaron ejemplos, revisaron los borradores del manuscrito y dieron su opinión al respecto.

Financiación: El estudio no contó con financiación.

Conflictos de interés: Todos los autores participan en la divulgación del sistema GRADE, cuyo éxito tiene una influencia positiva en su carrera académica. Los autores citados en el pie de autor han recibido dietas para los gastos de viaje y honorarios por presentaciones que incluyeron una revisión de la estrategia GRADE para valorar la calidad de la evidencia y clasificar las recomendaciones. GHG es consultor de UpToDate; su función consiste en ayudar a la empresa a usar el sistema GRADE. HJS es documents editor y experto en metodología de la American Thoracic Society; una de sus funciones es contribuir a implementar el uso del sistema GRADE. HJS recibe la beca «The human factor, mobility and Marie Curie actions scientist reintegration European Commission: IGR 42192—GRADE». AL ayuda a diferentes instituciones del Servicio Italiano de Salud a usar el sistema GRADE y lo ha implementado para elaborar recomendaciones clínicas en oncología a través de la beca N.º 249 (2005-7), Bando Ricerca Finalizzata, Ministero della Salute, Roma (Italia).

Procedencia y revisión por expertos: No solicitada; revisión externa por expertos.

1 Fleisher LA, Bass EB, McKeown P. Methodological approach: American College of Chest Physicians guidelines for the prevention and management of postoperative atrial fibrillation after cardiac surgery. *Chest* 2005;128:17-23S.
2 O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2003;(1):CD001431.
3 Geerts W, Ray JG, Colwell CW, Bergqvist D, Pineo GF, Lassen MR, et al. Prevention of venous thromboembolism. *Chest* 2005;128: 3775-6.
4 Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. *BMJ* 2001;323:1218-22.
5 Schunemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ* 2003;169:677-80.

BMJ (Ed Esp) 2008; 2(8):455-460

ANÁLISIS

VALORACIÓN DE LA CALIDAD DE LA EVIDENCIA Y FUERZA DE LAS RECOMENDACIONES

GRADE: Calificación de la calidad de la evidencia y la fuerza de las recomendaciones sobre pruebas y estrategias diagnósticas

El sistema GRADE puede servir para valorar la calidad de la evidencia y la fuerza de las recomendaciones sobre las pruebas o estrategias diagnósticas. En este artículo se explica de qué manera en este proceso se tienen en cuenta los resultados relevantes para el paciente

En este cuarto artículo de un total de cinco, explicamos cómo los expertos que elaboran directrices utilizan el sistema GRADE para evaluar la calidad de la evidencia y, basándose en ella, hacen recomendaciones sobre pruebas o estrategias diagnósticas. Aunque las recomendaciones sobre el diagnóstico se basan en los principios lógicos utilizados en las recomendaciones de otras intervenciones, plantean retos singulares. En el presente artículo se explica por qué los expertos que elaboran directrices deben tener cautela al utilizar la evidencia sobre la exactitud de los estudios («exactitud del estudio») como base para las recomendaciones, y por qué la evidencia sobre la exactitud de los estudios es, a menudo, una evidencia de baja calidad para hacer recomendaciones.

LAS PRUEBAS DIAGNÓSTICAS CONTRIBUYEN DE DIVERSAS MANERAS A LA ASISTENCIA MÉDICA

Los médicos utilizan pruebas —entre ellas signos y síntomas, estudios por imágenes y análisis bioquímicos— para identificar trastornos biológicos, establecer un pronóstico, hacer el seguimiento de enfermedades y documentar diagnósticos[1]. Este artículo se centra en el diagnóstico: el empleo de pruebas para determinar si existe o no una enfermedad (como la tuberculosis), un trastorno concreto (como la deficiencia de hierro) o un síndrome (como el de Cushing).

Los médicos suelen utilizar pruebas diagnósticas como un paquete o estrategia. Por ejemplo, en la asistencia a los pacientes con cáncer pulmonar en principio operable, pueden proceder directamente a la toracotomía o aplicar una estrategia de estudios por imágenes del cerebro, el sistema óseo, el hígado y las glándulas suprarrenales, y el tratamiento dependerá de sus resultados. En consecuencia, en muchos casos puede considerarse la evaluación o la recomendación no solo con respecto a un estudio, sino a una estrategia diagnóstica. Al considerar una prueba o estrategia diagnóstica, Los expertos que elaboran directrices deberán comenzar por identificar a los pacientes, la intervención diagnóstica (estrategia), la comparación y los resultados de interés (recuadro)[2,3].

LA EXACTITUD DEL ESTUDIO ES UN INDICADOR INDIRECTO DE LOS RESULTADOS RELEVANTES PARA LOS PACIENTES

La principal contribución de este artículo es que presenta un marco de referencia para analizar la calidad de la evidencia sobre las pruebas diagnósticas teniendo en cuenta su repercusión en los resultados que son relevantes para el paciente («resultados relevantes para el paciente»). Por lo general, cuando los médicos piensan en pruebas diagnósticas, se centran en la exactitud (sensibilidad y especificidad); es decir, en la eficacia con que el estudio clasifica correctamente a los pacientes como portadores o no portadores de una enfermedad. No obstante, la suposición básica es que si se tiene una idea más clara sobre si el paciente presenta o no un determinado trastorno, se logrará un mejor resultado. En los enfermos que presentan cáncer pulmonar operable, se supone que las pruebas adicionales evitarán la morbimortalidad inicial inherente a una toracotomía innecesaria. El ejemplo de la tomografía computarizada para la arteriopatía coronaria que muestra el recuadro ilustra otra justificación común de un nuevo estudio: el reemplazo de otro estudio (tomografía computarizada coronaria en lugar de angiografía convencional) para evitar las complicaciones inherentes a una alternativa más cruenta y costosa[6].

La mejor manera de evaluar cualquier estrategia diagnóstica —y, sobre todo, la nuevas estrategias con una exactitud supuestamente superior— es un ensayo aleatorizado comparativo en el cual los investigadores aleatoricen a los pacientes a enfoques diagnósticos experimentales o de referencia y en el que se de-

A Holger J Schünemann
Professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, 00144 Rome, Italy and CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Canadá) L8N 3Z5
Andrew D Oxman
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo (Noruega)
Jan Brozek
Research fellow, Department of Epidemiology, Italian National Cancer Institute Regina Elena, 00144 Roma (Italia)
Paul Glasziou
Professor, Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford OX3 7LF (Reino Unido)
Roman Jaeschke
Clinical professor, Department of Medicine, McMaster University, 1200 Main Street West, Hamilton, Ontario (Canadá) L8N 3Z5
Gunn E Vist
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo (Noruega)
John W Williams
Jr Professor, Department of Medicine, Duke University and Durham VA Medical Center, Durham, NC 27705 (Estados Unidos)
Regina Kunz
Associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basilea (Suiza)

Continúan los autores en la siguiente página

Éste es el cuarto de una serie de cinco artículos que explican el sistema GRADE de evaluación de la calidad de la evidencia y la fuerza de las recomendaciones

Tabla 1 | Ejemplos e implicaciones de diversas situaciones relacionadas con las pruebas

Ejemplo de una nueva prueba y de una prueba o estrategia de referencia	Posible beneficio de la nueva prueba	Exactitud diagnóstica	
		Sensibilidad	Especificidad
Versión más breve de la prueba de demencia frente al miniexamen del estado mental original para el diagnóstico de demencia	Prueba más simple, menos tiempo	Igual	Igual
Tomografía computarizada helicoidal para cálculos frente a la urografía excretora (UE)	Detección de un mayor número de cálculos (pero más pequeños)	Mayor	Igual
Tomografía computarizada para la arteriopatía coronaria frente a la angiografía coronaria	Pruebas menos cruentas, pero se pasan por alto algunos casos	Levemente menor	Menor

Véase la explicación de los términos en el texto.

