Localizador web
Artículo 117.901

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

GRADE Working Group

M. Marzo-Castillejo[a,b,c] y P. Alonso-Coello[a,d]

Las guías de práctica clínica son sólo tan buenas como las evidencias y los juicios en las que están basadas. El objetivo de la clasificación GRADE es facilitar la valoración de los juicios que hay detrás de las recomendaciones.

Resumen

Los usuarios de guías de práctica clínica y otras recomendaciones necesitan saber hasta qué punto pueden confiar en ellas. Los juicios llevados a cabo mediante una metodología sistemática y explícita permiten disminuir los errores y mejorar la comunicación. Hemos desarrollado un sistema para clasificar la calidad de las evidencias y la fuerza de las recomendaciones que pueda aplicarse a una amplia gama de intervenciones y contextos. En este artículo se presenta un resumen de nuestro enfoque desde la perspectiva del usuario de una guía. Los juicios sobre la fuerza de una recomendación deben tener en cuenta el balance entre beneficios y riesgos, la calidad de la evidencia, la aplicación de esta evidencia en circunstancias específicas y la situación de riesgo basal. Antes de elaborar una recomendación también es importante considerar los costes (utilización de recursos). Las inconsistencias entre los sistemas que clasifican la calidad de la evidencia y la fuerza de las recomendaciones dificultan la evaluación crítica y la comunicación de estos juicios. Nuestro sistema, para llevar a cabo estos complejos juicios, equilibra la sencillez con la valoración global y transparente de todos los aspectos importantes.

[a]Grupo de trabajo GRADE y miembros de REDEGUÍAS.

[b]Divisió d'Atenció Primària. Institut Català de la Salut. Barcelona. España.

[c]Comité científico y grupo MBE de la semFYC.

[d]Colaborador Centro Cochrane Iberoamericano.

Traducción al castellano del artículo:
Education and debate
Grading quality of evidence and strength of recommendations.
GRADE Working Group

Referencia Pubmed:
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 GRADE Working Group. BMJ. 2004;328:1490.

Revisada por:
Rosa Rico, Rafael Rotaeche, Arritxu Etxeberria, Itziar Pérez y Eulali Mariñelarena, miembros de REDEGUÍAS, y por Pilar Aizpurúa, Manel Ferran, Carmen Cabezas, Ana González, Rafael Bravo y Josep Jiménez

Correspondencia:
M. Marzo Castillejo
Correo electrónico: mmarzo@ics.scs.es

Introducción

Los juicios acerca de las evidencias y las recomendaciones son complejos. Consideremos, por ejemplo, la elección entre los inhibidores selectivos de la recaptación de la serotonina y los antidepresivos tricíclicos para el tratamiento de la depresión moderada. Los médicos deben decidir qué resultados se han de considerar, qué evidencia se debe tener en cuenta para cada uno de los resultados observados, cómo evaluar la calidad de la evidencia y cómo determinar si los inhibidores selectivos de la recaptación de serotonina proporcionan, comparados con los tricíclicos, más beneficio que riesgo. Dado que los recursos siempre son limitados y que el dinero utilizado en los inhibidores selectivos de la recaptación de serotonina ya no está disponible para otros fines, los médicos también pueden tener que decidir si un incremento de los beneficios en la salud justifica los costes adicionales.

No es práctico que médicos y pacientes, de forma individual y sin ayuda, realicen estos juicios para cada decisión clínica. A menudo, médicos y pacientes utilizan las guías de práctica clínica como herramienta de apoyo. Estas guías incluyen recomendaciones elaboradas de manera sistemática por grupos de trabajo que tienen acceso a la evidencia disponible, conocen el problema clínico y la metodología de investigación, y disponen de tiempo para la reflexión.

Los usuarios de las guías elaboradas de manera sistemática necesitan conocer hasta qué punto pueden confiar en las evidencias y recomendaciones. El grupo de trabajo ha descrito los principios en los que basar nuestra confianza, así como un enfoque sistemático para llevar a cabo los complejos juicios que, implícita o explícitamente, requieren las guías de práctica clínica u otras recomendaciones para la salud. En este artículo, y con el objetivo de simplificar, no se comentan todos los matices ni se presentan de manera pormenorizada todos los aspectos que los grupos que elaboran guías necesitan para aplicar este enfoque. Esta información se puede obtener de los autores (http://www.gradeworkinggroup.org/).

Los miembros del GRADE Working Group están relacionados al final del artículo.

TABLA 1 **Comparación de la clasificación de GRADE con otros sistemas**

Elemento	Otros sistemas	GRADE	Ventajas de la clasificación de GRADE*
Definiciones	Definiciones implícitas de calidad (nivel) de evidencia y fuerza de recomendación	Definiciones explícitas	Clarifica el significado que tienen los grados y lo que se debería tener en cuenta al realizar estos juicios
Juicios	Juicios implícitos sobre qué resultados son importantes, calidad de la evidencia para cada resultado importante, calidad de la evidencia global, balance entre beneficios y riesgos y valor del incremento de beneficios	Juicios secuenciales, explícitos	Clarifica cada uno de estos juicios y reduce el riesgo de introducir errores o sesgos que pueden surgir cuando los juicios se realizan implícitamente
Componentes esenciales de la calidad de la evidencia	No son considerados para cada resultado importante. Los juicios sobre la calidad de la evidencia están a menudo basados únicamente en el diseño de estudio	Consideración sistemática y explícita del diseño del estudio, calidad del estudio, consistencia y si la evidencia es directa o indirecta, en los juicios sobre la calidad de la evidencia	Asegura que estos factores se consideran de manera apropiada
Otros factores que pueden afectar la calidad de la evidencia	No se tienen en cuenta de manera explícita	Consideración explícita de los datos imprecisos o escasos, sesgo de información, fuerza de la recomendación, evidencia sobre el gradiente dosis-respuesta y posibles factores de confusión	Asegura la consideración de otros factores
Calidad global de la evidencia	Implícitamente basados en la calidad de la evidencia sobre los beneficios	Basado en la calidad más baja de la evidencia para cualquiera de los resultados clave para tomar una decisión	Disminuye la probabilidad de clasificar mal la calidad global de la evidencia cuando la evidencia para un resultado clave no está disponible
Importancia relativa de los resultados	Considerada implícitamente	Juicios explícitos sobre qué resultados son claves, importantes pero no claves, y poco importantes y que incluso pueden ser ignorados	Al clasificar la calidad global de la evidencia y la fuerza de las recomendaciones, asegura la consideración apropiada de cada resultado
Balance entre beneficios de salud y riesgos	No considerados explícitamente	Consideración explícita de balance entre beneficios y riesgos, calidad de la evidencia de éstos, traslación de la evidencia a las circunstancias específicas e incertidumbre del riesgo basal	Clarifica y mejora la transparencia de los juicios sobre beneficios y riesgos
Si el incremento de beneficios en salud merece los costes	No considerados explícitamente	Consideración explícita tras la consideración inicial de si hay beneficios netos para la salud	Asegura que los juicios sobre el valor de los beneficios netos para la salud sean transparentes
Resúmenes de la evidencia y los hallazgos	Presentación inconsistente	Tablas de síntesis de evidencia GRADE consistentes, incluyendo evaluación de la calidad y resumen de los hallazgos	Asegura que todos los miembros del grupo basan sus juicios en la misma información y que ésta es accesible a otros
Alcance de su utilización	Raramente usado por más de una organización y escasa, si alguna, evaluación empírica	Colaboración internacional de diversas organizaciones que las desarrollan y evalúan	Construido sobre la experiencia previa para conseguir un sistema que sea más razonable, fiable y ampliamente aplicable

