formación

GESTIÓN DEL CONOCIMIENTO

Niveles de evidencia y fuerza de las recomendaciones: necesidad de homogeneización

J. González de Dios

Hospital General Universitario. Alicante. Codirector de la revista *Evidencias en Pediatría*

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 de dicha intervención, por un lado, y los riesgos, los inconvenientes y los costes, por el otro, es lo suficientemente favorable como para que merezca la pena aplicarse. Cada vez es más habitual tomar decisiones médicas fundamentadas en el mejor nivel de evidencia (indica hasta qué punto nuestra confianza en la estimación de un efecto es adecuada para apoyar una recomendación) y la mayor fuerza de recomendación (indica hasta qué punto podemos confiar en que poner en práctica la recomendación conllevará más beneficios que inconvenientes).

El **nivel de evidencia** se ha relacionado, generalmente, con el diseño del estudio (estudios descriptivos o analíticos, observacionales o experimentales) y su calidad. La meta de la investigación es la agudeza en la medición, lo que implica precisión (limitar el error aleatorio) y validez (limitar el error sistemático). En este sentido, por las características propias de cada diseño, el «nivel» de evidencia será mayor en los estudios analíticos que en los descriptivos, y superior en los estudios experimentales (p. ej., ensayo clínico) que en los observacionales (p. ej., estudios de cohortes y estudios de casos y controles). Sin embargo, no toda pregunta clínica se puede abordar con el mismo diseño científico: el ensayo clínico es el patrón de referencia para intervenciones terapéuticas, pero no será el diseño apropiado para preguntas sobre diagnóstico o pronóstico.

Se establecen unos criterios de calidad propios para cada tipo de diseño. Así, podemos considerar cinco criterios de calidad en el ensayo clínico: definición clara de la población de estudio, intervención y resultado de interés; correcta aleatorización; adecuado enmascaramiento; seguimiento completo (menos del 20% de pérdidas), y análisis correcto (análisis por intención de tratar y control de covariables no equilibradas con la aleatorización). Éstos serán diferentes a los criterios de calidad barajados en el caso de estudios de valoración de pruebas diagnósticas: comparación con un patrón de referencia válido; muestra representativa; descripción completa de los métodos de realización de la prueba diagnóstica; control de sesgos (comparación ciega e independiente; control de sesgos de incorporación, verificación diagnóstica y revisión), y análisis correcto (datos que permitan calcular indicadores de validez). Asimismo, pueden ser distintos de los criterios establecidos en los estudios de cohortes: cohortes representativas de la población con y sin exposición, libres del efecto o enfermedad de interés; medición independiente, ciega y válida de exposición y efecto; seguimiento suficiente (superior al 80%), completo y no diferencial; control de la relación temporal de los acontecimientos (exposición/efecto) y de la relación entre el nivel de exposición y el grado de efecto (dosis/respuesta), y análisis correcto (control de factores de confusión y modificadores de efecto).

La **fuerza de las recomendaciones** indica hasta qué punto podemos confiar en que poner en práctica la recomendación conllevará más beneficios que inconvenientes. En la elaboración de las recomendaciones se debe tener en cuenta, en primer lugar, el nivel de evidencia, pero también otras consideraciones: balance entre beneficios y riesgos, consistencia de los estudios, aplicabilidad práctica en un paciente o población (incluido el riesgo basal en mi población), valores y preferencias de la población diana a la cual va dirigida, costes, etc. Establecer una recomendación a favor o en contra de una intervención no significa que todos los pacientes deban ser tratados de la misma manera, pues en la toma de decisión la evidencia procedente de la investigación es sólo uno de los cuatro círculos en una toma de decisiones basada en pruebas

Correspondencia: J. González de Dios. Prof. Manuel Sala, 6. 3.º A. 03003 Alicante.
Correo electrónico: javier.gonzalezdedios@gmail.com

Figura 1. Modelo actualizado en la toma de decisiones basada en pruebas

(figura 1), en donde será muy importante considerar el estado clínico y las circunstancias del paciente, sus preferencias y acciones, todo ello encuadrado en la experiencia clínica del profesional.

