Rev. Chilena de Cirugía. Vol 61 - Nº 6, Diciembre 2009; pág. 582-595

DOCUMENTOS

Cómo interpretar los "Niveles de Evidencia" en los diferentes escenarios clínicos*

Drs. CARLOS MANTEROLA D.[1,2,3], DANIELA ZAVANDO M. [4,5]. GRUPO MINCIR.

[1] Departamento de Cirugía, Universidad de La Frontera.
[2] Centro Colaborador UFRO de la Red Cochrane Iberoamericana.
[3] CIGES (Capacitación, Investigación y Gestión para la Salud Basada en Evidencia), Facultad de Medicina, Universidad de La Frontera.
[4] Programa de Magíster en Ciencias Médicas, Universidad de La Frontera.
[5] Programa de Pos-Graduaçao em Morfologia, Escola Paulista de Medicina, Universidade Federal de Sao Paulo, Brasil (UNIFESP).
Temuco, Chile.

Evidence-based clinical practice, levels of evidence

Introducción

El uso de la evidencia en ciencias se atribuye a la medicina tradicional china, en los tiempos del emperador Qianlong, cuando ya se señalaba el método "kaozheng" que representa la "búsqueda de evidencia práctica"[1], pero su desarrollo, como se conoce en la actualidad, se originó con la creación en 1976 de la Canadian Task Force on Preventive Health Care (CTFPHC), quienes fueron los primeros en generar y organizar los niveles de evidencia y los grados de recomendación para pacientes asintomáticos, indicando cuáles procedimientos eran los más adecuados y cuáles debían ser evitados[2]. Esta metodología permitió tomar conciencia de la existencia de un orden jerárquico en la calidad de la evidencia entre los estudios científicos, donde lógicamente aquellos que presentan más sesgos, no debieran justificar acciones clínicas[3]. También en Canadá, desde 1992, un grupo de médicos internistas y epidemiólogos clínicos de la Facultad de Medicina de la Universidad de McMaster, definieron, sistematizaron y consolidaron el concepto de medicina basada en la evidencia o "medicina basada en pruebas" (MBE), contribuyendo al cambio paradigmático para estudiar y ejercer las ciencias médicas, pues, entre otros tantos hechos, esta práctica se hacía cada vez más compleja, debido al incremento progresivo y abrumador de las publicaciones científicas que no se asociaban a calidad o a un aumento del tiempo necesario para leer y valorar de manera apropiada dicha información[4-6].

El análisis constante de la evidencia disponible desde la perspectiva de los diferentes escenarios clínicos, permite establecer grados de recomendación para el ejercicio de procedimientos diagnósticos, terapéuticos, preventivos y económicos en salud; e indica la forma de valorar el conocimiento en función de etiología, daño, morbilidad y complicaciones; pronóstico, historia natural y curso clínico de una enfermedad o evento de interés. Estos han de actualizarse permanentemente en relación al avance del conocimiento, del desarrollo tecnológico y al estado del arte.

Se estima que hasta la fecha se han descrito y propuesto alrededor de 100 sistemas distintos para valorar la evidencia[7], lo que nos orienta acerca del valor que le otorga la comunidad científica a esta

*Recibido el 17 de Agosto de 2009 y aceptado el 22 de septiembre de 2009.

Financiado por proyecto DI09-0060 de la Dirección de Investigación Universidad de La Frontera.

Correspondencia: Dr. Carlos Manterola D.
Manuel Montt 112, Oficina 408. Temuco, Chile.
E-mail: cmantero@ufro.cl

forma de hacer medicina; pero que, por otra parte contradice el principio de simpleza y practicidad que inspiró al paradigma de la MBE. Esta situación puede confundir al clínico en la búsqueda de la mejor evidencia aplicable a su realidad, debido a que, a lo anterior se agrega la dificultad propia del lenguaje epidemiológico en que se expresa la información, siendo en ocasiones la comprensión de estos tópicos, sólo privilegio de algunos.

El objetivo de este artículo es entregar información referente a los niveles de evidencia para cada tipo de estudio según el escenario clínico o ámbito de la práctica clínica que corresponda, interpretándolos en un lenguaje clínico comprensible, especialmente para quienes se acercan a la jerarquización de la evidencia y sus recomendaciones, desde la práctica clínica.

Clasificación de la evidencia y tipos de estudio

No todos los conocimientos provenientes de los artículos científicos publicados, tienen el mismo impacto o valor sobre la toma de decisiones en materia de salud; por ello, se hizo necesario evaluar la calidad de la evidencia. Esto es, en términos simples, el análisis de la validez de los hallazgos en virtud de la calidad metodológica de las investigaciones que los soportan, garantizándonos por una parte un acercamiento a la veracidad científica; y por otra, a que esta verdad pueda traducirse en recomendaciones que a partir de la valoración crítica de los estudios, nos permitan aplicarlas a la problemática clínica o evento de interés que nos ocupe. Resulta importante señalar, en este momento, que no todos los diseños tienen el mismo poder para formular una recomendación; y que más aún, un mismo diseño de investigación puede tener un nivel de evidencia y grado de recomendación diferente según el escenario clínico o ámbito de la práctica clínica que corresponda.

Por lo anteriormente expuesto, nos parece que vale la pena aclarar algunos conceptos que seguirán utilizándose de aquí en adelante:

Se define como diseño de investigación a los distintos tipos de estudios que con sus características metodológicas propias, permiten llevar a cabo una investigación clínica[8].

Como escenario, ámbito o entorno, al ambiente en el que se desarrolla la situación clínica que se está evaluando; es decir: tratamiento, prevención, etiología, daño, pronóstico e historia natural, diagnóstico diferencial, prevalencia, estudios económicos y análisis de decisión.

Como niveles de evidencia, a herramientas, instrumentos y escalas que clasifican, jerarquizan y valoran la evidencia disponible, de forma tal que en base a su utilización se pueda emitir juicios de recomendación.

Como grados de recomendación a una forma de clasificación de la sugerencia de adoptar o no la adquisición o puesta en marcha de tecnologías sanitarias según el rigor científico de cada tipo de diseño[4-6].