*La mayoría de los otros sistemas no incluyen ninguna de estas ventajas, aunque algunos incorporan algunas.

Al llevar a cabo juicios sobre la calidad de la evidencia y la fuerza de las recomendaciones, la aplicación de un enfoque sistemático y explícito puede ayudar a prevenir errores, facilitar la valoración crítica de estas decisiones y mejorar la comunicación de esta información. Desde la década de los años setenta, cada vez un mayor número de organizaciones utilizan algún sistema para clasificar la calidad (el nivel) de la evidencia y la fuerza de las recomendaciones[1-28]. Lamentablemente, las diferentes organizaciones utilizan sistemas distintos y, dependiendo del sistema empleado, una misma evidencia y recomendación podría clasificarse como II-2 y B; C+ y 1, o evidencia sólida y fuertemente recomendada. Esto resulta confuso e impide una comunicación efectiva.

El grupo de trabajo de GRADE (los miembros del grupo se citan al final del artículo) comenzó como una colaboración informal entre personas interesadas en abordar las deficien-

Recuadro 1.

Etapa previa

1. Establecer las bases: por ejemplo, priorización de los problemas, selección de un grupo de trabajo, declaración de los conflictos de interés y funcionamiento del grupo

Etapa preparatoria

2. Revisión sistemática: el primer paso es identificar y valorar críticamente, o preparar las revisiones sistemáticas a partir de la mejor evidencia disponible para todos los resultados importantes
3. Preparar una tabla de síntesis de la evidencia para todos los resultados importantes. Las tablas son necesarias para cada subpoblación o grupo de riesgo, deben estar basadas en los resultados de las revisiones sistemáticas e incluir una evaluación de la calidad y un resumen de las conclusiones

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

4. Calidad de la evidencia para cada resultado: la calidad se debe evaluar a partir de la información resumida en la tabla de síntesis de la evidencia y según los criterios de la tabla 2
5. Importancia relativa de los resultados: en la tabla de síntesis de la evidencia sólo se deberían incluir los resultados importantes. Éstos deberían clasificarse como resultados clave o resultados importantes (pero no clave) para cada una de las decisiones
6. Calidad global de la evidencia: la calidad global de la evidencia debería valorarse a partir de los resultados y basarse en la calidad más baja para cualquiera de los resultados clave
7. Balance entre beneficios y riesgos: el balance entre los beneficios en salud y los riesgos debería clasificarse como beneficios netos, beneficios con aceptación de los riesgos, beneficios inciertos con aceptación de los riesgos y ausencia de beneficios netos
8. Balance entre beneficio neto y costes: ¿para un incremento de los beneficios en salud merecen la pena los costes? Ya que los recursos son siempre limitados, cuando se elabora una recomendación es importante considerar los costes (utilización de recursos)
9. Fuerza de la recomendación: las recomendaciones deberían ser formuladas para que reflejen su fuerza, esto es, hasta qué punto se puede confiar que el poner en práctica esa recomendación será más beneficioso que perjudicial

Etapas posteriores

10. Implementación y evaluación: por ejemplo, utilizando estrategias de implantación efectivas que contemplen barreras para el cambio, evaluación de la implementación e incorporación de las actualizaciones

cias de los actuales sistemas de clasificación. En la tabla 1 se resumen estas deficiencias y como las hemos solucionado. El sistema GRADE permite llevar a cabo juicios más consistentes y la comunicación de estos juicios respalda las opciones mejor fundamentadas en la atención sanitaria. En el recuadro 1 se describen las etapas de la elaboración e implementación de las guías, desde la priorización de los problemas hasta la evaluación de su implementación. Este artículo se centra en la clasificación de la calidad de la evidencia y la fuerza de las recomendaciones.

Definiciones

Hemos utilizado las siguientes definiciones: la calidad de la evidencia indica hasta qué punto podemos confiar en que el estimador del efecto es correcto. La fuerza de una recomendación indica hasta qué punto podemos confiar en que poner en práctica la recomendación conllevará más beneficio que riesgo.

Para llevar a cabo juicios acerca de la calidad de la evidencia se necesita valorar la validez de los resultados de los estudios individuales para los resultados importantes. Para llevar a cabo estos juicios se deberían utilizar criterios explícitos[26,29-32]. Las etapas que se deben seguir en este enfoque permiten realizar juicios secuenciales acerca de:

– La calidad de la evidencia en los diferentes estudios para cada uno de los resultados importantes.
– Los resultados que son claves para una decisión.
– La calidad global de la evidencia para estos resultados clave.
– El balance entre beneficios y riesgos.
– La fuerza de las recomendaciones.

Todos estos juicios dependen de que dispongamos de una pregunta definida con claridad y de que se consideren todos los resultados que probablemente sean importantes para las personas afectadas. La pregunta debería identificar qué opciones se están comparando (p. ej., los inhibidores selectivos de la recaptación de serotonina y los antidepresivos tricíclicos), a quién van dirigidas (pacientes adultos con depresión moderada) y en qué ámbito (atención primaria en Inglaterra).

Calidad de la evidencia para cada resultado importante

La revisión sistemática de la evidencia disponible debería guiar estos juicios. Los revisores deberían considerar 4 elementos esenciales: el diseño del estudio, su calidad, la consistencia y si la evidencia es directa o indirecta.