EJEMPLO DE UNA PREGUNTA CLÍNICA SENSATA

Ante una sospecha de arteriopatía coronaria, ¿puede sustituir la tomografía computadorizada helicoidal multicorte de las arterias coronarias a la angiografía coronaria cruenta convencional, a fin de reducir las complicaciones con tasas aceptables de falsos negativos asociadas a complicaciones coronarias y falsos positivos que conduzcan a tratamientos innecesarios y complicaciones?[4,5]

termine la mortalidad, la morbilidad, los síntomas y la calidad de vida (figura)[7].

Cuando se dispone de estudios de intervención diagnóstica —en condiciones ideales, ensayos aleatorizados comparativos, pero también estudios de observación— que comparan la repercusión de estrategias diagnósticas alternativas en los resultados relevantes para el paciente, los expertos que elaboran directrices clínicas pueden utilizar el sistema GRADE descrito en artículos previos de esta serie[12,13].

Cuando no se cuenta con tales estudios, los expertos deben basarse en estudios sobre la exactitud de las pruebas y hacer deducciones sobre su posible repercusión en los resultados relevantes para el paciente[14]. Las preguntas clave son si se reducirán los resultados falsos negativos (casos pasados por alto) y los falsos positivos (y en qué medida pueden incrementarse los positivos y los negativos reales), con qué exactitud se clasifican los pacientes similares o diferentes mediante estrategias de análisis alternativas y qué resultados se producen en los pacientes que se consideren casos o no portadores de la enfermedad. La tabla 1 presenta ejemplos que ilustran estas preguntas.

Jonathan Craig
Associate professor, Screening and Test Evaluation Program, School of Public Health, University of Sydney, Department of Nephrology, Children's Hospital at Westmead, Sydney (Australia)
Victor M Montori
Associate professor, Knowledge and Encounter Research Unit, Department of Medicine, Mayo Clinic College of Medicine, Rochester, MN 55905 (Estados Unidos)
Patrick Bossuyt
Professor, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam 1100 DE (Países Bajos)
Gordon H Guyatt
Professor, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (Canadá)L8N 3Z5
Para el grupo de trabajo GRADE

Correspondencia:
schunch@mcmaster.ca

EMPLEO DE PRUEBAS INDIRECTAS PARA DEDUCIR LA REPERCUSIÓN EN LOS RESULTADOS RELEVANTES PARA EL PACIENTE

Para deducir de los datos disponibles que la exactitud de una prueba o estrategia diagnóstica mejora los resultados relevantes para el paciente es necesario disponer de un tratamiento eficaz[1]. Como alternativa, aun cuando no se disponga de él, una prueba exacta puede resultar útil si reduce los efectos adversos relacionados con la prueba o la ansiedad, o si la confirmación de un diagnóstico mejora el bienestar de los pacientes gracias a la información que brinda para el pronóstico.

Por ejemplo, los resultados de las pruebas genéticas en la corea de Huntington, un trastorno resistente al tratamiento, pueden tranquilizar a un paciente si se le comunica que no sufre el trastorno o permitirle planificar su futuro sabiendo que lo presentará. La posibilidad de planificar equivale a un tratamiento eficaz y los beneficios de la planificación deben sopesarse considerando los inconvenientes de conocer un diagnóstico oportuno[15-17]. A continuación, se describen los factores que influyen en el equilibrio entre las

Enfoque en la exactitud

Resultados de los pacientes y supuesta repercusión sobre el tratamiento				Equilibrio entre los supuestos resultados, las complicaciones de las pruebas y el coste
Positivos reales	**Negativos reales**	**Falsos positivos**	**Falsos negativos**	
Supuesta influencia sobre los resultados relevantes para el paciente				Por lo general, el tiempo más breve y la exactitud similar de la prueba (y, por tanto, los desenlaces para el paciente) indicarían que las nuevas pruebas son útiles
Beneficio dudoso del diagnóstico y el tratamiento en una etapa precoz	Beneficio casi seguro, por la tranquilidad que brinda al paciente	Probable ansiedad y morbilidad por pruebas y tratamiento adicionales	Posible perjuicio por diagnóstico tardío	
Relación directa de la evidencia (resultados de la prueba) con respecto a los resultados relevantes para el paciente				
Incertidumbre relativa	Ninguna incertidumbre	Incertidumbre relativa	Incertidumbre importante	
Supuesta influencia sobre los resultados relevantes para el paciente				El menor número de complicaciones e inconvenientes con respecto con la UE indicarían que la nueva prueba es útil, pero no está claro que exista un equilibrio entre los efectos beneficiosos y adversos en vista de las consecuencias indeterminadas de identificar cálculos más pequeños
Cierto beneficio para los cálculos más grandes y beneficios menos claros para los cálculos más pequeños; puede resultar un tratamiento innecesario	Beneficio casi seguro, porque se evitan pruebas innecesarias	Probable perjuicio debido a la realización de pruebas cruentas innecesarias	Posible perjuicio para los cálculos grandes, que es menos claro para los pequeños, a pesar de que la realización de pruebas cruentas innecesarias por otras posibles causas de molestias representaría un perjuicio	
Relación directa de la evidencia (resultados de la prueba) con respecto a los resultados relevantes para el paciente				
Cierta incertidumbre	Ninguna incertidumbre	Ninguna incertidumbre	Incertidumbre importante	
Supuesta influencia sobre los resultados relevantes para el paciente				Las consecuencias indeseables de más falsos positivos y falsos negativos con la tomografía computarizada no son aceptables pese a la mayor tasa de complicaciones raras (infarto y defunción) y el mayor coste de la angiografía
Beneficio del tratamiento y el menor número de complicaciones	Beneficio, por la tranquilidad que brinda a los pacientes y el menor número de complicaciones	Perjuicio debido a por tratamientos innecesarios	Perjuicio debido al diagnóstico tardío o la lesión miocárdica	
Relación directa de la evidencia (resultados de la prueba) con respecto a los resultados importantes para el paciente				
Ninguna incertidumbre	Ninguna incertidumbre	Ninguna incertidumbre	Cierta incertidumbre	

consecuencias favorables y adversas, de acuerdo con la calidad de la evidencia. Para ello, se utiliza método simplificado que clasifica los resultados de las pruebas como verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

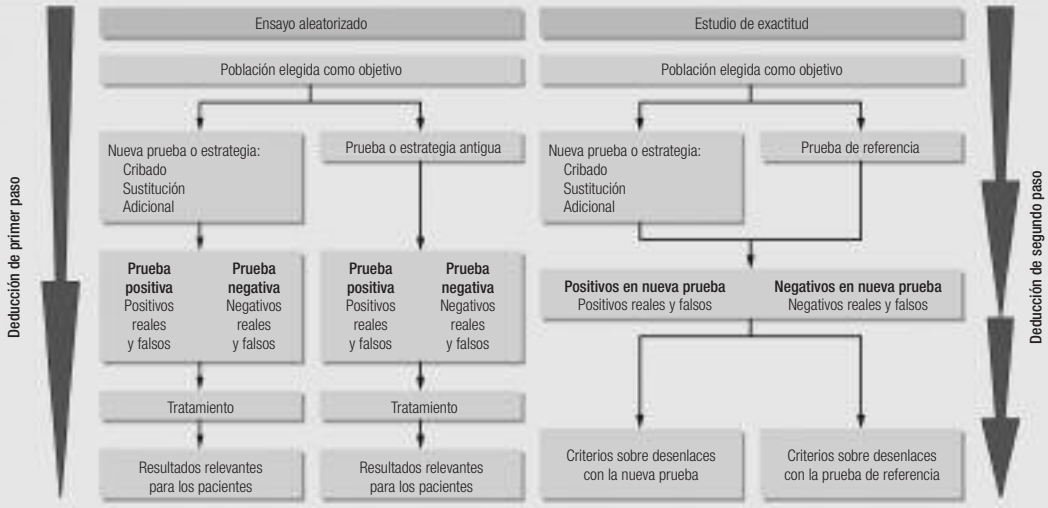
EVALUACIÓN DE LA CALIDAD DE LA EVIDENCIA SUBYACENTE

Diseño y limitaciones del estudio (riesgo de sesgo)