Dependiendo de los tipos de diseños de investigación clínica utilizados, podemos observar diferentes niveles o gradación de la calidad de estos según el escenario de investigación clínica a la que se refiera. Por una parte, llama la atención que las intervenciones terapéuticas sean las más abordadas desde esta concepción, situación que posiblemente no sea por azar, puesto que es un hecho que entre el 50% y el 60% de las publicaciones versan sobre el ámbito del tratamiento o los procedimientos terapéuticos[9-11]. No obstante ello, es deseable que todo el quehacer en salud sea analizado bajo la misma perspectiva. Lo anterior se relaciona al grado de avance que los diversos grupos de expertos (autores de clasificaciones de la evidencia), han ido adquiriendo y que han adecuado, paulatinamente, estos fundamentos a los diferentes escenarios y diseños. Entonces, debemos advertir al lector que sólo las clasificaciones propuestas por Sackett, National Institute for Health and Clinical Excellence (NICE) y el Oxford Centre for Evidence-Based Medicine (OCEBM), consideran otras áreas de la investigación clínica además de la concerniente a tratamiento. De esta forma, NICE adiciona evaluación de la evidencia para diagnóstico; mientras que Sackett considera cuatro grandes grupos temáticos: terapia, prevención, etiología y daño; pronóstico e historia natural; diagnóstico; y estudios económicos. OCEBM, que es una modificación "complementaria" de la clasificación de Sackett, incluye terapia, prevención, etiología y daño en un subgrupo; pronóstico e historia natural en otro; diagnóstico en el siguiente; diagnóstico diferencial y estudios de prevalencia en otro; y estudios económicos y análisis de decisión en otro.

A continuación se mencionan y describen las propuestas de jerarquización de la evidencia más utilizada en la actualidad (Tabla 1).

Canadian Task Force on Preventive Health Care (CTFPHC)[2,12]

Esta propuesta de clasificación de la evidencia busca generar recomendaciones de una manera práctica, adoptando una posición binaria, "*hágalo o no",* pero sólo en el ámbito de la prevención. Este grupo de estudio generó su primer informe en el año 1979, en el cual se divulgó el análisis de la

Tabla 1. Propuestas de jerarquización de la evidencia analizadas

Propuesta	Tratamiento	Prevención	Etiología	Daño	Pronóstico	Diagnóstico	Prevalencia	Económicos
CTFPHC		X						
SACKETT	X	X	X	X	X	X		X
USPSTF						X		
OCEBM	X	X	X	X	X	X	X	X
SIGN	X							
NICE	X					X		

Tabla 2. Grados de recomendación para intervenciones de prevención (CTFPHC)[12]

Grados de recomendación	Interpretación
A	Existe buena evidencia para recomendar la intervención clínica de prevención
B	Existe moderada evidencia para recomendar la intervención clínica de prevención
C	La evidencia disponible es conflictiva y no permite hacer recomendaciones a favor o en contra de la intervención clínica preventiva; sin embargo, otros factores podrían influenciar en la decisión
D	Existe moderada evidencia para recomendar en contra de la intervención clínica de prevención
E	Existe buena evidencia para recomendar en contra la intervención clínica de prevención
I	Existe evidencia insuficiente (en cantidad y en calidad) para hacer una recomendación; sin embargo, otros factores podrían influenciar en la decisión

evidencia hasta esa fecha para 78 enfermedades, lo que permitió ordenar y producir planes de salud para la población canadiense desde la perspectiva de "la prevención". La metodología de este grupo hace énfasis en el tipo de diseño utilizado y la calidad de los estudios publicados, basándose finalmente en tres elementos claves:

1. Un orden para los grados de recomendación, establecido por letras del abecedario donde, las letra A y B indican que existe evidencia para ejercer una acción (*se recomienda hacer*); D y E indican que no debe llevarse a cabo una maniobra o acción determinada (*se recomienda no hacer*); la letra C, indica que la evidencia es "conflictiva", o sea, que existe contradicción. Y la letra I que indica insuficiencia en calidad y cantidad de evidencia para establecer una recomendación (Tabla 2).
2. Niveles de evidencia clasificados según diseño de estudio de I a III, disminuyendo en calidad según se acrecienta numéricamente. Para el número II se subdivide en números arábigos del 1 al 3 (Tabla 3).
3. Niveles de evidencia clasificados según la validez interna o calidad metodológica del estudio, en buena, moderada e insuficiente (Tabla 4).

Como todo sistema de clasificación, este presenta algunas debilidades que se mencionan a continuación:

- No abarca toda la dimensionalidad de la problemática de la "prevención", respecto a las condiciones particulares de quienes son sujeto de la aplicación de medidas preventivas.
- En su análisis no se incorpora el ámbito financiero para la factibilidad de las intervenciones preventivas.
- La propuesta sólo se basó en población canadiense, por ende, el uso de las recomendaciones sólo puede ser extrapolada a poblaciones similares a ésta; por ello, esta propuesta tiene un inconveniente relacionado con su validez externa que ha de ser valorado al momento de pretender aplicarla en otros escenarios, pues de lo contrario se corre el riesgo de incurrir en esfuerzos económicos con resultados erráticos.

Tabla 3. Niveles de evidencia e interpretación de los tipos de estudio para intervenciones de prevención (CTFPHC)[12]

Niveles de evidencia	Interpretación
I	Evidencia a partir de EC con asignación aleatoria
II-1	Evidencia a partir de EC sin asignación aleatoria
II-2	Evidencia a partir de estudios de cohortes y casos y controles, preferiblemente realizados por más de un centro o grupo de investigación
II-3	Evidencia a partir de comparaciones en el tiempo o entre sitios, con o sin la intervención; podrían incluirse resultados espectaculares provenientes de estudios sin asignación aleatoria
III	Opinión de expertos, basados en la experiencia clínica; estudios descriptivos o informes de comités de expertos

Tabla 4. Validez interna e interpretación de los tipos de estudio para intervenciones de prevención (CTFPHC)[12]

Validez interna	Interpretación
Buena	Un estudio (incluyendo la RS y el meta-análisis) que cumple los criterios específicos de estudio bien diseñado
Moderada	Un estudio (incluyendo la RS y el meta-análisis) que no cumple (o no está claro que cumpla) al menos uno de los criterios específicos de estudio bien diseñado*, aunque no tiene "defectos fatales"
Insuficiente	Un estudio (incluyendo la RS y el meta-análisis) que tiene en su diseño al menos un "defecto fatal" o no cumple (o no está claro que cumpla) al menos uno de los criterios específicos de estudio bien diseñado*, aunque no presenta "errores fatales" o una acumulación de defectos menores que hagan que los resultados del estudio no permitan elaborar las recomendaciones

- No contempla la relación del paciente, sus expectativas y su medio, para establecer las recomendaciones.
- Puede inducir a errores al momento de valorar las recomendaciones para su puesta en práctica en los sistemas estatales de salud, esto quiere decir que una recomendación B puede ser menospreciada pudiendo tener un beneficio importante para la población.
- No contempla otro tipo de áreas de investigación como tratamiento, etiología, daño, pronóstico, etc.