Diseño del estudio

El diseño del estudio se refiere al tipo de estudio, que en términos generales hemos clasificado en observacional o ensayo aleatorizado. Esta clasificación se fundamenta en argumentos lógicos y en evidencias empíricas[33-36]. A pesar de que los estudios observacionales a menudo presentan resultados similares a los de los ensayos aleatorizados, esto no siempre es así. Un ejemplo significativo es la diferencia entre los resultados de los estudios observacionales, que sugerían que la terapia hormonal sustitutiva disminuía el riesgo de enfermedad coronaria, y los posteriores ensayos aleatorizados que no han confirmado dicha reducción, e incluso muestran un aumento del riesgo[37,38]. Desafortunadamente, no es posible conocer con antelación si los estudios observacionales van a predecir los hallazgos de los posteriores ensayos aleatorizados. Una vez se dispone de los resultados de ensayos aleatorizados de alta calidad, resulta difícil mantener recomendaciones basadas en ensayos no aleatorizados con resultados discrepantes.

Por otro lado, los ensayos aleatorizados no siempre son factibles y en ocasiones los estudios observacionales pueden proporcionar mejores evidencias, como generalmente ocurre en el caso de los efectos adversos poco comunes. Ade-

más, los resultados de los ensayos aleatorizados no siempre se pueden aplicar; por ejemplo, esto ocurre cuando los participantes están muy seleccionados y más motivados que la población de interés. Por tanto, es esencial considerar la calidad del estudio, la consistencia de los resultados entre los estudios, si la evidencia es directa o indirecta y la adecuación del diseño del estudio. Por ejemplo, un estudio de serie de casos bien diseñado puede proporcionar evidencia de alta calidad acerca de las tasas de complicaciones quirúrgicas (muertes ocurridas durante las intervenciones) u otros procedimientos (perforaciones a partir de la colonoscopia), y esta evidencia es más relevante que la que proporcionan los ensayos aleatorizados. De manera similar, los estudios de cohortes pueden proporcionar evidencia de alta calidad sobre la tasa de recitación o la tasa de intervenciones realizadas como consecuencia de un resultado de cribado falso positivo (p. ej., la tasa de biopsias a partir de la mamografía).

Calidad del estudio

La calidad del estudio hace referencia a los métodos utilizados y a su realización. Los revisores deberían aplicar criterios apropiados para la evaluación de la calidad del estudio para cada resultado importante[26,29-32]. Por ejemplo, en el caso de los ensayos aleatorizados, podrían usar criterios sobre la adecuación de la ocultación de la asignación, el enmascaramiento y el seguimiento. Los revisores deberían hacer explícitos los motivos para bajar el nivel de calidad. Por ejemplo, pueden plantear que la falta de enmascaramiento de pacientes y médicos reduce la calidad de la evidencia del impacto de una intervención sobre la intensidad de dolor, y lo consideran una limitación importante.

Consistencia

La consistencia se refiere a la similitud en las estimaciones del efecto entre los estudios. Cuando en los resultados hay una inconsistencia inexplicada importante, disminuye nuestra confianza en el estimador del efecto para este resultado. Las diferencias en la dirección del efecto, su magnitud e importancia ayudan a determinar (inevitablemente de forma un tanto arbitraria) si hay una inconsistencia importante. La magnitud del efecto se debería estimar por separado cuando, de forma convincente, la inconsistencia se explica por pertenecer a un subgrupo diferente. Por ejemplo, los diferentes resultados de la endarterectomía de carótida en la estenosis de alto y bajo grado nos orientan hacia la separación del estimador del efecto para estos 2 subgrupos.

Evidencia directa o indirecta (directness)

La evidencia directa o indirecta indica hasta qué punto los sujetos, las intervenciones y las medidas de resultado de los estudios son similares a aquellos de nuestro interés. Por ejemplo, puede haber incertidumbre sobre si la evidencia es directa cuando la población de interés es mayor, está más enferma o presenta más comorbilidad que la población de los estudios[39]. Para determinar si la incertidumbre es importante, nos podemos preguntar sobre posibles motivos que justifiquen que encontremos diferencias importantes en la magnitud del efecto. No es necesario aplicar criterios excesivamente estrictos para decidir si la evidencia es directa, ya que muchas intervenciones tienen efectos similares en la mayoría de los grupos de pacientes. Para algunos tratamientos, por ejemplo, en las terapias conductuales, en las que las diferencias culturales pueden ser importantes, estaría indicado aplicar criterios más estrictos.

Asimismo, los revisores pueden valorar si hay incertidumbre sobre si la evidencia es directa para fármacos que son distintos a los de los estudios, pero que pertenecen a su misma clase. Dudas similares surgen para otros tipos de intervenciones. Por ejemplo, ¿se pueden generalizar los resultados de llevar a cabo un consejo sanitario de menor intensidad que el utilizado en un estudio?, o ¿se puede utilizar una técnica quirúrgica alternativa a la aplicada en el estudio? Llevar a cabo estos juicios puede ser difícil[40], y es importante que los investigadores expliquen el razonamiento que han seguido para elaborar sus conclusiones.

Por otro lado, los estudios que utilizan resultados intermedios, generalmente proporcionan menos evidencia directa que los que emplean resultados finales que son importantes para las personas. Por tanto, al valorar si la evidencia a partir de los estudios con resultados intermedios es directa, es prudente usar criterios mucho más estrictos. Ejemplos de evidencia indirecta a partir de estudios con resultados intermedios y cuya evidencia, sobre la base de los resultados de posteriores ensayos, se ha mostrado que era engañosa incluye: la supresión de la arritmia cardíaca como resultado intermedio para estimar la mortalidad en pacientes que han presentado un infarto de miocardio[41], valor de las lipoproteínas en la enfermedad coronaria[37] y densidad ósea de las mujeres posmenopáusicas en la disminución de fractura[42].

La validez de una prueba diagnóstica también es un resultado intermedio de otros resultados importantes que puede verse afectada por un diagnóstico preciso, ya sea por una mejoría en términos de salud debida a un tratamiento adecuado y a una reducción del daño producido por un resultado falso positivo. En la valoración del diseño de los estudios de validez diagnóstica se han de utilizar otros criterios. No obstante, para valorar si la evidencia es directa debemos basarnos en el grado de confianza que tenemos en la relación entre ser clasificado correctamente (verdadero positivo o verdadero negativo) o incorrectamente (falso positivo o falso negativo) y las importantes consecuencias que conlleva. Por ejemplo, disponemos de evidencia consistente derivada de estudios bien diseñados de que, ante la sospecha diagnóstica de un cólico nefrítico agudo, la tomografía helicoidal sin contraste presenta menos resultados falsos negativos que la pielografía intravenosa[43]. Sin

Recuadro 2. Criterios para asignar el grado de evidencia

Tipo de evidencia

Ensayo aleatorizados: alta
Estudio observacional: baja
Cualquier otra evidencia: muy baja