Desde que hace más de 25 años el Canadian Task Force on Preventive Health Care (CTFPHC) introdujo el primer sistema de clasificación de los niveles de evidencia y fuerza de las recomendaciones, numerosas organizaciones e instituciones han ido desarrollando sus propios sistemas. Y, actualmente, se contabilizan más de cien, en general bastante similares, pero que pueden generar confusión en el lector, pues se catalogan con letras y/o números diferentes, con significados no siempre iguales. Entre las limitaciones presentes en los sistemas actuales se encuentran la falta de transparencia en el paso de la evidencia a la recomendación y en la valoración del balance entre beneficios y riesgos. Otra limitación es que la mayoría de los sistemas de clasificación se orientan principalmente a intervenciones terapéuticas y/o preventivas, con escasa consideración de otros tipos de intervenciones sanitarias (diagnósticas, pronósticas, etc.).

Desde el año 2000, un grupo internacional de expertos en metodología, epidemiólogos y clínicos, muchos de ellos procedentes de las organizaciones que establecieron las clasificaciones más conocidas en la formulación de recomendaciones (CTFPHC, US Preventive Service Task Force, Oxford Center for Evidence Based Medicine [CEBM], Scottish Intercollegiate Guidelines Network, National Institute for Clinical Excellence, Organización Mundial de la Salud, etc.), ha elaborado una nueva propuesta que tiene como objetivo consensuar un sistema común que supere las limitaciones detectadas hasta el momento en los sistemas previos. Este grupo de profesionales constituye el grupo de trabajo GRADE (Grading of Recommendations Assessment, Development and Evaluation) y puede ser el futuro para homogeneizar el área de los niveles de evidencia y la fuerza de las recomendaciones, algo que se considera ya más que necesario.

En la tabla 1 se expone, como ejemplo de los múltiples sistemas de clasificación de los niveles de evidencia y fuerza de las recomendaciones en ciencias de la salud, el sistema basado en el CEBM, uno de los más exhaustivos y que considera también en detalle los diferentes tipos de estudios (etiología, diagnóstico, tratamiento, prevención, efectos adversos, pronóstico, análisis económicos y análisis de decisiones).

El **sistema GRADE** tiene algunos aspectos diferenciales que se pueden resumir en los siguientes puntos:

1. Inicialmente categoriza las variables de resultado y su importancia relativa. Se sugiere clasificar su importancia mediante la siguiente escala de nueve puntos: 1-3 (variable de resultado no importante), 4-6 (variable de resultado importante, pero no clave para la toma de decisiones) y 7-9 (variable de resultado clave para la toma de decisiones).
2. Posteriormente evalúa la calidad de la evidencia en cuatro categorías para cada una de las variables de resultado, y después la calidad global de la evidencia. Inicialmente la evaluación de la evidencia comienza por considerar el diseño de los estudios y su adecuación para responder a cada tipo de pregunta. En este caso, los ensayos clínicos corresponden de entrada a una «calidad alta» y los observacionales a una «calidad baja». En los ensayos clínicos se deben valorar cinco aspectos que pueden disminuir la calidad, mientras que en los estudios observacionales son tres las circunstancias que pueden aumentarla.

Los aspectos que pueden disminuir la calidad de los ensayos clínicos son los siguientes:

- Limitaciones en el diseño o la ejecución: ausencia de ocultamiento de la secuencia de aleatorización, enmascaramiento inadecuado, pérdidas de seguimiento importantes, ausencia de análisis por intención de tratar, y finalización prematura del estudio por razones de beneficio.
- Resultados inconsistentes: las estimaciones muy diferentes del efecto del tratamiento (heterogeneidad o variabilidad en los resultados) entre los estudios disponibles sugieren diferencias reales en dichas estimaciones. Éstas pueden deberse a diferencias en la población, la intervención, las variables de resultado o la calidad de los estudios.
- Ausencia de evidencia directa, que puede presentarse en varias circunstancias: ausencia de comparaciones directas entre dos tratamientos (la evidencia disponible puede provenir de una comparación indirecta de cada

Tabla 1. Niveles de evidencia y grado de recomendación (CEBM 2009)