Clasificación de la Evidencia según Sackett[13]

Esta sistematización propuesta por el epidemiólogo David L. Sackett (la que se emplea generalmente), jerarquiza la evidencia en niveles que van del 1 a 5; siendo el nivel 1 la "mejor evidencia" y el nivel 5 la "peor, la más mala o la menos buena", según como se quiera leer (Tabla 5).

Ésta fue la primera propuesta que consideró otros escenarios clínicos o ámbitos de la práctica clínica diferentes de la prevención. Incorporó los análisis económicos, el diagnóstico y el pronóstico.

Hasta hoy, ha sido ampliamente utilizada por diferentes grupos científicos. A cada ámbito o escenario clínico le otorga el diseño de estudio más apropiado para la elaboración de las recomendaciones. Así, en el escenario de terapia, los diseños más puntuados corresponden a las revisiones sistemáticas (RS) de ensayos clínicos controlados con asignación aleatoria (EC); en escenarios de pronóstico, los estudios de cohortes; en escenarios de diagnóstico, los estudios de pruebas diagnósticas con estándar de referencia, etc.

Presenta desventajas que son comunes a todas las propuestas existentes, como que puede llevar a despreciar recomendaciones que se basan en evidencias inferiores a las de nivel 1; por ejemplo un

estudio de nivel de evidencia 4, aún cuando se trate de un artículo cuyo aporte sea novedoso. No obstante, los autores son enfáticos en que lo que debe valorarse es "la mejor evidencia disponible actual"; pues "lo actual puede variar en el día a día"; y de este modo puede ocurrir que ante determinadas situaciones "la mejor evidencia disponible actual" sea una serie de casos y no un EC; y dos meses después, aparezca un estudio de cohorte prospectivo que dará "la mejor evidencia disponible actual".

Esta clasificación fue pionera y ha servido de base para el desarrollo de clasificaciones más completas, como la propuesta por el OCEBM[14].

Tabla 5. Clasificación de los niveles de evidencia según Sackett[13]

Recomendación	Nivel	Terapia, prevención, etiología y daño	Pronóstico	Diagnóstico	Estudios económicos
A	1a	RS con homogeneidad y Meta-análisis de EC	RS con homogeneidad y Meta-análisis de estudios de cohortes concurrente	RS de estudios de diagnóstico nivel 1	RS de estudios económicos de nivel 1
	1b	EC individuales con intervalo de confianza estrecho	Estudio individual de cohorte concurrente con seguimiento superior al 80% de la cohorte	Comparación independiente y enmascarada de un espectro de pacientes consecutivos sometidos a la prueba diagnóstica y al estándar de referencia	Análisis que compara los desenlaces posibles, contra una medida de costos. Incluye un análisis de sensibilidad
B	2a	RS con homogeneidad de estudios de cohortes	RS de cohortes históricas	RS de estudios diagnósticos de nivel mayor a 1	RS de estudios económicos de nivel mayor a 1
	2b	Estudio de cohortes individual. EC de baja calidad	Estudio individual de cohortes históricas	Comparación independiente enmascarada de pacientes no consecutivos, sometidos a la prueba diagnóstica y al estándar de referencia	Comparación de un número limitado de desenlaces contra una medida de costo. Incluye análisis de sensibilidad
	3a	RS con homogeneidad de estudios de casos y controles			
	3b	Estudio de casos y controles individuales		Estudios no consecutivos o carentes de un estándar de referencia	Análisis sin una medida exacta de costo, pero incluye análisis de sensibilidad
C	4	Series de casos. Estudio de cohortes y casos y controles de mala calidad	Series de casos. Estudios de cohortes de mala calidad	Estudios de casos y controles sin la aplicación de un estándar de referencia	Estudio sin análisis de sensibilidad
D	5	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, o en investigación teórica	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, o en investigación teórica	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, o en investigación teórica	Opinión de expertos sin evaluación crítica explícita, o basada en teoría económica

Por homogeneidad se entiende una RS que está libre de variaciones (heterogeneidad) en las direcciones o grados de resultados entre los estudios individuales.

U.S. Preventive Services Task Force (USPSTF)[15]

Este grupo de expertos jerarquizó y estableció la fuerza de sus recomendaciones a partir de la calidad de la evidencia y del beneficio neto; es decir, beneficios menos perjuicios de la medida evaluada para su aplicación en "exámenes periódicos de salud". Por otro lado, analizó el coste-efectividad de las intervenciones, por ende su aporte vino a complementar lo que había generado el grupo de CTFPHC[2,12]. La fuerza de las recomendaciones va desde la letra A hasta la E, otorgándose una A cuando existe buena evidencia que respalda la recomendación; y una E, que indica que existe buena evidencia que sustenta la recomendación de evitar la intervención (Tabla 6). La calidad de la evidencia es valorada en buena, justa o insuficiente en la fuerza de las recomendaciones y se basa en la consideración sistemática de tres criterios: el concepto de "costo de sufrimiento de la condición estudiada", las "características de la intervención" y la "efectividad de la intervención" demostrada en investigaciones clínicas publicadas; recibiendo un especial énfasis la efectividad de la intervención. En la revisión de los estudios clínicos, se utilizan criterios estrictos para la selección de la evidencia admisible y se otorga un acento especial en la calidad de los diseños de los estudios. Para jerarquizar la calidad de la evidencia, se otorga mayor peso a aquellos diseños que metodológicamente ofrecen menor riesgo de sesgos y errores aleatorios (Tabla 7). Es interesante observar que la correlación entre la fuerza de la recomendación y el nivel de evidencia no es exacta; por ejemplo, puede haber un buen nivel de evidencia que no prueba que una intervención es efectiva, como el caso de la mamografía en mujeres menores de 50 años; que recibe una recomendación "C". Por otra parte una recomendación tipo "A" fue otorgada al Papanicolaou para detección precoz de cáncer cérvico-uterino basado en el "costo de sufrimiento de la enfermedad" y un nivel II de evidencia sosteniendo dicha intervención[14]. Así, las recomendaciones permiten un cierto grado de flexibilidad y se amoldan de acuerdo al contexto imperante.