Disminuir el grado si

– Limitación importante (–1) o muy importante (–2) en la calidad del estudio
– Inconsistencia importante (–1)
– Alguna (–1) o máxima (–2) incertidumbre sobre si la evidencia es directa o indirecta
– Información imprecisa o escasa (–1)
– Alta probabilidad de sesgo de información (–1)

Aumentar el grado si

– Evidencia de asociación fuerte: riesgo relativo significativo > 2 (< 0,5) basado en evidencia consistente derivada de 2 o más estudios observacionales, sin factores de confusión plausibles (+1)[46]
– Evidencia de asociación muy fuerte: riesgo relativo significativo > 5 (< 0,2) basado en evidencia directa, sin amenazas importantes para la validez (+2)[46]
– Evidencia de un gradiente dosis respuesta (+1)
– Todos los factores de confusión plausibles habrían reducido el efecto (+1)

embargo, hay una incertidumbre considerable acerca de si ello tiene consecuencias importantes para la salud. Por eso, la evidencia a la hora de elaborar una recomendación se podría considerar de calidad baja.

Otro tipo de evidencia indirecta se plantea cuando no hay comparaciones directas entre las intervenciones y los investigadores deben hacer comparaciones entre los estudios. Por ejemplo, esto ocurre en caso de haber ensayos aleatorizados que comparan los inhibidores selectivos de la recaptación de serotonina con placebo y los tricíclicos con placebo, pero que no hubieran ensayos que comparan los inhibidores selectivos de la serotonina con los tricíclicos. Las comparaciones indirectas siempre conllevan mayor incertidumbre que las directas, debido a que todas las diferencias entre los estudios pueden afectar a los resultados[45].

Combinación de los 4 componentes

La calidad de la evidencia para cada resultado importante se puede determinar una vez considerados cada uno de los componentes mencionados: diseño del estudio, calidad, consistencia y si la evidencia es directa o indirecta. Nuestro enfoque primero determina la evidencia según el diseño de estudio y la categoriza en ensayos aleatorizados o en estudios observacionales (estudios de cohorte, estudios de casos y controles, análisis de series temporales interrumpidas y estudios controlados antes-después). Posteriormente, es aconsejable valorar si los estudios tienen limitaciones importantes en su diseño, inconsistencias importantes en sus resultados o si hay incertidumbre en cuanto a si la evidencia es directa (recuadro 2).

Para clasificar la calidad de la evidencia sugerimos las siguientes definiciones:

– Alta: es muy poco probable que nuevos estudios cambien la confianza que tenemos en el resultado estimado.
– Moderada: es probable que nuevos estudios tengan un impacto importante en la confianza que tenemos en el resultado estimado y que puedan modificar el resultado.
– Baja: es muy probable que nuevos estudios tengan un impacto importante en la confianza que tenemos en el resultado estimado y que puedan modificar el resultado.
– Muy baja: cualquier resultado estimado es muy incierto.

Las limitaciones en la calidad de los estudios, las inconsistencias importantes en los resultados o la incertidumbre en cuanto a si la evidencia es directa pueden disminuir el grado de evidencia. Por ejemplo, si todos los estudios de los que disponemos presentan limitaciones importantes, el grado de evidencia disminuirá un nivel, y si todos los estudios presentan limitaciones muy importantes, el grado de evidencia disminuirá 2 niveles. Los estudios con deficiencias graves pueden excluirse.

Otras factores que pueden disminuir la calidad de la evidencia son: datos imprecisos y escasos (recuadro 3) y riesgo elevado de sesgo de información. Entre las razones que pueden aumentar la calidad de la evidencia se incluyen:

– Presencia de una asociación muy fuerte (p. ej., un riesgo 50 veces superior de sobredosis mortal con los antidepresivos tricíclicos que con los inhibidores selectivos de la recaptación de serotonina (tabla 2) o de una asociación fuerte (p. ej., un riesgo 3 veces superior de lesiones en la cabeza entre los ciclistas que no utilizan casco comparado con los que sí lo usan)[47].
– Presencia de un gradiente dosis respuesta.
– La presencia de factores de confusión plausibles reduciría el efecto. Por ejemplo, al no haber ajustado por todos los posibles factores explicativos en los estudios que comparan las tasas de mortalidad de los hospitales con y sin ánimo de lucro, podríamos estar reduciendo el efecto observado[48]. De haber ajustado por todos los posibles factores, la evidencia de que el riesgo de mortalidad es mayor en los hospitales con ánimo de lucro sería aún más convincente.

Todos estos factores son acumulativos. Por ejemplo, si los ensayos aleatorizados presentan limitaciones importantes y también incertidumbre acerca de si la evidencia es directa, el grado de evidencia descendería 2 niveles, del alto al moderado y del moderado al bajo.

Estas reglas se deberían aplicar a los juicios sobre la calidad de la evidencia para riesgos y beneficios. Los riesgos importantes se deberían incluir en los resúmenes de las tablas de evidencia y se debería tener en cuenta si la evidencia que los hace plausibles es indirecta. Por ejemplo, si hay preocupación por la ansiedad que puede generar el cribado del melanoma, y no se dispone de evidencia directa, podría ser apropiado considerar la evidencia derivada de otros estudios de cribado.

Recuadro 3. Datos escasos o imprecisos

No se dispone de una base empírica para definir los datos como imprecisos o escasos. Dos posibles definiciones son:

Los datos son escasos si los resultados incluyen únicamente unos pocos eventos u observaciones y no son informativos
Los datos son imprecisos si los intervalos de confianza son suficientemente amplios para que un estimador del efecto sea consistente, tanto con los riesgos como con los beneficios importantes

Estas diferentes definiciones pueden conllevar juicios diferentes. Aunque pueda no ser posible reconciliar estas diferencias, para considerar si se ha de disminuir la calidad de la evidencia debido a que los datos son escasos o imprecisos, ofrecemos la siguiente guía:
El umbral para considerar datos imprecisos o escasos debería ser más bajo cuando sólo se dispone de un estudio. Si tenemos un solo estudio, cuyo tamaño de muestra es pequeño (o con escasos eventos) y con intervalos de confianza amplios, que indican tanto la presencia de riesgo como de beneficio, se debe considerar como datos imprecisos o escasos
Los intervalos de confianza que son suficientemente amplios para que, con independencia de otros resultados, la estimación sea compatible con recomendaciones contradictorias, se deben considerar como datos imprecisos o escasos

Los juicios sobre la calidad de la evidencia para resultados importantes a partir de los estudios pueden y deberían llevarse a cabo en el contexto de revisiones sistemáticas, tales como las revisiones Cochrane. Los juicios sobre la calidad global de la evidencia, el balance entre beneficios y riesgos y las recomendaciones requieren en general información adicional más allá de los resultados de una revisión.