Las cuatro categorías de la calidad de la evidencia del sistema GRADE constituyen un gradiente de certidumbre en las estimaciones del efecto de una estrategia de pruebas diagnósticas sobre los resultados relevantes para el paciente[13]. La tabla 2 describe de qué manera el sistema GRADE aborda las dificultades específicas de evaluar la calidad de la evidencia con respecto a estrategias diagnósticas alternativas. Según se ha comentado, los ensayos aleatorizados de métodos diagnósticos alternativos representan el diseño de estudio ideal que proporciona información para las recomendaciones. No obstante, en el sistema GRADE, los estudios válidos sobre la exactitud de las pruebas

Tabla 2 | Factores que disminuyen la calidad de la evidencia de los estudios de exactitud diagnóstica y grado en que difieren de la evidencia de otras intervenciones

Factores que determinan y pueden disminuir la calidad de la evidencia	Explicaciones y diferencias derivadas de la calidad de la evidencia de otras intervenciones
Diseño del estudio	Criterios distintos para los estudios de exactitud; los estudios transversales o de cohortes con pacientes en los que existe una incertidumbre diagnóstica y la comparación directa de los resultados de las pruebas con una norma de referencia apropiada se consideran una evidencia de gran calidad, que puede cambiar a moderada, baja o muy baja en función de otros factores
Limitaciones (riesgo de sesgo)	Criterios distintos para los estudios de exactitud; debe incorporarse a pacientes consecutivos como una sola cohorte, sin clasificarlos según su estado patológico, y debe definirse claramente los procesos de selección y de remisión[7]. Deberán realizarse pruebas a todos los pacientes en la misma población para la nueva prueba y la norma de referencia bien descrita; los evaluadores no podrán conocer los resultados de la prueba alternativa y la norma de referencia
Carácter indirecto:	
Desenlaces	Criterios similares; a menudo, los grupos de expertos que evalúan las pruebas diagnósticas no disponen de evidencia directa sobre la repercusión en los desenlaces relevantes para el paciente. Basándose en estudios de pruebas diagnósticas sobre el equilibrio entre las supuestas influencias en los desenlaces relevantes para el paciente, deben deducir cualesquiera diferencias en los positivos y negativos reales y falsos en relación con las complicaciones y los costes de la prueba. Por tanto, los estudios de exactitud suelen proporcionar evidencia de baja calidad para las recomendaciones como consecuencia de la cualidad indirecta de los resultados, como ocurre con los resultados indirectos para los tratamientos
Poblaciones de pacientes, prueba diagnóstica, prueba de comparación y comparaciones indirectas	Criterios similares; la calidad de la evidencia puede reducirse si existen diferencias importantes entre las poblaciones estudiadas y aquellas a quienes está dirigida la recomendación (en pruebas previas, gama de enfermedades o trastornos concomitantes); si existen diferencias importantes en las pruebas estudiadas y en la destreza diagnóstica de personas que las aplican en estudios con respecto a los contextos en los cuales se aplicarán las recomendaciones, o si las pruebas se comparan por separado con una norma de referencia en diferentes estudios, y no directamente en los mismos estudios
Contradicciones importantes en los resultados del estudio	Criterios similares; para los estudios de exactitud, las contradicciones no explicables en cuanto a la sensibilidad, la especificidad o los índices de probabilidad (más que en cuanto al riesgo relativo o las diferencias medias) pueden reducir la calidad de la evidencia
Evidencia imprecisa	Criterios similares; para los estudios de exactitud, los intervalos de confianza amplios para los cálculos de la exactitud de la prueba o las tasas de positivos y negativos reales o falsos pueden reducir la calidad de la evidencia
Alta probabilidad de sesgo de publicación	Criterios similares; el riesgo elevado de sesgo de publicación (p. ej., evidencia de estudios pequeños para nuevas intervenciones o pruebas, o asimetría en la gráfica de embudo) puede reducir la calidad de la evidencia



Ejemplo
Ensayos aleatorizados comparativos (EAC) que analizaron una estrategia diagnóstica basándose en el empleo del péptido natriurético de tipo B (BNP), utilizado para ayudar al diagnóstico de la insuficiencia cardiaca, frente a ningún uso del BNP en pacientes que acudieron al servicio de urgencias con disnea aguda[8,9]. Posteriormente, se comprobó que el grupo aleatorizado a recibir BNP fue hospitalizado por un período más breve y a un coste menor, sin que esto aumentase la morbilidad ni la mortalidad.

Ejemplo
La evidencia uniforme de los estudios bien diseñados muestra indica que la tomografía computarizada (TC) helicoidal sin contraste da lugar a menos resultados falsos negativos que la urografía excretora (UE) en el diagnóstico de urolitiasis aguda sospechada[10]. Sin embargo, los cálculos ureterales que detecta la TC pero que «pasa por alto» la UE son más pequeños y, por tanto, es probable que sean expulsados con más facilidad. Puesto que no se dispone de EAC que evalúen los desenlaces de pacientes tratados por cálculos más pequeños, siguen existiendo dudas respecto a los beneficios para la salud de la reducción de los casos no detectados (falsos negativos) y el seguimiento de los datos nuevos no relacionados con los cálculos renales mediante el empleo de la TC[11]

Dos formas genéricas con las que evaluar una prueba o una estrategia diagnóstica. A la izquierda, los pacientes son aleatorizados a una prueba o estrategia nuevas o a una prueba o estrategia antiguas. Los que dan un resultado positivo en la prueba (casos detectados) son aleatorizados (o fueron previamente aleatorizados) a recibir el mejor tratamiento disponible (no se muestra el segundo paso de aleatorización para el tratamiento). Los investigadores evaluaron y compararon los resultados relevantes para el paciente en todos los casos de los dos grupos[6]. A la derecha, los pacientes fueron sometidos a una nueva prueba y a una prueba de referencia (prueba diagnóstica o estrategia antiguas o de comparación). Los investigadores pueden, entonces, calcular la exactitud de la prueba con respecto a la de referencia (primer paso). Para evaluar la importancia que tiene esta información para los pacientes, los individuos con una prueba (o estrategia) diagnóstica positiva de cualquiera de los dos grupos son sometidos (o lo han sido en estudios previos) a tratamiento o a ningún tratamiento; a continuación, los investigadores evalúan y comparan los resultados relevantes para el paciente en todos los casos de los dos grupos (segundo paso).

Tabla 3 | Datos fundamentales de los estudios sobre exactitud diagnóstica. ¿Debería utilizarse la tomografía computadorizada helicoidal multicorte en lugar de la angiografía coronaria convencional* para diagnosticar la arteriopatía coronaria en una población con una probabilidad baja (20 %) previa a la prueba?[5]

Variable	Resultados de la prueba (IC del 95 %)
Sensibilidad acumulada	0,96 (de 0,94 a 0,98)
Especificidad acumulada	0,74 (de 0,065 a 0,84)
Índice de probabilidad positivo†	5,4 (de 3,4 a 8,3)
Índice de probabilidad negativo†	0,05 (de 0,03 a 0,09)

*Asumiendo que la norma de referencia, que es la angiografía, no produce resultados falsos positivos o falsos negativos.
†Índice de probabilidad promedio de Hamon y cols.[5]

también comienzan como una gran calidad en el marco de referencia diagnóstico. Sin embargo, tales estudios tienen limitaciones y, a menudo, proporcionan evidencia de baja calidad para las recomendaciones, ya que las pruebas que brindan sobre la repercusión que tienen en los resultados relevantes para el paciente son indirectas.

Los estudios válidos sobre la exactitud de las pruebas diagnósticas incluyen a pacientes representativos y consecutivos sobre quienes existe una incertidumbre diagnóstica legítima; es decir, la clase de pacientes a quienes los médicos realizarían la prueba durante el curso de su ejercicio clínico habitual. Si los estudios no cumplen con este criterio —y, por ejemplo, incorporan casos graves e individuos de referencia sanos—, es probable que la exactitud manifiesta de un estudio sea engañosamente elevada[18,19].