Esta clasificación tiene la ventaja que contempla nuevos aspectos que condicionan la elaboración de recomendaciones ya no sólo a la calidad de la evidencia (siempre controversial) sino también al beneficio neto y al "costo de sufrimiento de padecer el evento de interés en estudio", aproximándose a una perspectiva más global que emerge desde la posición de quien padece la condición de salud o evento de interés en estudio. Dentro de las desventajas podemos señalar que sólo se consideran estudios para diagnóstico y, aún resultan insuficientes los aspectos abordados para establecer las recomendaciones; por ejemplo no se considera la factibilidad de cobertura de salud estatal.

Tabla 6. Recomendaciones a partir de calidad de evidencia para exámenes periódicos de salud (USPSTF)[15]

Recomendación	Interpretación
A	La USPSTF recomienda claramente que los clínicos proporcionen la intervención a los pacientes que cumplan los criterios. La USPSTF ha encontrado buena evidencia de que la medida mejora de manera importante los resultados en salud y concluye que los beneficios superan ampliamente a los riesgos
B	La USPSTF recomienda que los clínicos proporcionen la intervención a los pacientes. La USPSTF ha encontrado evidencia moderada de que la medida mejora de manera importante los resultados en salud y concluye que los beneficios superan a los riesgos
C	La USPSTF no recomienda a favor o en contra de la intervención. La USPSTF ha encontrado al menos evidencia moderada de que la medida puede mejorar los resultados en salud, pero los beneficios son muy similares a los riesgos y no puede justificarse una recomendación general
D	La USPSTF recomienda en contra que los clínicos proporcionen la intervención a los pacientes asintomáticos. La USPSTF ha encontrado al menos evidencia moderada de que la medida es ineficaz o que los riesgos superan a los beneficios
I	La USPSTF concluye que la evidencia es insuficiente para recomendar a favor o en contra de la intervención. No existe evidencia de que la intervención es ineficaz, o de calidad insuficiente, o conflictiva y que el balance entre los riesgos y los beneficios no se puede determinar

Tabla 7. Niveles de evidencia e interpretación de estos para exámenes periódicos de salud (USPSTF)[15]

Evidencia	Interpretación
Buena	La evidencia incluye resultados consistentes a partir de estudios bien diseñados y realizados en poblaciones representativas que directamente evalúan efectos sobre resultados de salud
Moderada	La evidencia es suficiente para determinar efectos sobre resultados de salud, pero la fuerza de la evidencia es limitada por el número, la calidad, o la consistencia de los estudios individuales, la generalización a la práctica rutinaria, o la naturaleza indirecta de la evidencia sobre los resultados de salud
Insuficiente	La evidencia es insuficiente para evaluar los efectos sobre los resultados de salud debido al número limitado o al poder de estudios, defectos importantes en su diseño o realización, inconsistencias en la secuencia de la evidencia, o falta de información sobre resultados de salud importantes

Centre for Evidence-Based Medicine, Oxford (OCEBM)[14]

Esta propuesta se caracteriza por valorar la evidencia según el área temática o escenario clínico y el tipo de estudio que involucra al problema clínico en cuestión. Lo anterior es una innovación y es complementaria a lo expuesto por las otras iniciativas. Esta, tiene la ventaja que gradúa la evidencia de acuerdo al mejor diseño para cada escenario clínico, otorgándole intencionalidad, agregando las RS en los distintos ámbitos. Así por ejemplo, al tratarse de un escenario clínico relacionado con pronóstico de un evento de interés, la evidencia será valorada a partir de una RS de estudios de cohortes con homogeneidad, o en su defecto de estudios de cohortes individuales con un seguimiento superior al 80% de la cohorte; en cambio, si el escenario se refiere a terapia o tratamiento, la evidencia se valorará principalmente a partir de RS de EC, o en su defecto de EC individuales con intervalos de confianza estrechos.

Esta clasificación tiene la ventaja que nos asegura el conocimiento más atingente a cada escenario, por su alto grado de especialización. Además tiene la prerrogativa de aclarar cómo afecta la falta de rigurosidad metodológica al diseño de los estudios, disminuyendo su valoración no sólo en la gradación de la evidencia, sino que también en la fuerza de las recomendaciones.

No obstante lo cual, presenta algunos inconvenientes para su práctica habitual. Por una parte, vemos como en su estructura se presentan términos epidemiológicos poco amigables y con múltiples aclaraciones que hacen su lectura poco fluida y, que rápidamente pueden frustrar a quien se aproxima a ella por primera vez. En su intento por abarcar todos los aspectos con la máxima exhaustividad, pierde la simpleza para hacerla aplicable (Tabla 8).

Scottish Intercollegiate Guidelines Network (SIGN)[16,17]

Esta propuesta, se origina también teniendo como foco de interés la temática del tratamiento. Se diferencia de las anteriores por su particular énfasis en el análisis cuantitativo que involucra a las RS y otorga importancia a la reducción del error sistemático. Se compone de niveles de evidencia y grados de recomendación según esos niveles (Tablas 9 y 10). Como fortaleza, es interesante el hecho que considera la calidad metodológica de los estudios que componen las RS, situación que es de interés dada la alta producción anual de revisiones. Como debilidad podemos señalar que no considera en la elaboración de las recomendaciones la realidad científica y tecnológica del momento, pues estas se crean con una rigidez que puede ser peligrosa en el sentido de quienes usan con ortodoxia las recomendaciones para la implementación de políticas de salud, los que pueden evitar invertir con argumento en la evidencia. Por otro lado, se basa de forma puntual en los aspectos metodológicos y de diseño, pero no así en la dimensión de la perspectiva del padecer una enfermedad o considerar las implicancias económicas de las medidas recomendadas; situación que pone en riesgo la factibilidad de su utilización en la práctica médica latinoamericana.