Fuerza de la recomendación	Nivel de evidencia	Tratamiento, prevención, etiología y complicaciones	Pronóstico	Diagnóstico	Análisis económico y análisis de decisiones
A	1a	RS de ECA (con homogeneidad)[a]	RS (con homogeneidad)[a] de estudios de cohortes; RDC[b] validadas en diferentes poblaciones	RS (con homogeneidad)[a] de estudios de diagnóstico de nivel 1; RDC[b] con estudios 1b de diferentes centros clínicos	RS (con homogeneidad)[a] de estudios económicos de nivel 1
	1b	Un ECA (con intervalo de confianza estrecho)[c]	Un estudio de cohortes con seguimiento >80%; RDC[b] validada en una sola población	Estudios de cohortes que validen la calidad de una prueba específica[h], con unos buenos[g] estándares de referencia; o RDC probadas en un solo centro clínico	Análisis basados en los costes clínicos o en sus alternativas; RS de la evidencia; e inclusión de análisis de sensibilidad de múltiples vías
	1c	Todos o ninguno[d]	Series de casos (todos o ninguno)	Pruebas diagnósticas con especificidad tan alta que un resultado positivo confirma el diagnóstico y con sensibilidad tan alta que un resultado negativo descarta el diagnóstico	Análisis en términos absolutos de riesgos y beneficios clínicos: claramente tan buenas o mejores, pero más baratas; claramente tan malas o peores pero más caras
B	2a	RS (con homogeneidad)[a] de estudios de cohortes	RS (con homogeneidad)[a] de estudios de cohortes retrospectivos o de ECA con grupos control no tratados	RS (con homogeneidad)[a] de estudios de diagnóstico de nivel >2	RS (con homogeneidad)[a] de estudios económicos de nivel >2
	2b	Un estudio de cohortes (incluido un ECA de baja calidad; p. ej., seguimiento <80%)	Estudio de cohortes retrospectivo o seguimiento de controles no tratados en un ECA; derivación de una RDC[b] o RDC validada sólo en una muestra aislada[f]	Estudios exploratorios[h] de cohortes con buenos[g] estándares de referencia; RDC[b] tras derivación, o validada sólo en una muestra aislada[f] o en bases de datos	Análisis basados en costes clínicos o en sus alternativas; RS con evidencia limitada; estudios individuales; e inclusión de análisis de sensibilidad de múltiples vías
	2c	Investigación de resultados en salud; estudios ecológicos	Investigación de resultados en salud[i]		Auditoría de resultados en salud
	3a	RS (con homogeneidad)[a] de estudios de casos y controles		RS (con homogeneidad)[a] de estudios 3b y mejores	RS (con homogeneidad)[a] de estudios 3b y mejores
	3b	Un estudio de casos y controles		Estudio no consecutivo, o en el que el estándar de referencia no se aplica a todos los pacientes del estudio	Análisis sin medidas de coste precisas, pero incluyen un análisis de sensibilidad que incorpora variaciones clínicamente sensibles en las variables importantes
C	4	Series de casos (y estudios de cohortes y de casos y controles de baja calidad)[e]	Series de casos (y estudios de cohortes de pronóstico, de baja calidad)[i]	Estudios de casos y controles con estándares de referencia de poca calidad o no independientes	Análisis que no incluye análisis de la sensibilidad

Continúa en la página siguiente

Tabla 1. Niveles de evidencia y grado de recomendación (CEBM 2009) *(continuación)*

Fuerza de la recomendación	Nivel de evidencia	Tratamiento, prevención, etiología y complicaciones	Pronóstico	Diagnóstico	Análisis económico y análisis de decisiones
D	5	Opinión de expertos sin valoración crítica explícita, ni basada en fisiología ni en investigación juiciosa ni en los principios fundamentales	Opinión de expertos sin valoración crítica explícita, ni basada en fisiología ni en investigación juiciosa ni en los principios fundamentales	Opinión de expertos sin valoración crítica explícita, ni basada en fisiología ni en investigación juiciosa ni en los principios fundamentales	Opinión de expertos sin valoración crítica explícita, o basada en la teoría económica o en los principios fundamentales