Calidad global de la evidencia

Otros sistemas habitualmente basan los juicios sobre la calidad global de la evidencia en la calidad de la evidencia sobre los beneficios de las intervenciones. Cuando el riesgo de un efecto adverso es clave para una decisión y la evidencia en cuanto a este riesgo es más débil que la evidencia de beneficio, el ignorar la incertidumbre acerca de este riesgo es problemático. Nosotros consideramos que la calidad más baja de la evidencia disponible para cualquiera de los resultados clave para tomar una decisión debería ser la base para graduar la calidad global de la evidencia.

Los resultados que son importantes pero no claves deberían incluirse en las tablas de síntesis de evidencia y deberían considerarse al llevar a cabo juicios sobre el balance entre beneficios y riesgos, pero no al graduar la calidad global de la evidencia. Decidir si un resultado es clave, importante pero no clave, o no importante, es un juicio de valor. Siempre que sea posible, a la hora de hacer estos juicios se deberían tener en cuenta los valores de las personas que se verán afectadas al poner en práctica las recomendaciones.

Decidir qué es un resultado clave puede ser difícil. La plausibilidad de los resultados adversos puede influir en la decisión sobre si estos resultados se consideran claves. Una evidencia débil sobre posibles riesgos poco probables no debería disminuir el grado global de evidencia. Las decisiones sobre si un determinado riesgo es plausible pueden provenir de evidencia indirecta. Por ejemplo, si a partir de estudios realizados en animales hay una preocupación importante sobre los efectos adversos graves de un medicamento, la calidad global de la evidencia, basada en cualquier evidencia que haya disponible en humanos para aquel efecto adverso en concreto, debería descender un grado. A veces, no disponer de evidencia sobre un determinado riesgo plausible no permite evaluar el beneficio neto de una intervención. En tales circunstancias, el grupo de elaboración de la guía puede optar por recomendar que se realicen nuevos estudios.

Si la evidencia para todos los resultados clave favorece la misma alternativa y hay evidencia de alta calidad sólo para algunos de estos resultados, la calidad global de la evidencia todavía se podría considerar alta. Por ejemplo, hay evidencia de calidad alta de que los antiagregrantes reducen el riesgo de ictus y de infarto de miocardio en los pacientes que han tenido un infarto de miocardio. Aunque la evidencia para la mortalidad por todas las causas es de calidad moderada, la calidad global de la evidencia todavía podría considerarse alta, incluso a pesar de que la mortalidad por todas las causas se considere un resultado clave.

Recomendaciones

¿Produce la intervención más beneficios que riesgos?

Al llevar a cabo las recomendaciones se considera el balance entre beneficios y riesgos. Este balance conlleva de manera inevitable asignar, implícita o explícitamente, un valor relativo a cada resultado. A menudo es difícil juzgar qué peso asignar a los diferentes resultados observados, y con frecuencia diferentes personas expresan valores distintos. Las personas que llevan a cabo los juicios en nombre de otras personas afectadas podrán hacerlo mejor si conocen sus valores. Por ejemplo, las personas que elaboren recomendaciones sobre la quimioterapia para mujeres con cáncer de mama diagnosticado en un estadio temprano lo podrán hacer mejor si conocen la importancia que esas mujeres dan a la disminución del riesgo de recurrencia de cáncer de mama en relación con la importancia que otorgan a los efectos adversos de la quimioterapia.

Es aconsejable que los juicios sobre el balance entre los beneficios importantes en salud y los riesgos se lleven a cabo antes de considerar los costes. ¿Son superiores los beneficios de la intervención a los riesgos? Cuando los beneficios y los riesgos varían en diferentes ámbitos o grupos de pacientes, las recomendaciones se deben adaptar a cada ámbito específico y a cada grupo de pacientes en concreto. Por ejemplo, consideremos si se de-

be recomendar warfarina en pacientes con fibrilación auricular para reducir el riesgo de ictus, a pesar de que podría incrementar el riesgo de sangrado. Las recomendaciones o su fuerza diferirán entre los ámbitos de práctica clínica donde se pueda controlar de forma regular el grado de anticoagulación y los ámbitos donde no se pueda controlar. Además, las recomendaciones o su fuerza probablemente serán diferentes para los pacientes con riesgo muy bajo de ictus (menores de 65 años sin ninguna comorbilidad) que para los que presentan un riesgo más alto (más mayores con insuficiencia cardíaca), a causa de las diferencias en la reducción absoluta del riesgo. Por tanto, las recomendaciones deben ser definidas para cada grupo de pacientes y cada ámbito de práctica clínica. Al elaborar recomendaciones, es especialmente importante considerar las particularidades de las poblaciones más desfavorecidas y, si es apropiado, modificar las recomendaciones para tener en cuenta estas desigualdades.

Para categorizar el balance entre beneficios y riesgos aconsejamos usar las siguientes definiciones:

– Beneficios netos: la intervención claramente comporta más beneficios que riesgos.
– Beneficios con aceptación de los riesgos: la intervención comporta beneficios y riesgos.
– Beneficios inciertos con aceptación de los riesgos: no está claro que la intervención comporte más beneficios que riesgos.
– Ausencia de beneficios netos: la intervención claramente no comporta más beneficios que riesgos.

Las personas que elaboran recomendaciones deberían tener en cuenta 4 factores importantes:

– El balance entre beneficios y riesgos, teniendo en cuenta la magnitud del efecto estimado para los resultados importantes, los intervalos de confianza de las estimaciones y la importancia relativa asignada a cada resultado.
– La calidad de la evidencia.
– El trasladar la evidencia a la práctica clínica en un ámbito específico, teniendo en cuenta los factores importantes que podrían modificar la magnitud del efecto esperado, como puede ser la proximidad a un hospital o el disponer de la experiencia necesaria.
– La incertidumbre sobre el riesgo basal de la población de interés.

Nuestra confianza en una recomendación podría disminuir cuando haya incertidumbre para trasladar la evidencia a la práctica clínica en un ámbito específico, o cuando haya incertidumbre sobre el riesgo basal de la población. Por ejemplo, ante una intervención que comporta efectos adversos graves pero también beneficios importantes, la recomendación probablemente será más incierta si se desconoce el riesgo basal de la población de interés.
Las categorías que sugerimos en cuanto a las recomendaciones son las siguientes:

–«Hazlo» o «No lo hagas»: se refiere a la decisión que tomaría la mayoría de personas bien informadas.
– «Probablemente hazlo» o «Probablemente no lo hagas»: se refiere a la decisión que tomaría la mayoría de personas bien informadas, aunque una minoría considerable no lo haría.