National Institute for Health and Clinical Excellence (NICE)[18]

Esta iniciativa que nace del National Health Service del Reino Unido (NHS), abarca la temática de la terapia y el diagnóstico. Adapta la clasificación hecha por SIGN para terapia y utiliza la de la OCEBM para diagnóstico; de tal modo que se efectúa una valoración de la evidencia disponible con base en estas dos herramientas. Por lo tanto, queda patente la semejanza con las clasificaciones an-

Tabla 8. Clasificación de los niveles de evidencia de Oxford (OCEBM)[14]

Grado de recomendación	Nivel de evidencia	Tratamiento, prevención, etiología y daño	Pronóstico e historia natural	Diagnóstico	Diagnóstico diferencial y estudios de prevalencia	Estudios económicos y análisis de decisión
A	**1a**	RS con homogeneidad de EC controlados con asignación aleatoria	RS de estudios de cohortes, con homogeneidad, o sea que incluya estudios con resultados comparables, en la misma dirección y validadas en diferentes poblaciones	RS de estudios diagnósticos de nivel 1 (alta calidad), con homogeneidad, o sea que incluya estudios con resultados comparables y en la misma dirección y en diferentes centros clínicos	RS con homogeneidad de estudios de cohortes prospectivas	RS con homogeneidad de estudios económicos de nivel 1
	1b	EC individual con intervalo de confianza estrecho	Estudios de cohortes individuales con un seguimiento mayor de 80% de la cohorte y validadas en una sola población	Estudios de cohortes que validen la calidad de una prueba específica, con estándar de referencia adecuado (independientes de la prueba) o a partir de algoritmos de estimación del pronóstico o de categorización del diagnóstico o probado en un centro clínico	Estudio de cohortes prospectiva con buen seguimiento	Análisis basado en costes o alternativas clínicamente sensibles; RS de la evidencia; e incluyendo análisis de la sensibilidad
	1c	Eficiencia demostrada por la práctica clínica. Considera cuando algunos pacientes mueren antes de ser evaluados	Resultados a partir de la efectividad y no de su eficacia demostrada a través de un estudio de cohortes. 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	Series de casos todos o ninguno	Análisis absoluto en términos del mayor valor o peor valor
B	**2ª**	RS de estudios de cohortes, con homogeneidad	RS de estudios de cohorte retrospectiva o de grupos controles no tratados en un EC, con homogeneidad	RS de estudios diagnósticos de nivel 2 (mediana calidad) con homogeneidad	RS (con homogeneidad) de estudios 2b y mejores	RS (con homogeneidad) de estudios económicos con nivel mayor a 2
	2b	Estudio de cohortes individual con seguimiento inferior a 80% (incluye EC de baja calidad)	Estudio de cohorte retrospectiva o seguimiento de controles no tratados en un EC, o GPC no validadas	Estudios exploratorios que, a través de una regresión logística, determinan factores significativos, y validados con estándar de referencia adecuado (independientes de la prueba)	Estudios de cohortes retrospectivas o de seguimiento insuficiente	Análisis basados en costes o alternativas clínicamente sensibles; limitado a revisión de la evidencia; e incluyendo un análisis de sensibilidad

Tabla 8. Clasificación de los niveles de evidencia de Oxford (OCEBM)[14] (Continuación)

Grado de recomendación	Nivel de evidencia	Tratamiento, prevención, etiología y daño	Pronóstico e historia natural	Diagnóstico	Diagnóstico diferencial y estudios de prevalencia	Estudios económicos y análisis de decisión
	2c	Estudios ecológicos o de resultados en salud	Investigación de resultados en salud		Estudios ecológicos	Auditorías o estudios de resultados en salud
	3a	RS de estudios de casos y controles, con homogeneidad		RS con homogeneidad de estudios 3b y de mejor calidad	RS con homogeneidad de estudios 3b y mejores	RS con homogeneidad de estudios 3b y mejores
	3b	Estudios de casos y controles individuales		Comparación enmascarada y objetiva de un espectro de una cohorte de pacientes que podría normalmente ser examinado para un determinado trastorno, pero el estándar de referencia no se aplica a todos los pacientes del estudio. Estudios no consecutivos o sin la aplicación de un estándar de referencia		Estudio no consecutivo de cohorte, o análisis muy limitado de la población basado en pocas alternati vas o costes, estimaciones de datos de mala calidad, pero incluyendo análisis de la sensibilidad que incorporan variaciones clínicamente sensibles
C	**4**	Serie de casos, estudios de cohortes, y de casos y controles de baja calidad	Serie de casos y estudios de cohortes de pronóstico de poca calidad	Estudio de casos y controles, con escasos o sin estándares de referencia independiente	Series de casos o estándares de referencia obsoletos	Análisis sin análisis de sensibilidad
D	**5**	Opinión de expertos sin evaluación crítica explícita, ni basada en fisiología, ni en trabajo de investigación juicioso ni en "principios fundamentales"	Opinión de expertos sin evaluación crítica explícita, ni basada en fisiología, ni en trabajo de investigación juicioso ni en "principios fundamentales"	Opinión de expertos sin evaluación crítica explícita, ni basada en fisiología, ni en trabajo de investigación juicioso ni en "principios fundamentales"	Opinión de expertos sin evaluación crítica explícita, ni basada en fisiología, ni en trabajo de investigación juicioso ni en "principios fundamentales"	Opinión de expertos sin evaluación crítica o basado en teoría económica o en "principios fundamentales"

GPC: Guía de práctica clínica. Estudios con homogeneidad: Se refiere a que incluya estudios con resultados comparables y en la misma dirección.