Basada en: Oxford Centre for Evidence-based Medicine Levels of Evidence (marzo de 2009).
ECA: ensayo clínico aleatorizado; RDC: regla de decisión clínica; RS: revisión sistemática.
Notas: Los usuarios pueden añadir un signo menos (–) para marcar el nivel de lo que falla para poder dar una respuesta concluyente, porque se trata de un resultado aislado con un intervalo de confianza amplio o una RS con heterogeneidad importante. Esta evidencia no será concluyente y, por tanto, sólo genera un grado D de recomendación.
[a]Por homogeneidad se entiende una RS libre de variaciones preocupantes o importantes (heterogeneidad) en las direcciones y grados de los resultados entre los estudios individuales. No todas las RS con heterogeneidad estadísticamente significativa tienen que ser preocupantes, y no toda heterogeneidad preocupante ha de ser estadísticamente significativa. Como ya se ha señalado, los estudios que presentan heterogeneidad importante deben ser etiquetados con un signo menos (–) al final de su nivel adjudicado.
[b]Regla de decisión clínica (son algoritmos o sistemas de puntuación que dan lugar a una estimación del pronóstico o a una categoría diagnóstica)
[c]Véase la nota anterior como ayuda de cómo entender, clasificar y usar ensayos u otros estudios con intervalos de confianza amplios.
[d]Se da cuando todos los pacientes fallecieron antes de que el tratamiento estuviera disponible, pero ahora algunos sobreviven con él, o cuando algunos pacientes murieron antes de que el tratamiento estuviera disponible, pero ahora ninguno muere con él.
[e]Por estudio de cohortes de baja calidad se entiende el que no define claramente los grupos de comparación y/o no mide la exposición y el resultado de la misma manera objetiva (preferentemente ciega), tanto en individuos expuestos como en no expuestos, y/o no identifica o controla adecuadamente los factores de confusión conocidos y/o no lleva a cabo un seguimiento suficientemente largo y completo de los pacientes. Por estudio de casos y controles de baja calidad se entiende el que no define claramente los grupos de comparación y/o no mide la exposición y el resultado de la misma manera objetiva (preferentemente ciega) tanto en casos como en controles, y/o no identifica o controla adecuadamente los factores de confusión conocidos.
[f]La validación de una muestra aislada se consigue recogiendo toda la información junta, y dividiéndola después artificialmente en muestra de «derivación» y de «validación».
[g]Los buenos estándares de referencia son independientes de la prueba y se aplican con enmascaramiento o de modo objetivo a todos los pacientes. Los estándares de referencia de baja calidad se aplican de cualquier modo, pero también independientemente de la prueba. El uso de un estándar de referencia no independiente (donde la prueba o el test están incluidos en la «referencia», o el hecho de realizar la prueba afecta a la referencia) implica un nivel 4 del estudio.
[h]Los estudios de validación evalúan la calidad de una prueba diagnóstica específica, basándose en la evidencia anterior. Un estudio exploratorio recoge información y rastrea los datos (p. ej., utilizando un análisis de regresión) para determinar qué factores son «significativos».
[i]Por estudios de cohortes de pronóstico de baja calidad se entiende aquel cuya muestra está sesgada a favor de los pacientes que ya tienen el resultado que se desea medir, o que la medida del resultado se consiguió en menos del 80% de los pacientes del estudio, o aquel en el que no hay corrección de los factores de confusión.
Grados de recomendación:
A. Estudios de nivel 1 sistemáticamente.
B. Estudios de nivel 2 o 3 sistemáticamente, o extrapolaciones de estudios de nivel 1.
C. Estudios de nivel 4, o extrapolaciones de estudios de nivel 2 o 3.
D. Nivel 5 de evidencia, o estudios de cualquier nivel no coherentes o no concluyentes.
Se consideran «extrapolaciones» cuando los datos se utilizan en una situación que potencialmente tiene diferencias clínicas importantes con respecto a la situación del estudio original.

uno de ellos frente a placebo), extrapolación de los resultados de un estudio con un determinado fármaco al resto de fármacos de su misma familia (en ausencia de un efecto de clase demostrado), cuando existen grandes diferencias entre la población donde se aplicará y la incluida en los estudios evaluados.
- Imprecisión: cuando los estudios disponibles incluyen relativamente pocos episodios o pocos pacientes y, por tanto, presenta intervalos de confianza amplios.
- Sesgo de notificación: si se tiene la duda razonable de que los autores no han incluido todos los estudios (p. ej., en el contexto de una revisión sistemática) o todas las variables de resultado relevantes; es especialmente destacable en ensayos de pequeño tamaño y financiados por la industria farmacéutica.