Establecer una recomendación, ya sea a favor o en contra de una intervención, no significa que todos los pacientes deban ser tratados de la misma manera. Tampoco significa que los médicos no puedan involucrar a los pacientes en la decisión, ni explicarles las ventajas de las alternativas. Sin embargo, como la mayoría de pacientes bien informados escogerán la misma opción, las explicaciones sobre las ventajas de las alternativas pueden ser relativamente concisas. Una recomendación intenta facilitar que se tome una decisión apropiada para un paciente en concreto o para una población. Por tanto, una recomendación debería reflejar lo que las personas probablemente escogerían a partir de la evidencia disponible y de sus propios valores o preferencias en relación con los resultados esperados. Una recomendación de «Probablemente hazlo» implica que los médicos, al proponer una intervención, consideren los valores y las preferencias de los pacientes de forma más completa y detenida.
En algunos casos, ya sea porque el balance entre beneficios y riesgos es incierto o porque no haya acuerdo (como se ilustra en el recuadro 4), puede no ser apropiado establecer una recomendación. Si esto ocurre por falta de evidencia de buena calidad, se debería recomendar poner en marcha investigaciones específicas que proporcionen la evidencia necesaria para fundamentar una recomendación.

¿Justifica los costes adicionales un incremento en los beneficios para la salud?
Dado que el dinero utilizado para un determinado fin deja de estar disponible para otros fines, las recomendaciones dependen, ya sea implícita o explícitamente, de los juicios sobre la relación entre el incremento de los beneficios para la salud y el aumento de los costes. Los costes –el valor monetario de los recursos utilizados– son un factor importante a la hora de elaborar recomendaciones, pero son específicos de cada contexto, cambian con el tiempo y son difíciles de cuantificar. Aun reconociendo la dificultad de cuantificar los costes de forma precisa, sugerimos que el incremento del coste de las alternativas de salud se considere de forma explícita, junto con los beneficios en salud y los riesgos esperados. Cuando sea necesario y posible, los costes deberían presentarse de forma desagregada (diferencias en los recursos utilizados) en la tabla de síntesis de

Recuadro 4. Los valores no son o correctos o incorrectos

El ejemplo siguiente muestra cómo distintas personas podrían elaborar recomendaciones distintas debido a las diferencias en los valores, incluso después de estar de acuerdo sobre la evidencia

Pregunta: ¿Se debería ofrecer el cribado del melanoma a la población general?

Ámbito: atención primaria en Estados Unidos

Riesgo basal: población general (la incidencia de melanoma en el año 1995 fue de 13,3 por 100.000)

Referencia: Helfand et al. Screening for skin cancer. Systematic evidence review No. 2
Rockville, MD: Agency for Healthcare Research and Quality. April 2001. (AHRQ Publication No 01-S002.)

Hay evidencia de calidad muy baja sobre la validez del cribado del melanoma y sus resultados en términos de mortalidad. Los riesgos potenciales del cribado incluyen las consecuencias de las pruebas falsas positivas, pero no se dispone de evidencia sobre aquéllos. Basándonos en esto, es posible concluir que la calidad de la evidencia es muy baja y que hay beneficios netos inciertos sobre este cribado. Basado en un solo estudio de casos y controles, se estimó una *odds ratio* de 0,37 para las personas cribadas frente a las no cribadas. En los varones blancos, el riesgo de morir de melanoma se estimó en 0,36%. Basándose en esta evidencia, muchas personas podrían hacer una recomendación de «no hacer cribado», ya que asignan un valor elevado a evitar el potencial, aunque desconocido, riesgo del cribado en las personas sanas, con relación a los inciertos beneficios. Sin embargo, algunas personas podrían recomendar «probablemente hacer cribado» debido a la asignación de un elevado valor sobre los pequeños, pero potencialmente importantes beneficios del cribado, en relación con los potenciales riesgos desconocidos. Bajo estas circunstancias, después de tener en cuenta los costes, un grupo de personas que desarrolle guías podría decidir no hacer una recomendación para la práctica clínica y realizar una recomendación específica en cuanto a la investigación que es necesaria para disminuir la incertidumbre y clarificar el balance entre los beneficios y riesgos.

Este es un ejemplo típico sobre los juicios de valor en que se basan las recomendaciones de cribado, aunque estas cuestiones también surgen al elaborar recomendaciones sobre tratamientos, ya sea en enfermedades agudas o crónicas, donde siempre es necesario buscar el balance entre los beneficios y los riesgos esperados, a la luz del valor relativo que se asigna a cada resultado importante y a la incertidumbre

la evidencia, junto con los resultados importantes observados. La calidad de la evidencia para las diferencias en el uso de los recursos se debería clasificar utilizando los mismos criterios descritos anteriormente para los otros resultados importantes.

¿Cómo funciona este sistema en la práctica?

En la tabla 2 se muestra un ejemplo de este sistema de clasificación aplicado a la evidencia procedente de una revisión sistemática llevada a cabo en 1997, que compara los inhibidores selectivos de la recaptación de serotonina con los antidepresivos tricíclicos[49]. Tras valorar la información disponible, acordamos que la evidencia sobre los efectos de los inhibidores selectivos de la recaptación de serotonina y de los antidepresivos tricíclicos, respecto a la gravedad de la depresión y el riesgo de sobredosis mortal, era de calidad moderada, y la evidencia para los efectos adversos transitorios de calidad alta. A continuación, acordamos que la calidad global de la evidencia era moderada y que había beneficios netos a favor de los inhibidores selectivos (sin diferencias respecto a la gravedad de la depresión y con menos efectos transitorios adversos y sobredosis mortales). A pesar de estar de acuerdo en que podía haber beneficios netos, se optó por una recomendación de tipo «Probablemente hazlo» respecto a la utilización de los inhibidores selectivos, reflejando así la incertidumbre sobre la calidad de la evidencia. Para realizar este ejercicio no dispusimos de evidencia sobre los costes que suponía utilizar los inhibidores selectivos de la recaptación de serotonina en comparación con los antidepresivos tricíclicos. De haber considerado los costes, la recomendación podría haber sido otra.