Tabla 9. Niveles de evidencia para estudio de tratamiento con análisis cuantitativo (SIGN)[17]

Nivel de evidencia	Interpretación
1++	Meta-análisis de alta calidad, RS de EC o EC de alta calidad con muy poco riesgo de sesgo
1+	Meta-análisis bien realizados, RS de EC o EC bien realizados con poco riesgo de sesgos
1–	Meta-análisis, RS de EC o EC con alto riesgo de sesgos
2++	RS de alta calidad de estudios de cohortes o de casos y controles. Estudios de cohortes o de casos y controles con riesgo muy bajo de sesgo y con alta probabilidad de establecer una relación causal
2+	Estudios de cohortes o de casos y controles bien realizados con bajo riesgo de sesgo y con una moderada probabilidad de establecer una relación causal
2–	Estudios de cohortes o de casos y controles con alto riesgo de sesgo y riesgo significativo de que la relación no sea causal
3	Estudios no analíticos, como informes de casos y series de casos
4	Opinión de expertos

Los estudios clasificados como 1- y 2- no deben usarse en el proceso de elaboración de recomendaciones por su alto potencial de sesgo.

Tabla 10. Grados de recomendación para estudios de tratamiento con análisis cuantitativo (SIGN)[17]

Grados de recomendación	Interpretación
A	Al menos un meta-análisis, RS o EC clasificado como 1++ y directamente aplicable a la población diana de la guía; o un volumen de evidencia científica compuesto por estudios clasificados como 1+ y con gran consistencia entre ellos
B	Un volumen de evidencia científica compuesta por estudios clasificados como 2 ++, directamente aplicable a la población diana de la guía y que demuestran gran consistencia entre ellos; o evidencia científica extrapolada desde estudios clasificados como 1 ++ ó 1+
C	Un volumen de evidencia científica compuesta por estudios clasificados como 2 + directamente aplicables a la población diana de la guía y que demuestran gran consistencia entre ellos; o evidencia científica extrapolada desde estudios clasificados como 2 ++
D	Evidencia científica de nivel 3 ó 4; o evidencia científica extrapolada desde estudios clasificados como 2+

teriores respecto a la importancia otorgada a las RS al momento de generar las recomendaciones (Tablas 11-14). Presenta el inconveniente que hace una relación tan directa entre calidad de la evidencia y grado de recomendación que puede generar confusión sobre esos constructos.

Ejemplo

Una forma de ejemplificar el *modus operandi* de las distintas propuestas analizadas, es a través de la aplicación de un ejemplo. Para ello se utilizará una publicación de interés nacional, pues a partir de ella se generó el protocolo que ha sido incorporado en algunos hospitales chilenos como tratamiento del cáncer gástrico. Se trata del estudio publicado por JS Macdonald y cols[19], y que dice relación con los resultados obtenidos con el uso de cirugía y quimioradioterapia comparado con cirugía exclusiva en pacientes con adenocarcinoma gástrico y de la unión esofagogástrica; por ende, se trata de un escenario de tratamiento, razón por la que se aplicarán las propuestas concernientes a dicho es-

cenario. Para ello, se construyó una tabla que permite comparar las clasificación del estudio según la propuesta con que se analice (Tabla 15).

Es así como se puede comentar que el artículo no se orienta a una pregunta claramente definida pues esta sólo se puede suponer (de hecho puede tratarse de un escenario de tratamiento, prevención o pronóstico). La población blanco es heterogénea, incluye pacientes con cáncer gástrico y gastroesofágico. La intervención está claramente descrita cirugía+quimioradioterapia (CQR), no así la variable de interés (no se explicita si es morbilidad, mortalidad, supervivencia global, supervivencia libre de enfermedad, efectos tóxicos, etc). No se describe

Tabla 11. Niveles de evidencia para estudios de terapia (NICE)[18]

Nivel de evidencia	Interpretación
1++	Meta-análisis de gran calidad, RS de EC con asignación aleatoria o EC con asignación aleatoria con muy bajo riesgo de sesgos
1+	Meta-análisis de gran calidad, RS de EC con asignación aleatoria o EC con asignación aleatoria con bajo riesgo de sesgos
1-	Meta-análisis de gran calidad, RS de EC con asignación aleatoria o EC con asignación aleatoria con alto riesgo de sesgos*
2++	RS de alta calidad de estudios de cohortes o de casos-controles, o estudios de cohortes o de casos-controles de alta calidad, con muy bajo riesgo de confusión, sesgos o azar y una alta probabilidad de que la relación sea causal
2+	Estudios de cohortes o de casos-controles bien realizados, con bajo riesgo de confusión, sesgos o azar y una moderada probabilidad de que la relación sea causal
2–	Estudios de cohortes o de casos y controles con alto riesgo de sesgo*
3	Estudios no analíticos, como informe de casos y series de casos
4	Opinión de expertos

* Los estudios con un nivel de evidencia '–' no deberían utilizarse como base para elaborar una recomendación. Adaptado de Scottish Intercollegiate Guidelines Network.

Tabla 12. Grados de recomendación para estudios de terapia (NICE)[18]

Grados de recomendación	Interpretación
A	Al menos un meta-análisis, o un EC con asignación aleatoria categorizados como 1++, que sea directamente aplicable a la población diana; o una RS o un EC con asignación aleatoria o un volumen de evidencia con estudios categorizados como 1+, que sea directamente aplicable a la población diana y demuestre consistencia de los resultados. Evidencia a partir de la apreciación de NICE
B	Un volumen de evidencia que incluya estudios calificados de 2++, que sean directamente aplicables a la población objeto y que demuestren globalmente consistencia de los resultados, o extrapolación de estudios calificados como 1++ o 1+
C	Un volumen de evidencia que incluya estudios calificados de 2+, que sean directamente aplicables a la población objeto y que demuestren globalmente consistencia de los resultados, o extrapolación de estudios calificados como 2++
D	Evidencia nivel 3 o 4, o extrapolación de estudios calificados como 2+, o consenso formal

D (BPP): Un buen punto de práctica (BPP) es una recomendación para la mejor práctica basado en la experiencia del grupo que elabora la guía. IP: Recomendación a partir del manual para procedimientos de intervención de NICE.