Los aspectos que pueden aumentar la calidad de los estudios observacionales son los siguientes:
- Efecto importante: cuando el efecto observado muestra una asociación fuerte (riesgo relativo [RR] >2 o <0,5) o muy fuerte (RR >5 o <0,2) y consistente, basada en estudios sin factores de confusión, es improbable que éste se deba únicamente al diseño más débil del estudio. En estas ocasiones podemos considerar la calidad como moderada o incluso alta.
- La presencia de un gradiente dosis-respuesta.
- Situaciones en que todos los posibles factores de confusión podrían haber reducido el efecto observado. Por ejemplo, si los pacientes que reciben la intervención de interés presentan un peor pronóstico y, aun así, obtienen mejores resultados que el grupo control, es probable que el efecto real sea mayor.

3. La calidad de la evidencia será valorada como alta, moderada, baja y muy baja, categorías que representan el gradiente de confianza que tenemos en que la estimación del efecto observado es cierta.
4. Finalmente, gradúa la fuerza de las recomendaciones en dos únicas categorías (recomendaciones fuertes o débiles, y ambas pueden ir a favor o en contra de una determinada intervención). En el caso de una recomendación fuerte, el grupo elaborador confía en que los efectos beneficiosos superan a los perjudiciales. En el caso de una recomendación débil, concluye que los efectos beneficiosos de llevar a cabo la recomendación probablemente superan a los perjudiciales, aunque no está completamente seguro. Los factores que se deben tener en cuenta en la graduación de las recomendaciones son los siguientes: a) balance entre beneficios y riesgos; b) calidad de la evidencia (si ésta no es buena, a pesar de que la magnitud sea importante, disminuirá nuestra confianza y, por tanto, la fuerza con la que realizamos una recomendación); c) valores y preferencias de la población diana a la cual va dirigida la intervención (más aún si reflejan los de los médicos o los de los individuos o la sociedad en general), y d) costes (variables en el tiempo, según el área geográfica, así como sus implicaciones; las recomendaciones altamente determinadas por los costes pueden cambiar en el tiempo en la medida en que las implicaciones de los recursos varíen).

Así pues, entre tanta maraña de niveles de evidencia y fuerza de las recomendaciones, resulta oportuno adaptar y homogeneizar el sistema GRADE en la toma de decisiones clínicas, pues es sencillo y permite diferenciar las implicaciones de las recomendaciones en los pacientes, los clínicos y los gestores:
- Recomendación fuerte:
  – Pacientes. La inmensa mayoría de las personas estaría de acuerdo con la acción recomendada y únicamente una pequeña parte no lo estaría.
  – Clínicos. La mayoría de los pacientes debería recibir la intervención recomendada.
  – Gestores. La recomendación puede ser adoptada como política sanitaria en la mayoría de las situaciones.
- Recomendación débil:
  – Pacientes. La mayoría de las personas estaría de acuerdo con la acción recomendada, pero un número importante de ellas no.
  – Clínicos. Reconoce que diferentes opciones serán apropiadas para distitntos pacientes y que el profesional sanitario tiene que ayudar a cada uno a adoptar la decisión más consistente con sus valores y preferencias.
  – Gestores. Existe la necesidad de un debate importante y la participación de los grupos de interés.

Bibliografía

Alonso Coello P, Rotaeche del Campo R, Etxebarría Agirre A. El sistema GRADE para la evaluación de la calidad de la evidencia y la graduación de la fuerza de las recomendaciones. Fisterra [en línea] [consultado el 24-1-2010]. Disponible en: http://www.fisterra.com/guias2/fmc/grade.asp

Jiménez JF. Niveles de evidencia y grados de recomendación. Psicoevidencias [en línea] [citado el 26-1-2010] [consultado el 24-1-2010]. Disponible en: http://www.psicoevidencias.es/Evidencia/ASBE/niveles-de-evidencia-y-grados-de-recomendacion.html

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: 457-460.

Primo J. Niveles de evidencia y grados de recomendación (I/II). Enfermedad Inflamatoria Intestinal al Día. 2003; 2: 39-42.

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