Conclusiones

Cualquier sistema para clasificar la calidad de la evidencia y fuerza de la recomendación necesita equilibrar sencillez y claridad. Al disminuir la complejidad de un sistema es probable que también reduzcamos su claridad ya que, probablemente, los juicios en los sistemas más sencillos se llevan a cabo de forma más implícita que explícita. Por otro lado, los intentos por mejorar la claridad y llevar a cabo juicios más transparentes probablemente den como resultado una mayor complejidad. En el sistema que hemos descrito se ha intentado encontrar un equilibrio entre la sencillez y la claridad. A pesar de lo sencillo o complejo que sea un sistema, siempre es necesario llevar a cabo juicios. El enfoque que

Puntos clave

- Las organizaciones han utilizado múltiples sistemas para clasificar la calidad de la evidencia y la fuerza de las recomendaciones.
- Las diferencias y limitaciones de estos sistemas de clasificación pueden confundirnos e impedir una comunicación eficaz.
- Presentamos un enfoque sistemático y explícito para hacer juicios sobre la calidad de la evidencia y la fuerza de las recomendaciones.
- El enfoque tiene en cuenta el diseño del estudio, su calidad, la consistencia y si la evidencia es directa o indirecta, a la hora de valorar la calidad de la evidencia para cada resultado importante.
- El balance entre beneficios y riesgos, la calidad de la evidencia, si la evidencia es directa o indirecta y el riesgo basal, son aspectos a tener en cuenta cuando se llevan a cabo juicios sobre la fuerza de las recomendaciones.

TABLA 2 **Evaluación de la calidad de los ensayos que comparan los inhibidores selectivos de la recaptación de serotonina con los antidepresivos tricíclicos en el tratamiento de la depresión moderada en atención primaria**

	Evaluación de la calidad					Resumen de los hallazgos					
						N.º de pacientes		Efecto			
N.º de estudios	Diseño	Calidad	Consistencia	Evidencia directa o indirecta	Otros factores a tener en cuenta[a]	ISRS	Tricíclicos	Relativo (IC del 95%)	Absoluto	Calidad	Importancia
Gravedad de la depresión (medida con la escala de depresión de Hamilton, tras 4-12 semanas de tratamiento)											
Citalopram (8)	Ensayo clínico aleatorizado	Sin limitaciones importantes	Inconsistencia no importante	Alguna incertidumbre acerca de si la evidencia es directa (medida de resultado)[b]	Ninguna	5.044	4.510	DMP 0,034 (–0,007 a 0,075)	Sin diferencias	Moderada	Clave
Fluoxetina (38)											
Fluvoxamina (25)											
Nefazodona (2)											
Paroxetina (18)											
Sertralina (4)											
Venlafaxina (4)											
Efectos adversos transitorios que causan interrupción del tratamiento											
Citalopram (8)	Ensayo clínico aleatorizado	Limitaciones no importantes	Inconsistencia no importante	Directa	Ninguna	1.948/ 7.032 (28%)	2.072/ 6.334 (33%)	RRR 13% (5-20%)	5/100	Alta	Clave
Fluoxetina (50)											
Fluvoxamina (27)											
Nefazodona (4)											
Paroxetina (23)											
Sertralina (6)											
Venlafaxina (5)											
Intoxicación fatal[d]											
Oficina del Reino Unido para las estadísticas nacionales (1)	Datos observacionales	Limitaciones importantes[c]	Solo un estudio	Directa	Asociación muy fuerte	1/100 000/ años de tratamiento	58/100.000 años de tratamiento	RRR 98% (97-99%)[d]	6/10.000	Moderada	Clave

DMP: diferencia de medias ponderadas; RRR: reducción riesgo relativo; ISRS: inhibidores selectivos de la recaptación de serotonina; tricíclicas: antidepresivos tricídicos.
[a]Datos escasos o imprecisos, asociación fuerte o muy fuerte, riesgo elevado de sesgo de información, gradiente dosis-respuesta, efecto de las variables de confusión plausibles residuales.
[b]Hay incertidumbre acerca de si la evidencia de la variable de resultado es directa debido a la breve duración del ensayo.
[c]Es posible que las personas de menor riesgo tomen con mayor probabilidad ISRS y es incierto que de cambiar los antidepresivos se habrían podido disuadir los intentos de suicidio.
[d]Hay incertidumbre sobre el riesgo basal para los casos de sobredosis.

describimos proporciona un marco para una reflexión estructurada y puede ayudar a asegurar que se realizan los juicios pertinentes, pero no elimina la necesidad de realizarlos.

Miembros del grupo de trabajo para la elaboración, evaluación, desarrollo y evaluación de los grados de recomendación (GRADE) que han contribuido en este artículo: David Atkins, Dana Best, Peter A. Briss, Martin Eccles, Yngve Falck-Ytter, Signe Flottorp, Gordon H. Guyatt, Robin T. Harbour, Margaret C. Haugh, David Henry, Suzanne Hill, Roman Jaeschke, Gillian Leng, Alessandro Liberati, Nicola Magrini, James Mason, Philippa Middleton, Jacek Mrukowicz, Dianne O'Connell, Andrew D. Oxman, Bob Phillips, Holger J Schünemann, Tessa Tan-Torres Edejer, Helena Varonen, Gunn E. Vist, John W. Williams Jr y Stephanie Zaza.

El National Institute for Clinical Excellence (NICE) de Inglaterra y País de Gales y el Polish Institute for Evidence-Based Medicine (PIEBM) han proporcionado apoyo para las reuniones del grupo de trabajo GRADE. Las instituciones a las que están afiliados los miembros del grupo de trabajo han proporcionado apoyo institucional. La participación de Alessandro Liberati en las actividades de GRADE ha sido apoyada por una beca del Ministero Università e Ricerca Scientifica (M.I.U.R., Progetto COFIN 2001).

Contribuidores

Todos los miembros del grupo de trabajo GRADE citados anteriormente han contribuido en la preparación de este manuscrito y el desarrollo de las ideas incluidas en él, han participado en al menos una reunión y han leído y comentado los borradores de este artículo. G.H.G. y A.D.O. lideraron el proceso. G.E.V. preparó las tablas de síntesis de evidencia utilizadas en el estudio piloto y coordinó el proceso.

Conflictos de interés

La mayoría de los miembros del grupo de trabajo GRADE tienen un interés personal en otros sistemas de clasificación de la calidad de la evidencia y la fuerza de las recomendaciones.

Agradecimientos

Agradecemos las aportaciones y comentarios de todas las personas que participaron en el primer seminario GRADE que se realizó en Vitoria-Gasteiz en el mes de abril de 2005.