Tabla 13. Niveles de evidencia para estudios diagnóstico (NICE)[18]

Nivel de evidencia	Interpretación
Ia	RS con homogeneidad* de estudios de nivel 1†
Ib	Estudios de nivel 1†
II	Estudios de nivel 2 ‡ RS de estudios de nivel 2
III	Estudios de nivel 3 § RS de estudios de nivel 3
IV	Consenso, informes de comités de expertos u opiniones y/o experiencia clínica sin valoración crítica explícita; o en base a la psicología, difusión de la investigación o "principios básicos"

*** Homogeneidad** significa que no hay variaciones o estas son pequeñas en la dirección y grado de los resultados entre los estudios individuales que incluye la RS. † **Estudios de nivel 1** son aquellos que utilizan una comparación enmascarada de la prueba con un estándar de referencia validado, en una muestra de pacientes que refleja a la población a quien se aplicaría la prueba. ‡ **Estudios nivel 2** son aquellos que presentan una sola de esta características: población reducida (la muestra no refleja las características de la población a la que se le va a aplicar la prueba; utilizan un estándar de referencia pobre (definido como aquel donde la 'prueba' es incluida en la 'referencia', o aquel en que las 'pruebas' afectan a la 'referencia'; la comparación entre la prueba y la referencia no está enmascarada; o estudios de casos y controles. **§ Estudios de nivel 3** son aquellos que presentan al menos dos o tres de las características señaladas anteriormente.

Tabla 14. Grados de recomendación para estudios diagnóstico (NICE)[18]

Grados de recomendación	Interpretación
A (EPD)	Estudios con un nivel de evidencia Ia o Ib
B (EPD)	Estudios con un nivel de evidencia II
C (EPD)	Estudios con un nivel de evidencia III
D (EPD)	Estudios con un nivel de evidencia IV

EPD = Estudios de pruebas diagnósticas

Tabla 15. Niveles de evidencia y grado de recomendación del estudio de Macdonald[19], según clasificación

Clasificación	Nivel de evidencia	Grado de recomendación
Tabla 2. CTFPHC[12]	—	C
Tablas 3 y 4. CTFPHC[12]	II-1	Insuficiente (validez)
Tabla 5. Sackett[13]	2b	B
Tabla 6 y 7. USPSTF[15]	Moderada	C
Tabla 8. OCEBM[14]	2b	B
Tabla 9 y 10. SIGN[17]	1 -	No recomendable
Tabla 11 y 12. NICE[18]	1 -	No recomendable

qué método de asignación aleatoria fue utilizado, sólo se especifica que 281 pacientes fueron asignados a CQR y 275 a cirugía exclusiva (CE); lo que ocurrió en el postoperatorio; tampoco se menciona si es que se mantuvo oculta la secuencia de asignación. Por otra parte, los pacientes fueron analizados en los grupos a los que fueron asignados, pero el seguimiento sólo se completó en el grupo de CE (35,6% de los pacientes del grupo de CQR no completó tratamiento y se desconoce qué ocurrió con ellos).

Por otra parte, es evidente que los resultados no pueden ser aplicados en el cuidado de mis pacientes, debido a que existen diferencias biodemográficas evidentes entre la población estudiada y la chilena (la población de raza negra, asiática y anglosajona en nuestra realidad es escasa; y el grupo denominado "otros" por Macdonald y que representa la población latina es sólo el 4% de la muestra); más del 50% de las lesiones son de localización antral, y en nuestro medio un porcentaje mayoritario son de fondo gástrico; al mismo

tiempo, en nuestra realidad existe una baja prevalencia de lesiones T1 y T2 (mayoritaria en el artículo en evaluación).

Tampoco fueron considerados todos los resultados de importancia clínica en el estudio, pues variables como calidad de vida, efectos adversos de la terapia en evaluación, mortalidad por efectos deletéreos, morbilidad y mortalidad quirúrgica y costos involucrados no fueron consideradas (al menos no aparecen consignadas en el reporte del estudio). No se reporta información sobre complicaciones quirúrgicas; por otra parte, un 36% de pacientes no completaron el esquema asignado, por lo que las estimaciones finales se realizan en base al 64% de los pacientes del grupo de CQR; lo que dificulta la valoración de potenciales beneficios o efectos deletéreos del esquema CQR al compararlo con CE.

En resumen, se trata de un EC multicéntrico, sin asignación aleatoria precisa, ni enmascaramiento; cuya validez interna se encuentra afectada por lo que su nivel de evidencia es 2b. Por otra parte, en relación a la validez externa; si se omitiesen los problemas relacionados con la validez interna, se constata que la inferencia de los resultados sólo aplicaría a poblaciones de características similares a las del estudio. En este caso, no a la población chilena o a otras similares[20].

Discusión

La evidencia ha tenido un desarrollo diacrónico marcado por el interés de la comunidad científica por ordenar y valorar de una manera exhaustiva el conocimiento, siendo uno de los objetivos primordiales el contestar preguntas clínicas surgidas del quehacer médico diario. En principio, la idea principal fue dar directrices para que el conocimiento más válido sea conseguido de manera expedita por el clínico. Paradójicamente, el desarrollo de múltiples propuestas de clasificar la evidencia y formular recomendaciones y el lenguaje epidemiológico empleado, ha llevado consigo confusión e incomprensión por parte de los clínicos, quienes ven en esta gran variedad de opciones más conflictos que ayuda al desarrollo de la práctica profesional. En este artículo hemos presentado y desarrollado de forma somera algunas de las clasificaciones existentes, utilizando como criterio de selección, el nivel de utilización de cada cual, el que se asocia directamente con el grado de aceptación de ellas por parte de la comunidad científica.

El problema de la jerarquización de la evidencia también ha planteado desafíos a los comités editoriales de las revistas biomédicas respecto de cómo estandarizar la valoración de la evidencia, para la aceptación de trabajos científicos. Entendemos la necesidad de simplificar y aunar criterios, incluso sería deseable que las recomendaciones que nacen de la valoración de la evidencia tuvieran un carácter universal, pero el generalizar también trae consigo algunos inconvenientes. ¿Por qué hay políticas de salud que fracasan aún cuando han contemplado en su elaboración la evidencia disponible? Una de las tantas razones es que algunas recomendaciones basadas en la evidencia pueden no ser aplicables a nuestro contexto latinoamericano. De esta forma, al momento de acercarnos a la evidencia es necesario considerar la existencia de características propias de nuestra región, tanto poblacionales, culturales, económicas, tecnológicas como ambientales; es decir, darle relevancia al concepto de validez externa y no "importar" todo lo que parece que resulta en otras latitudes sin probar previamente en nuestra realidad; en otras palabras dejar de pensar que "todo lo que brilla es oro". No obstante lo anterior, sorprendentemente vemos que se hace poco esfuerzo por establecer sistemas de valoración de la evidencia aplicada a nuestra propia población y contexto. Por otro lado, la enseñanza de las ciencias médicas no incentiva la validación de instrumentos extranjeros para la valoración de la lectura crítica, ni menos la creación de los propios. Así, nos vemos obligados no sólo a traducir el lenguaje epidemiológico desde el inglés, sino que también a adaptar e implementar esas traducciones a las problemáticas sanitarias regionales.