Los estudios válidos son los que comparan la prueba o las pruebas que se están considerando y una norma de referencia apropiada (denominada, en ocasiones, «óptima»). Si los investigadores no realizan tal comparación en todos los pacientes, el riesgo de sesgo es mayor. Este riesgo aumenta más cuando las personas que llevan a cabo o interpretan la prueba conocen los resultados de la prueba de referencia, o viceversa. Los expertos que elaboran directrices pueden utilizar instrumentos disponibles para evaluar el riesgo de sesgo en estudios en los que se evalúan la exactitud de las pruebas diagnósticas, y pueden reducir el grado de la calidad de la evidencia si existen limitaciones importantes[20-22].

LA VALORACIÓN DIRECTA

La valoración directa es, tal vez, el aspecto más difícil para los especialistas que elaboran directrices y recomendaciones sobre pruebas diagnósticas. Por ejemplo, un nuevo estudio puede ser más sencillo de realizar, conllevar menos riesgo y coste, pero puede dar lugar a falsos positivos y falsos negativos. Considérese las consecuencias de reemplazar la angiografía cruenta por la tomografía computarizada coronaria para el diagnóstico de la arteriopatía coronaria (tablas 3 y 4).

Los resultados verdaderos positivos conducirán a la administración de tratamientos de eficacia conocida (fármacos, angioplastia y endoprótesis, procedimiento de derivación coronaria), mientras que los verdaderos negativos evitarán a los pacientes los posibles efectos adversos de la prueba estándar de referencia. Sin embargo, los falsos positivos producirán efectos adversos (fármacos e intervenciones innecesarias, incluida la posibilidad de una angioplastia de seguimiento) sin un beneficio manifiesto, y los falsos negativos evitarán que no se prescriban intervenciones disponibles que ayudarían a reducir el riesgo posterior de complicaciones coronarias.

Por consiguiente, es relativamente evidente que minimizar los falsos positivos y los falsos negativos proporciona beneficios a los pacientes. La repercusión de los resultados de las pruebas no concluyentes es menos clara, pero, sin duda, estas pruebas no son convenientes. Asimismo, las complicaciones de la angiografía cruenta —infarto y muerte— pese a ser raras, son indudablemente importantes. Cuando los expertos que elaboran directrices sopesan las consecuencias favorables y adversas de las pruebas diagnósticas, deben considerar la importancia de estas consecuencias para los pacientes. En el caso de los pacientes con una probabilidad relativamente baja de arteriopatía coronaria, la tomografía computarizada produce un gran número de falsos positivos que generan una ansiedad innecesaria y pruebas adicionales (tabla 4), y hacen que se pase por alto el 1% (falsos negativos) de los pacientes con arteriopatía coronaria.

Al considerar aspectos del diagnóstico, los expertos que elaboran directrices afrontan la misma serie de retos con relación a los datos indirectos que los especialistas que hacen recomendaciones para otras intervenciones[2]. La exactitud de la prueba puede variar en diferentes poblaciones de pacientes, de manera que los expertos deben considerar en qué grado las poblaciones incluidas en los estudios corresponden a la población a la que está dirigida la recomendación. Asimismo, deben considerar la posible equiva-

Tabla 4 | Consecuencias de los datos fundamentales de los estudios de exactitud diagnóstica. ¿Debería utilizarse la tomografía computadorizada helicoidal multicorte en lugar de la angiografía coronaria convencional* para diagnosticar arteriopatía coronaria en una población con una probabilidad baja (20 %) previa a la prueba?[6]

Consecuencias	N.º por cada 1.000 pacientes	Importancia†
Positivos reales ‡	192	8
Negativos reales §	592	8
Falsos positivos¶	208	7
Falsos negativos**	8	9
Resultados no concluyentes††§§	–	5
Complicaciones‡‡§§	–	5
Costes§§	–	5

Todos los resultados por 1.000 pacientes sometidos a prueba para una prevalencia de 20% y los índices de probabilidad que se muestran en la tabla 3.
*Asumiendo que la norma de referencia, es decir, la angiografía, no produce falsos positivos o falsos negativos.
†En una escala de 9 puntos, el sistema GRADE recomienda clasificar los resultados como no relevantes (calificación 1-3), relevantes (4-6) y críticos (7 a 9) para una decisión[13,18,19].
‡Relevantes porque obligan a usar fármacos, angioplastia y endoprótesis y procedimiento de derivación.
§Relevantes porque evitan intervenciones innecesarias que se acompañan de efectos adversos para los pacientes.
¶Relevantes porque los pacientes están expuestos innecesariamente a posibles efectos adversos de fármacos y procedimientos cruentos.
**Relevantes porque aumentan el riesgo de complicaciones coronarias, ya que no se prescriben tratamientos eficaces.
††Resultados de la prueba no interpretables, indeterminados o intermedios; relevantes porque generan ansiedad, incertidumbre respecto a cómo proceder, pruebas adicionales y posibles consecuencias negativas del tratamiento o de la ausencia de tratamiento.
‡‡No se comunican de forma fiable; relevantes porque, aunque son raras, pueden ser graves.
§§Aunque los datos de estas categorías no se muestran para facilitar los cálculos o porque no se conocen exactamente de acuerdo con los datos disponibles, son relevantes.

CONCEPTOS BÁSICOS

Al igual que en otras intervenciones, el sistema GRADE para determinar el grado de la calidad de la evidencia y la fuerza de las recomendaciones con respecto a las pruebas o estrategias diagnósticas es un enfoque exhaustivo y claro para formular dichas recomendaciones.

Los estudios transversales o de cohortes proporcionan pruebas de alta calidad sobre la exactitud de las pruebas diagnósticas.

Sin embargo, la exactitud de las pruebas diagnósticas es un indicador indirecto de los resultados relevantes para el paciente, de manera que tales ensayos proporcionan, a menudo, evidencia de baja calidad para las recomendaciones sobre las pruebas diagnósticas, aun cuando no tengan limitaciones importantes.

Para deducir de los datos sobre la exactitud que una prueba o estrategia diagnóstica mejora los resultados relevantes para el paciente es necesario que se disponga de un tratamiento eficaz, se reduzcan los efectos adversos relacionados con la prueba o de la ansiedad o mejore el bienestar de los pacientes gracias a que se les informa de su pronóstico.

Por tanto, se necesitan criterios para evaluar el carácter directo de los resultados de la prueba en relación con las consecuencias de las recomendaciones diagnósticas relevantes para los pacientes.

lencia de las nuevas pruebas y las pruebas de referencia en relación con las utilizadas en las circunstancias en las cuales se hacen las recomendaciones. Por último, al evaluar dos o más nuevas pruebas o estrategias diagnósticas, deben considerar si estas estrategias diagnósticas han sido comparadas directa (en un estudio) o indirectamente (en estudios diferentes) con una norma de referencia común[25-27].

¿CÓMO LLEGAR A LA ESENCIA DE LA CALIDAD DE UN ESTUDIO?

En la tabla 5 se muestra el resumen de la evidencia y la evaluación de la calidad para todos los desenlaces importantes de la angiotomografía computarizada como sustitución de la angiografía cruenta. La incertidumbre sobre el carácter directo de la evidencia (para los resultados de la prueba) de resultados relevantes para el paciente es baja o nula por lo que respecta a verdaderos positivos, falsos positivos y verdaderos negativos (tabla 1). Sin embargo, cierta incertidumbre sobre el grado en el que la exactitud de la prueba tendrá consecuencias perjudiciales en los desenlaces relevantes para el paciente por lo que respecta a los falsos negativos llevó a reducir el grado de la calidad de la evidencia de elevado a moderado (tabla 5, v. www.bmj.com). La heterogeneidad inexplicable en los resultados de los diferentes ensayos redujo más todavía la calidad de la evidencia para todas las variables. La incertidumbre importante sobre la repercusión de los resultados falsos negativos en los resultados relevantes para el paciente habría llevado a una reducción del grado de la calidad de la evidencia de elevado a bajo para los otros ejemplos que se muestran en la tabla 1.