Bibliografía

1. Canadian Task Force on the Periodic Health Examination. The periodic health examination. CMAJ. 1979;121:1193-254.
2. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1986;89 Suppl 2:S2-3.
3. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Arch Intern Med. 1986; 146:464-5.
4. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989;95:S2-4.
5. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Antithrombotic therapy consensus conference. Chest. 1992;102 Suppl 4:S305-11.
6. US Department of Health and Human Services, Public Health Service, Agency Health Care Policy and Research. Acute pain management: operative or medical procedures and trauma. Rockville, MD: Agency for Health Care Policy and Research Publications, 1992. (AHCPR Pub 92-0038.)
7. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Oxman AD, Scott EA, Millson ME, et al. An approach to the development of practice guidelines for community health interventions. Can J Public Health. 1994;85 Suppl 1:S8-13.
8. Hadorn DC, Baker D. Development of the AHCPR-sponsored heart failure guideline: methodologic and procedural issues. Jt Comm J Qual Improv. 1994;20:539-54.
9. Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ. Clinical recommendations using levels of evidence for antithrombotic agents. Chest. 1995;108 Suppl 4:S227-30.
10. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, et al. Users' guide to the medical literature IX: a method for grading health care recommendations. JAMA. 1995;274:1800-4.
11. Scottish Intercollegiate Guidelines Network (SIGN). Forming guideline recommendations. En: A guideline developers' handbook. Edinburgh: SIGN, 2001. (Publication No 50.) [citado 8 Feb 2004]. Disponible en: www.sign.ac.uk/guidelines/fulltext/50/section6.html
12. US Preventive Services Task Force. Guide to clinical preventive services. 2nd ed. Baltimore: Williams and Wilkins; 1996. p. 39-49.
13. Eccles M, Clapp Z, Grimshaw J, Adams PC, Higgins B, Purves I, et al. North of England evidence based guidelines development project: methods of guideline development. BMJ. 1996;312:760-2.
14. Centro per la Valutazione della Efficacia della Assistenza Sanitaria (CeVEAS). Schema di grading CeVEAS [citado 18 May 2004]. Disponible en: http://web1.satcom.it/interage/ceveas/html/doc/45/GLICO.pdf
15. Guyatt G, Schünemann H, Cook D, Jaeschke R, Pauker S, Bucher H. Grades of recommendation for antithrombotic agents. Chest. 2001;119:S3S-7 [citado 8 Feb 2004]. Disponible en: www.chestjournal.org/content/vol119/1_suppl/
16. Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M. Levels of evidence and grades of recommendations. Oxford: Oxford Centre for Evidence-Based Medicine [citado 8 Feb 2004]. Disponible en: www.cebm.net/levels_of_evidence.asp
17. National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. Canberra: AusInfo, 2000 [citado 8 Feb 2004]. Disponible en: www.health.gov.au/nhmrc/publications/pdf/cp69.pdf
18. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ. 2001;323:334-6.
19. Roman SH, Silberzweig SB, Siu AL. Grading the evidence for diabetes performance measures. Eff Clin Pract. 2000;3:85-91.
20. Woloshin S. Arguing about grades. Eff Clin Pract. 2000;3:94-5.
21. Guyatt GH, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, et al. Grades of recommendation for antithrombotic agents. Chest. 2001;119:S3-7.
22. Atkins D, Best D, Shapiro EN, editores. Third US Preventive Services Task Force: background, methods and first recommendations. Am J Prev Med. 2001;20:3 Suppl:1-108.
23. Woolf SH, Atkins D. The evolving role of prevention in health care: contributions of the US Preventive Services Task Force. Am J Prev Med. 2001;20:3 Suppl:13-20.
24. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20:3 Suppl:21-35.
25. Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Agüero L, et al. Developing an evidence-based guide to community preventive services—methods. Am J Prev Med. 2000;18 Suppl 1:35-43.
26. Zaza S, Wright-De A, Briss PA, Truman BI, Hopkins DP, Hennessy MH, et al. Data collection instrument and procedure for systematic reviews in the guide to community preventive services. Am J Prev Med. 2000;18 Suppl 1:44-74.
27. Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. Jt Comm J Qual Improv. 2000;26:700-12.
28. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Rockville: Agency for Healthcare Research and Quality. 2002. p. 64-88. (AHRQ publication No 02-E016.)
29. Guyatt G, Drummond R, eds. Users' guide to the medical literature. Chicago: AMA Press; 2002. p. 55-154.
30. Clarke M, Oxman AD, eds. Assessment of study quality. Cochrane reviewers' handbook 4.1.5 section 6. En: Cochrane Library. Issue 4. Oxford: Update Software; 2002.
31. Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. En: Egger M, Davey Smith G, Altman DG, editores. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001. p. 87-121.
32. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Rockville: Agency for Healthcare Research and Quality; 2002. p. 51-63. (AHRQ publication No 02-E016.)

33. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in health-care trials (Cochrane methodology review). En: Cochrane Library Issue 4. Oxford: Update Software; 2002.
34. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001;286: 821-30.
35. Kleijnen J, Gøtzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomisation? En: Chalmers I, Maynard A, editores. Non-random reflections on health care research: on the 25th anniversary of Archie Cochrane's effectiveness and efficiency. London: BMJ; 1997. p. 93-106.
36. Lacchetti C, Guyatt G. Surprising results of randomized controlled trials. En: Guyatt G, Drummond R, editores. Users' guide to the medical literature. Chicago: AMA Press; 2002. p. 247-65.
37. 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. JAMA. 1998;280:605-13.
38. Writing Group for the Women's Health Initiative Investigators. 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.
39. Dans A, McAlister F, Dans L, Richardson WS, Straus S, Guyatt G. Applying results in individual patients. En: Guyatt G, Drummond R, editores. Users' guide to the medical literature. Chicago: AMA Press; 2002. p. 369-84.
40. McAlister F, Laupacis A, Wells G. Drug class effects. En: Guyatt G, Drummond R, editores. Users' guide to the medical literature. Chicago: AMA Press; 2002. p. 415-31.
41. 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.
42. Riggs BL, Hodgson SF, O'Fallon WM, Chao EY, Wahner HW, Muhs JM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. N Engl J Med. 1990;322:802-9.
43. 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.
44. Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? Can Assoc Radiol J. 2002;53:241.
45. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. BMJ. 2003;326:472.
46. Bross IDJ. Pertinency of an extraneous variable. J Chron Dis. 1967;20:487-95.
47. Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. Cochrane Database Syst Rev. 2000;2:CD001855.
48. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schünemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ. 2002;166:1399-406.
49. North of England Evidence Based Guideline Development Project. Evidence based clinical practice guideline: the choice of antidepressants for depression in primary care. Newcastle upon Tyne: Centre for Health Services Research; 1997.