Es cierto que el análisis de la evidencia incluye sutilmente, la adaptación a diferentes realidades; pero, complicando un poco más la contingencia ¿qué repercusiones podemos tener en el ámbito legal? De a poco vamos observando cómo la *lex artis* ya no se basa en opiniones de expertos, cada día se hace más próxima a la MBE, lo cual implica la contemplación de recomendaciones al momento de fallos judiciales, pero ¿cómo nos aseguraremos que el acercamiento sea justo y se considere la realidad local? Necesitamos establecer recomendaciones basadas en estudios propios. No podemos pretender que las recomendaciones establecidas por científicos con distinto grado de desarrollo tecnológico y de habilidades, puedan ser entendidas por quienes ejecutan las leyes sin el riesgo de creerlas como definitorias.

Siguiendo con nuestro análisis nos preguntamos: ¿Cuál es el nivel de influencia de la evidencia en el quehacer médico en Chile? ¿Entendemos bien lo que es la evidencia, sus niveles e interpretaciones? ¿Cuál de las propuestas de clasificación de la evidencia es más adecuada para nosotros? Desde nuestro punto de vista, destacamos la clasificación

ofrecida por OCEBM por sobre las otras, por ser capaz de asignar una valoración más completa de la evidencia según cada tipo de escenario. Sin embargo, es absolutamente imprescindible entender que las recomendaciones, son consejos desde la más alta perspectiva científica, no importa cuál propuesta usemos sino cómo las empleemos, qué juicios hagamos y cómo interrelacionamos la evidencia con los factores propios de nuestro entorno.

El propósito de los autores es a través de este artículo, acercar las clasificaciones de la evidencia y sus recomendaciones a quienes se aproximan a la jerarquización de la misma e invitar a reflexionar respecto a este tema para abordarlo de manera equilibrada.

Referencias

1. Sackett DL, Richardson S, Rosenberg W, Haynes RB. Medicina baseada em evidências: práctica e ensino. 2ª ed. 2003, Porto Alegre: Artmed Editora S.A. 270.
2. Canadian Task Force on Preventive Health Care. New grades for recommendations from the Canadian Task Force on Preventive Health Care. CMAJ 2003; 169: 207-220.
3. Upshur RE. Are all evidence-based practices alike? Problems in the ranking of evidence. CMAJ 2003; 169: 672-673.
4. Evidence-based medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992; 268: 2420-2425.
5. Manterola C. Medicina basada en la evidencia o medicina basada en pruebas. Generalidades acerca de su aplicación en la práctica clínica cotidiana. Rev Med Clin Condes 2009; 20: 125-130.
6. Manterola C. Medicina basada en la evidencia. Conceptos generales y razones para aplicación en cirugía. Rev Chil Cir 2002; 54: 550-554.
7. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to Rate the Strength of Scientific Evidence. Health Services/Technology Assessment Text, National Library of Medicine. AHRQ Publication No. 02-E016, 2002. Available from: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.70996. Visitado el 22 de junio de 2009.
8. Manterola C. Investigación clínica, por qué realizarla y cómo desarrollarla. Rev Med Clin Condes 2009; 20: 233-239.
9. Pineda V, Manterola C, Vial M, Losada H. ¿Cuál es la calidad metodológica de los artículos referentes a terapia publicados en la Revista Chilena de Cirugía? Rev Chil Cir 2005; 57: 500-507.
10. Manterola C, Pineda V, Vial M, Losada H; the MINCIR Group. What is the methodologic quality of human therapy studies in ISI surgical publications? Ann Surg 2006; 244: 827-832.
11. Manterola C, Busquets J, Pascual M, Grande L. What is the methodological quality of articles on therapeutic procedures published in Cirugía Española? Cir Esp 2006; 79: 95-100.
12. Task Force Revitalization Process (CTFPHC). Evidence-Based Clinical Prevention, Updated August 17, 2005. Available from: http://www.ctfphc.org. Visitado el 22 de junio de 2009.
13. Sackett DL, Wennberg JE. Choosing the best research design for each question. BMJ 1997; 315(7123): 1636.
14. Oxford Centre for Evidence-based Medicine (CEBM). Centre for Evidence Based Medicine - Levels of Evidence (March 2009). Available from: http://www.cebm.net/index.aspx?o=1025. Visitado el 22 de junio de 2009.
15. Task Force Ratings. Guide to Clinical Preventive Services, Second Edition. Available from: http://odphp.osophs.dhhs.gov/pubs/guidecps/PDF/APPA.PDF. Visitado el 22 de junio de 2009.
16. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001; 323: 334-336.
17. Scottish Intercollegiate Guidelines Network (SIGN). SIGN 50, A guideline developer's handbook, revisited edition 2008. Available from: http://www.sign.ac.uk/pdf/sign50.pdf. Visitado el 23 de junio de 2009.
18. National Institute for Health and Clinical Excellence (NICE). The guidelines manual 2009. Available from: http://www.nice.org.uk/media/5F2/44/The_guidelines_manual_2009_-_All_chapters.pdf. Visitado el 22 de junio de 2009.
19. MacDonald JS, Smalley SR, Benedetti J, Hundahl SA, Estes NC, Stemmermann GN et al. Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or the gastroesophageal junction. N Engl J Med 2001; 345: 725-730.
20. Manterola C, Torres R, Burgos L, Vial M, Pineda V. Calidad metodológica de un artículo de tratamiento de cáncer gástrico adoptado como protocolo por algunos hospitales chilenos. Rev Méd Chile 2006; 134: 920-926.
21. Saha S, Hoerger TJ, Pignone MP, Teutsch SM, Helfand M, Mandelblatt JS; Cost Work Group, Third US Preventive Services Task Force. The art and science of incorporating cost effectiveness into evidence-based recommendations for clinical preventive services. Am J Prev Med 2001; 20 (3 Suppl): 36-43.