CÓMO SE LLEGA A LA RECOMENDACIÓN

Sopesar los resultados supuestamente relevantes para el paciente a causa de los positivos y negativos verdaderos y falsos con las complicaciones de la prueba determinará si los expertos que elaboran las directrices harán una recomendación a favor o en contra de utilizar una determinada prueba diagnóstica[12]. Otros factores que influyen en la fuerza de una recomendación son la calidad de la evidencia, la incertidumbre sobre los valores y preferencias inherentes a las pruebas diagnósticas y los resultados supuestamente relevantes para el paciente, así como su coste.

La angiotomografía coronaria evita las consecuencias adversas de la angiografía cruenta, que pueden incluir un infarto de miocardio y la muerte. Sin embargo, estas consecuencias son muy raras. Por consiguiente, los expertos que preparen las directrices, al evaluar la angiotomografía coronaria como sustitución de la angiografía coronaria, podrían, pese a su menor coste, hacer una recomendación débil para que no se utilice en lugar de la angiografía coronaria cruenta.

Esta recomendación obedece al gran número de falsos positivos y el riesgo de pasar por alto a pacientes con arteriopatía coronaria que podrían ser tratados eficazmente (falsos negativos). También se fundamenta en que la evidencia para el empleo de la nueva prueba es de baja calidad y en la consideración de los valores. Pese a la preferencia general por las pruebas menos cruentas con riesgos de complicaciones más bajos, es probable que la mayoría de los pacientes prefiriese el método más cruento (angiografía), dados los riesgos relacionados con los falsos positivos y negativos.

CONCLUSIÓN

Al igual que con otras recomendaciones terapéuticas, el sistema GRADE para calificar la calidad de la evidencia y la fuerza de las recomendaciones sobre pruebas diagnósticas constituye un método exhaustivo y transparente para formular dichas recomendaciones. Reconocer que los resultados de la prueba son indicadores indirectos de los resultados relevantes para el paciente es esencial en este enfoque. La aplicación del método exige un cambio en la manera de pensar de los médicos a fin de que reconozcan claramente que, con independencia de su exactitud, las pruebas diagnósticas son de utilidad sólo si mejoran los desenlaces para los pacientes.

AGRADECIMIENTOS

Agradecemos a las numerosas personas y organizaciones que han contribuido a la evolución del sistema GRADE mediante la financiación de reuniones de trabajo y sus comentarios al trabajo descrito en este artículo.

Los miembros del grupo de trabajo de GRADE son Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Robin Harbour, Margaret Haugh, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Nicola Magrini, Merce Marzo, James Mason, Jacek Mrukowics, Andrew D Oxman, Susan Norris, Vivian Robinson, Holger J Schünemann, Jane Thomas, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams y James Woodcock.

Colaboradores: Todos los autores enumerados y el resto de miembros del grupo de trabajo GRADE contribuyeron al desarrollo de las ideas vertidas en el manuscrito, el cual leyeron y aprobaron. HJS redactó el primer borrador y compaginó los comentarios de los autores y los revisores en las versiones sucesivas. Todos los otros autores aportaron ideas sobre la estructura y el contenido del artículo, y dieron sus opiniones. HJS es el garante.

Financiación: Este trabajo fue financiado parcialmente por la subvención «The human factor, mobility and Marie Curie Actions Scientist Reintegration» de la Comisión Europea: IGR 42192-«GRADE» a HJS.

Conflictos de intereses: Los autores son miembros del grupo de trabajo de GRADE. El trabajo con este grupo tuvo, probablemente, una influencia favorable en las carreras aca-

démicas de algunos o de todos los autores y los miembros del grupo. Los autores enunciados han recibido compensaciones por gastos y honorarios por presentaciones, entre ellas un análisis del método GRADE para calificar la calidad de la evidencia y la fuerza de las recomendaciones. GHG es asesor de UpToDate; su trabajo incluye ayudar a UpToDate en su aplicación del sistema GRADE. HJS es editor de documentos y experto en metodología para la American Thoracic Society. Una de sus funciones en estos puestos es ayudar a implementar el empleo del sistema GRADE; además, colabora con la implementación de GRADE en organismos de todo el mundo. WMM ayuda a la implementación del sistema GRADE en diversas organizaciones profesionales norteamericanas no lucrativas.

1 Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.
2 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.
3 Mulrow C, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med* 1989;4:288-95.
4 Guyatt G, Montori V, Devereaux PJ, SchŸnemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club* 2004;140(1):A11-2.
5 Hamon M, Biondi-Zoccai GG, Malagutti P, Agostoni P, Morello R, Valgimigli M, et al. Diagnostic performance of multislice spiral computed tomography of coronary arteries as compared with conventional invasive coronary angiography: a meta-analysis. *J Am Coll Cardiol* 2006;48:1896-910.
6 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92.
7 Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000; 356:1844-7.
8 Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med* 2004;350:647-54.
9 Moe G, Howlett J, Januzzi JL, Zowall H, Canadian multicenter improved management of patients with congestive heart failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian prospective randomized multicenter IMPROVE-CHF study. *Circulation* 2007;115:3103-10.
10 Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med* 2002;40:280-6.
11 Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? *Can Assoc Radiol J* 2002;53:144-8.
12 Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schünemann HJ. Going from evidence to recommendations. *BMJ* 2008, doi: 10.1136/bmj.39493.646875.AE.
13 Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008, doi: 10.1136/bmj.39490.551019.BE.
14 Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006;144:850-5.
15 Maat-Kievit A, Vegter-van der Vlis M, Zoeteweij M, Losekoot M, van Haeringen A, Roos R. Paradox of a better test for Huntington's disease. *J Neurol Neurosurg Psychiatry* 2000;69:579-83.
16 Walker FO. Huntington's disease. *Semin Neurol* 2007;27:143-50.
17 Almqvist EW, Brinkman RR, Wiggins S, Hayden MR. Psychological consequences and predictors of adverse events in the first 5 years after predictive testing for Huntington's disease. *Clin Genet* 2003;64:300-9.
18 Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
19 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
20 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138:40-4.
21 Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;3:25.
22 Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006;6:9.
23 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
24 Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, 24 Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605-14.
25 Fletcher RH. Carcinoembryonic antigen. *Ann Intern Med* 1986; 104:66-73.
26 Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. *Am J Med* 1984;77:64-71.
27 Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation* 1990;81:815-20.

BMJ (Ed Esp) 2008; 2(8):461-464

ANÁLISIS

VALORACIÓN DE LA CALIDAD DE LA EVIDENCIA Y FUERZA DE LAS RECOMENDACIONES

GRADE: Incorporación de consideraciones sobre el empleo de recursos en la calificación de las recomendaciones

Los expertos que elaboran directrices no siempre opinan lo mismo sobre si el empleo de recursos debe influir en las decisiones que se toman sobre cada paciente. A medida que aumentan los costes de la atención médica, las consideraciones en torno al uso de recursos parecen más convincentes, y este es un reto que puede resultar difícil para los expertos

En esta última parte de una serie de artículos en que se describe el enfoque Grading of Recommendations Assessment, Development and Evaluation (GRADE) para establecer recomendaciones, analizamos de qué manera los expertos que preparan directrices y los médicos pueden incorporar cuestiones relacionadas con el empleo de recursos a las recomendaciones y el ejercicio clínico. Las recomendaciones clínicas implican, inevitablemente, decisiones sobre la asignación de recursos; a tales decisiones se les suele denominar *costes*. En este artículo, se abordan algunos de los retos implícitos en la consideración de los costes, se explican las razones que obligan a centrarse en el empleo de recursos más que en los costes y se analiza cómo pueden incorporarse las consideraciones sobre el empleo de recursos a las recomendaciones.

¿En qué difieren los costes de otras variables sanitarias?

- Los pacientes reciben beneficios para su salud y son los afectados por los desenlaces clínicos adversos, pero los costes sanitarios son compartidos por la sociedad en general (representada por el gobierno), los empresarios y los pacientes.
- Las actitudes en cuanto a si los costes deben influir en las decisiones del médico con respecto al tratamiento de cada paciente son variables.
- Los costes sanitarios pueden variar considerablemente entre áreas geográficas e incluso en el seno de cada una de ellas, y modificarse rápidamente.
- Lo que las sociedades pueden adquirir si descartan el empleo de recursos sanitarios (coste de oportunidad) varía ampliamente entre distintos países. Una dotación anual de un fármaco de coste elevado corresponde al salario de una enfermera en los Estados Unidos, pero en China permitiría pagar el salario de 30 enfermeras.
- Cuando los gastos asistenciales exigen recortar el gasto en otras partidas, las actitudes en torno a si es el sistema de salud, el erario público o la sociedad en general quien debe asumirlos son variables.
- Las cuestiones relacionadas con el uso de recursos tienen un alto componente político y pueden ocasionar conflictos de intereses a los grupos de expertos que elaboran directrices (p. ej., los expertos pueden tener vínculos con la industria o el gobierno).

Gordon H Guyatt
Professor, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, (Canadá) L8N 3Z5
Andrew D Oxman
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo (Noruega)
Regina Kunz
Associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, 4031 Basilea (Suiza)
Roman Jaeschke
Clinical professor, Department of Medicine, McMaster University, Hamilton, ON, (Canadá) L8N 3Z5
Mark Helfand
Professor of medicine, Portland VA Medical Center and OHSU Department of Medicine, Portland, Oregón 97201 (Estados Unidos)
Alessandro Liberati
Professor, Università di Modena e Reggio Emilia and Agenzia Sanitaria Regionale, Regione Emilia Romagna, 40127 Bolonia (Italia)
Gunn E Vist
Researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, 0130 Oslo (Noruega)
Holger J Schünemann
Associate professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, Roma (Italia)
Por el grupo de trabajo de GRADE

Correspondencia:
guyatt@mcmaster.ca

Este es el último de una serie de cinco artículos que explican el sistema GRADE de evaluación de la calidad de la evidencia y la fuerza de las recomendaciones. Puede consultarse más información al respecto en la versión publicada en www.bmj.com

LA CONSIDERACIÓN DEL COSTE COMO VARIABLE PLANTEA DIFICULTADES ESPECÍFICAS

En cierto sentido, el coste no es más que una variable potencialmente importante —como la mortalidad, la morbilidad y la calidad de vida— relacionada con formas alternativas de tratar los problemas del paciente. Además de estas variables clínicas, una intervención puede incrementar o disminuir los costes. Sin embargo, los costes implican algunos aspectos diferentes a los de otras variables (cuadro)[1]. En el presente análisis se analizan las implicaciones de estas diferencias, como la posible omisión legítima del coste como variable al considerar una recomendación de tratamiento.

DEBE UTILIZARSE UNA HOJA DE BALANCE PARA VALORAR LOS BENEFICIOS FRENTE A LOS COSTES

Pese a sus diferencias, los enfoques sobre el empleo de recursos son similares a otras variables, por cuanto las autoridades sanitarias necesitan calcular la diferencia entre el tratamiento y la referencia. Una hoja de balance es una forma sencilla pero eficiente de presentar las ventajas y los inconvenientes de las opciones de tratamiento que se están considerando, incluido el empleo creciente de recursos[2]. En las tablas 1 y 2 se presenta un ejemplo de un conjunto de evidencias derivadas de un amplio ensayo clínico internacional (realizado en 33 países) y un análisis económico específico para investigar la utilidad del sulfato de magnesio en mujeres con preeclampsia[3,4].

LOS CONJUNTOS DE EVIDENCIAS DEBEN PRESENTAR EL EMPLEO DE RECURSOS, Y NO SÓLO SU COSTE ECONÓMICO

Recomendamos a quienes preparan directrices que documenten los mejores cálculos del empleo de recursos, no los mejores cálculos de su coste. Los costes dependen de los recursos consumidos y del coste por unidad de recurso. Dada la amplia variabilidad en los costes de estas unidades, notificar únicamente los costes totales priva a los usuarios de la información necesaria para juzgar si los cálculos de los costes por unidad son aplicables a su situación.

Asimismo, si se especifican los recursos consumidos por estrategias de tratamiento alternativo se permite a los usuarios juzgar si el empleo de recursos refleja las pautas de procedimientos en su ámbito y centrarse en los aspectos de más relevancia para ellos (p. ej., el gasto en medicamentos para una farmacia o un hos-

pital para el administrador correspondiente). Por último, los usuarios pueden verificar si los costes por unidad son aplicables en su ámbito y si los recursos económicos son asignados posteriormente a los recursos utilizados; de lo contrario, podrán sustituirlos por costes por unidades que sí son asignados.

En las tablas 1 y 2 se muestra la importancia de documentar el empleo de recursos y especificar el contexto en que se brindan. Puede observarse una considerable variación en los costes inherentes al sulfato de magnesio, su administración y los costes hospitalarios asociados en distintos países con ingresos nacionales brutos elevados, medios y bajos. Nuestras tablas documentan estas diferencias, pero muchos análisis económicos no lo harán. A menos que se especifique el empleo de recursos, los usuarios en situaciones diferentes a aquellas en las que se enfocan los analistas no pueden calcular los costes crecientes inherentes a la intervención.

EL CONTEXTO ESPECÍFICO ES CRUCIAL PARA CONSIDERAR EL USO DE RECURSOS

La enorme variabilidad en los costes en función del tiempo y las áreas geográficas tiene varias implicaciones. En primer lugar, los grupos de expertos que preparan directrices deben especificar muy claramente la población de pacientes, las características de la intervención, el elemento de comparación y el contexto sanitario. La selección del elemento de comparación puede ser un problema importante en los análisis económicos. Si es inapropiado (p. ej., ningún tratamiento en vez de un tratamiento menos eficaz), las conclusiones pueden ser engañosas[5].

En segundo lugar, un grupo de trabajo que prepare una guía puede, legítimamente, no hacer consideraciones sobre el empleo de recursos y proponer recomendaciones basándose únicamente en otras ventajas e inconvenientes de las alternativas que se estén considerando. En tercer lugar, si los expertos contemplan el empleo de recursos, deben decidir, antes de tener en cuenta los costes en la ecuación, cuál es la calidad de la evidencia con relación a otras variables y sopesar sus ventajas y sus inconvenientes.

ES CONVENIENTE AMPLIAR LA PERSPECTIVA

Es posible que una recomendación pudiera dirigirse a un grupo de usuarios muy concreto, como la farmacia de un determinado hospital, un hospital o una organización para el mantenimiento de la salud (HMO, por sus siglas en inglés). Como alternativa, podría estar dirigida a una región sanitaria, un país o a un público internacional.

Sin embargo, pocos interesados en una recomendación estarían satisfechos con una perspectiva más reducida que la del sistema sanitario en su conjunto. Por ejemplo, en un sistema de salud financiado con recursos públicos, la perspectiva del paciente no tendría en cuenta la mayoría de los costes generados, la de una farmacia haría lo mismo los ahorros logrados en costes vinculados como resultado de la prevención de sucesos adversos (como el accidente cerebrovascular o el infarto del miocardio) gracias a un fármaco, mientras que la de un hospital no consideraría los costes de los pacientes ambulatorios, ni los reales ni los evitados[6].

La perspectiva más completa es la de la sociedad, puesto que incluye todos los costes, independientemente de quién los cubra. Esta perspectiva suele ser preferible, sobre todo si la intervención sanitaria tiene un efecto amplio (p. ej., una intervención para la insuficiencia cardíaca que mejora la actividad de los pacientes y reduce el tiempo y el coste relacionado con los cuidadores familiares). Lo que es más cuestionable es si los análisis sobre coste-efectividad deben incluir las implicaciones de los efectos sobre la salud, como los cambios en los ingresos. Las directrices económicas recomiendan que estas implicaciones se presenten por separado, en vez de cómo parte de un análisis formal de coste-efectividad.

Aunque un plan de salud específico puede no conllevar costes vinculados, es informativo y permite a las autoridades percatarse del empleo creciente de recursos a largo plazo inherentes a estrategias de tratamiento alternativas. Asimismo, aunque la responsabilidad de un médico que atiende a un paciente es para con éste y su familia, se asume en un contexto más amplio en el que existen limitaciones de recursos y costes de oportunidad: los recursos que se utilizan para una intervención no se pueden emplear para otras y pueden afectar a la capacidad del sistema de salud para cumplir mejor con las necesidades de los ciudadanos.

EVALUACIÓN DE LA CALIDAD DE LA EVIDENCIA PARA EL EMPLEO DE RECURSOS

Al igual que con la evidencia de efectos adversos raros pero importantes, la evidencia del uso de recursos puede provenir de fuentes distintas de las utilizadas para valorar beneficios para la salud. Esto puede deberse a que los ensayos sobre las intervenciones no informan por completo sobre el empleo de recursos, ya que la situación del ensayo puede no reflejar bien

Tabla 1 | Resumen de resultados con respecto a si los médicos deben utilizar el sulfato de magnesio para prevenir la eclampsia: variables clínicas

Variable	Gravedad de la preeclampsia	Riesgo del grupo de referencia típico	Efecto absoluto típico (IC del 95 %)	Riesgo relativo (IC del 95 %)	No. de participantes	Calidad de la evidencia
Eclampsia	Grave*	27/1.000	16 menos/1.000 (de 11 a 19)	0,41 (de 0,29 a 0,58)	11.444	Alta†
	No grave	15/1.000	9 menos/1.000 (de 6 a 11)			
Muerte materna	Grave	6/1.000	3 menos/1.000 (de 0,6 más a 4 menos)	0,54 (de 0,26 a 1,10)	10.795	Moderada‡
	No grave	3/1.000	1 menos/1.000 (de 0,3 más a 2 menos)			
Efecto secundario§	Grave y no grave	46/1.000	196 más/1.000 (de 165 a 231)	5,26 (de 4,59 a 6,03)	9.992	Alta†

*La eclampsia grave fue definida como (tensión arterial diastólica > 110 mmHg en dos ocasiones, o tensión arterial sistólica > 170 mmHg en dos ocasiones y proteinuria > 3+) o (tensión arterial diastólica > 100 mmHg en dos ocasiones, o tensión arterial sistólica > 150 mmHg en dos ocasiones y proteinuria > 2+ y, como mínimo, dos signos o síntomas de eclampsia inminente) o, para las mujeres que recibieron un antihipertensivo en las 48 h previas a la aleatorización: (en las 48 h antes de ingresar en el ensayo, tensión arterial diastólica máxima > 110 mmHg o tensión arterial sistólica máxima > 170 mmHg y proteinuria > 3+ en el momento de su inclusión en el ensayo) o (en las 48 h previas al ingreso en el estudio, tensión arterial diastólica más alta > 100 mmHg o tensión arterial sistólica más alta > 150 mmHg y proteinuria > 2+ y, como mínimo, dos signos o síntomas de eclampsia inminente).

†La evidencia se deriva de ensayos aleatorizados y no hubo ninguna razón para reducir su grado debido a limitaciones del estudio, imprecisiones, inconsistencias, datos indirectos o sesgo de publicación.

‡El intervalo de confianza fue amplio, de manera que se estableció un grado menor para la evidencia a causa de la imprecisión.

§Principalmente, rubefacción. Otros efectos secundarios son náuseas, vómitos, voz farfullante, debilidad muscular, mareos, somnolencia, confusión y cefalea.

Tabla 2 | Resumen de datos con respecto a si los médicos deben utilizar sulfato de magnesio para prevenir la preeclampsia: uso de recursos considerado desde la perspectiva del sistema sanitario

Recursos	Coste*	Efecto absoluto típico (IC del 95 %)	No. de participantes (estudios)	Calidad de la evidencia	Comentarios
Ampollas de sulfato de magnesio (ampollas de 6 × 10 ml/paciente)					
Contexto:					
Países de ingresos altos	20 dólares más/paciente		9.996	Alta†	
Países de ingresos medios	3 dólares más/paciente				
Países de ingresos bajos	5 dólares más/paciente				
Administración de sulfato de magnesio (1 ampolla/paciente)					
Contexto:					
Países de ingresos altos	66 dólares/paciente		9.996	Alta†	Los recursos para administrar sulfato de magnesio incluyeron el tiempo de trabajo de comadronas (coste principal), agujas y cánulas intravenosas, jeringas, líquidos intravenosos y el fármaco
Países de ingresos medios	14 dólares/paciente				
Países de ingresos bajos	8 dólares/paciente				
Otros recursos hospitalarios (variaron ampliamente)					
Ámbito:					
Países de altos ingresos	12.839 dólares	20 dólares menos/paciente (de 0 a 60)	9.996	Moderada‡	El empleo de otros recursos intrahospitalario fue muy variable tanto en los grupos de intervención como en los de referencia. El resto de costes hospitalarios han sido ajustados basándose en la influencia de la eclampsia, para tener en cuenta muchos otros factores que influyeron en estos costes
Países de medios ingresos	1.416 dólares	4 dólares menos/paciente (de 0 a 10)			
Países de bajos ingresos	157 dólares	2 dólares menos/paciente (de 1 a 3)			

*1 dólares = 0,7 euros.
†La evidencia se deriva de ensayos aleatorizados, y no hubo razones para asignar un grado menor a causa de limitaciones del estudio, imprecisiones, incoherencias, datos indirectos o sesgos de publicación.
‡El intervalo de confianza fue amplio, de manera que se asignó un menor grado de evidencia a causa de esta imprecisión.

las circunstancias —y, por tanto, el empleo de recursos— que esperaríamos en el ejercicio clínico, o porque el uso de recursos pertinentes puede extenderse más allá de la duración del ensayo.

Para el empleo de recursos comunicado en el contexto del ensayo, los criterios de valoración de la calidad son idénticos a los de otras variables, según se describe en el segundo artículo de esta serie; es el caso que se presenta en la tabla 1. Como ocurre con el resto de resultados de un ensayo, la calidad de la evidencia puede ser diferente en contextos con distintos recursos. Por ejemplo, al considerar el sulfato de magnesio en la preeclampsia, hay más certidumbre sobre el empleo de recursos relacionados con el fármaco y su administración que con respecto al empleo de los recursos del hospital (tabla 2).

EL MODELADO ECONÓMICO FORMAL PUEDE SER ÚTIL

El modelado económico formal da como resultado un coste por unidad de beneficio logrado: el coste por unidad natural, como el coste por accidente cerebrovascular prevenido (análisis de coste-efectividad), el coste por año de vida ganado ajustado con respecto a la calidad (análisis de coste-utilidad), o el coste como beneficios económicos (análisis de coste-beneficio). Estos resúmenes son de utilidad porque brindan información para establecer criterios. Lamentablemente, los análisis de coste-efectividad, sobre todo los de fármacos, son, en muchos casos, imperfectos, sesgados[7] y específicos de un contexto concreto.

Por tanto, los grupos de expertos que preparan directrices pueden considerar establecer su propio modelo económico formal. Sin embargo, para considerar esta opción, deben tener la experiencia y los recursos necesarios. Cuanto mayor sea la diferencia en los recursos consumidos por las estrategias de tratamiento alternativas, mayor será la incertidumbre con respecto a si los beneficios de una intervención justifican o no los costes crecientes, y cuanto mayor sea la calidad de la evidencia con respecto al consumo de recursos, más probabilidades habrá de que un modelo económico completo proporcione información para tomar una decisión.

El modelado, pese a ser necesario para tener en cuenta los aspectos complejos y las incertidumbres en el cálculo del coste por unidad de beneficio, reduce la transparencia. Además, cualquier modelo es tan satisfactorio como los datos en los cuales se basa. Cuando los cálculos de beneficios, daños o recursos utilizados se deriven de pruebas de baja calidad, los resultados de cualquier ejercicio de modelado serán muy teóricos.

Se dispone de criterios para valorar la credibilidad que debe otorgarse a resultados de modelos estadísticos de coste-efectividad o coste-utilidad[8-11]. Sin embargo, estos modelos suelen incluir un gran número de suposiciones y de evidencias de calidad variable para los distintos cálculos que comprende el modelo. Por estas razones, no recomendamos la inclusión de modelos de coste-efectividad o coste-utilidad en los conjuntos de evidencia. Sin embargo, pueden proporcionar información para que el grupo de trabajo encargado de elaborar directrices adopte criterios, o para que los gobiernos o las organizaciones sanita-

Tabla 3 | Coste creciente por cada episodio de eclampsia prevenido con sulfato de magnesio

	Gravedad de la eclampsia	
Ingreso nacional	Grave	No grave
Elevado	4.125 dólares	7.333 dólares
Medio	813 dólares	1.444 dólares
Bajo	688 dólares	1.222 dólares

rias consideren si es conveniente incluir una intervención entre los beneficios que ofrecen sus programas.

Las tablas 1 y 2 permiten calcular el coste creciente por episodio de eclampsia prevenido para la preeclampsia grave y no grave, en países de ingresos elevados, medios y bajos (tabla 3). Aun cuando los cálculos de coste-efectividad sean fiables —como ocurre en este caso—, no proporcionan respuestas claras respecto a las acciones apropiadas. Sin embargo, la mayoría de las personas considerarían que el coste por episodios de eclampsia prevenidos justificaría la inversión económica en el caso de la preeclampsia grave. Para la preeclampsia no grave, sobre todo en los países de bajos ingresos, la decisión es más difícil. En última instancia, las autoridades sanitarias deben sopesar la utilidad relativa de la prevención de la preeclampsia considerando los beneficios que recibirá el sistema sanitario o la sociedad al asignar recursos a la administración de sulfato de magnesio.

OBSERVACIONES FINALES

La toma de decisiones clínicas es un proceso complejo. Las directrices pueden ayudar a médicos y pacientes a decidir entre opciones complejas, mejorar la calidad de la atención y ayudar a garantizar el mejor empleo de recursos limitados. Para cerciorarse de que las directrices proporcionan información correcta, es importante que se basen en la mejor evidencia disponible y que los expertos que las preparan utilicen procesos sistemáticos y claros para evaluar sobre la calidad de la evidencia, fundamentando en ella sus recomendaciones y considerando cómo se utilizan los recursos.

Para los médicos y sus pacientes, serán más útiles las directrices que utilicen un enfoque como el que hemos descrito en esta serie para evaluar explícitamente la calidad de la evidencia y la fuerza de las recomendaciones. No es necesario que los médicos que atienden directamente a pacientes o los que elaboran directrices locales reproduzcan el trabajo de los expertos que preparan directrices y que disponen de recursos adecuados.

Sin embargo, para hacer un mejor uso de las directrices, deben comprender la evidencia y los criterios en que se basan. Deben tener acceso a resúmenes concisos de recomendaciones, que incluyan evaluaciones de la calidad de la evidencia subyacente y de la fuerza de la recomendación, y comprender el significado de los distintos niveles de evidencia y sus implicaciones en la toma de decisiones clínicas.

CONCEPTOS BÁSICOS

Los costes presentan diferencias respecto a otras variables sanitarias, ya que son compartidos por pacientes, empresarios y la sociedad en general. Por otra parte, los criterios con respecto de quién debe asumirlos son variables. Algunas personas consideran que los costes no deben influir en las decisiones de los médicos. Además, pueden variar en cada área geográfica y en el seno de cada una de ellas.

Las hojas de balance proporcionan información para establecer criterios con respecto a si los beneficios netos justifican los costes crecientes.

Las series de evidencias deben presentar siempre el uso de recursos, y no sólo su coste económico.

Los grupos de expertos que elaboran directrices pueden optar, legítimamente, por no tener en cuenta los costes.

El modelado económico formal puede o no ser de utilidad.

AGRADECIMIENTOS

Los miembros del grupo de trabajo GRADE son Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Margaret Haugh, Robin Harbour, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Nicola Magrini, Merce Marzo, James Mason, Jacek Mrukowics, Susan Norris, Andrew D Oxman, Vivian Robinson, Holger J Schünemann, Jane Thomas, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams y James Woodcock.

Colaboradores: Todos los autores enumerados y miembros del grupo de trabajo de GRADE ayudaron a desarrollar las ideas del manuscrito, que fue revisado y aprobado por todos ellos. CHG redactó el primer borrador e incorporó los comentarios de autores y revisores en los borradores sucesivos. El resto de autores mencionados contribuyeron con ideas sobre la estructura y el contenido del artículo, proporcionaron ejemplos, analizaron borradores sucesivos del manuscrito y dieron su opinión al respecto. CHG es el garante.

Financiación: Ninguna.

Conflicto de intereses: Todos los autores participan en la difusión del sistema GRADE, y el éxito de este sistema tiene una influencia positiva en sus carreras académicas. Los autores enumerados han recibido reembolsos de compensaciones y honorarios por presentaciones, entre ellas un análisis del enfoque GRADE para la evaluación de la calidad de la evidencia y la calificación de las recomendaciones. CHG es asesor de UpToDate; su trabajo incluye ayudar a UpToDate a aplicar el sistema GRADE. HJS es editor de los documentos y experto en metodología en la American Thoracic Society; una de sus funciones en estos puestos es ayudar a implementar el sistema GRADE. Además, recibe subvenciones de «The human factor, mobility and Marie Curie actions scientist reintegration European commission grant: IGR 42192—GRADE». AL está ayudando a poner en marcha el sistema GRADE en distintas instituciones del sistema público de salud de Italia y ha implementado dicho sistema para elaborar recomendaciones clínicas en oncología a través de la donación No. 249 (2005-7), Bando Recerca Finalizzata, Ministerio de Sanidad, Roma (Italia).

Procedencia y análisis de expertos: No solicitada; analizado por expertos externos.

1 Guyatt G, Baumann M, Pauker S, Halperin J, Maurer J, Owens DK, et al. Addressing resource allocation issues in recommendations from clinical practice guideline panels: suggestions from an American College of Chest Physicians task force. *Chest* 2006;129:182-7.
2 Eddy DM. Comparing benefits and harms: the balance sheet. *JAMA* 1990;263:2493, 2498, 2501.
3 Magpie Trial Collaborative Group. Do women with pre-eclampsia, and their babies, benefit from magnesium sulphate? The magpie trial: a randomised placebo-controlled trial. *Lancet* 2002;359:1877-90.
4 Simon J, Gray A, Duley L. Cost-effectiveness of prophylactic magnesium sulphate for 9996 women with pre-eclampsia from 33 countries: economic evaluation of the magpie trial. *Br J Obstet Gynaecol* 2006;113:144-51.
5 Montori VM, Jaeschke R, Schünemann HJ, Bhandari M, Brozek JL, Devereaux PJ, et al. Users' guide to detecting misleading claims in clinical research reports. *BMJ* 2004;329:1093-6.
6 Luce B, Manning W, Siegel J. Estimating costs in cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine.* Oxford: Oxford University Press, 1996: 176-213.
7 Friedberg M, Saffran B, Stinson TJ, Nelson W, Bennett CL. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *JAMA* 1999;282:1453-7.
8 Garber AM, Phelps CE. Economic foundations of cost-effectiveness analysis. *J Health Econ* 1997;16:1-31.
9 Owens DK. Interpretation of cost-effectiveness analyses. *J Gen Intern Med* 1998;13:716-7.
10 Gold M, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine.* Oxford: Oxford University Press, 1996.
11 O'Brien B, Drummond M, Richardson WS, Levine M, Heyland D, Guyatt G. Economic analysis. In: Guyatt G, Rennie D, eds. *Users' guides to the medical literature.* Chicago: AMA Press, 2002, 621-44.