Jerarquización de la evidencia. Niveles de evidencia y grados de recomendación de uso actual

Carlos Manterola, Claudia Asenjo-Lobos y Tamara Otzen

Hierarchy of evidence. Levels of evidence and grades of recommendation from current use

There are multiple proposals and classifications that hierarchize evidence, which may confuse those who are dedicated to generate it both in health technology assessments, as for the development of clinical guidelines, etc. The aim of this manuscript is to describe the most commonly used classifications of levels of evidence and grades of recommendation, analyzing their main differences and applications so that the user can choose the one that better suits your needs and take this health decisions basing their practice on the best available evidence. A systematic literature search was performed in PubMed and MEDLINE databases and in Google, Yahoo and Ixquick search engines. A wealth of information concerning levels of evidence and degrees recommendation was obtained. It was summarized the information of the 11 proposals more currently used (CTFPHC, Sackett, USPSTF, CEBM, GRADE, SIGN, NICE, NHMRC, PCCRP, ADA y ACCF/AHA), between which it emphasizes the GRADE WORKING GROUP, incorporated by around 90 national and international organizations such as the World Health Organization, The Cochrane Library, American College of Physicians, American Thoracic Society, UpToDate, etc.; and locally by the Ministry of Health to create clinical practice guidelines.

Key words: "Evidence-Based Practice"[Mesh], "Evidence-Based Medicine", levels of evidence, grades of recommendation, clinical recommendation.

Palabras clave: Práctica clínica basada en la evidencia; medicina basada en la evidencia niveles de evidencia, grados de recomendación, recomendación clínica.

Universidad de La Frontera, Temuco, Chile.
Departamento de Cirugía y Traumatología (CM).
Universidad de Concepción, Concepción, Chile.
Centro de Rehabilitación Avanzada e Implantología-C RAI- (CA).
Universidad Autónoma de Chile.
Escuela de Psicología (TO).
Conflictos de interés: Ninguno.
Financiamiento: Programa de Doctorado en Ciencias Médicas, Universidad de La Frontera.

Recibido: 30 de marzo de 2014
Aceptado: 31 de mayo de 2014

Correspondencia a:
Carlos Manterola Delgado
carlos.manterola@ufrontera.cl

Introducción

Los primeros indicios de la práctica de la medicina basada en la evidencia (MBE), se remontan a los tiempos del emperador Qianlong, cuando se utilizaba el método "kaozheng" que representa la "práctica de la investigación probatoria"[1,2]. Desde entonces, numerosas instancias aisladas se fueron desarrollando hasta que en la década de los ochenta, el grupo de la Universidad McMaster, liderado por David Sackett, desarrolló los principios de la enseñanza y práctica de la MBE que rigen hasta hoy, y que se pueden resumir como *"integrar la experiencia clínica individual con la mejor evidencia disponible a partir de la investigación sistemática"*[3-5].

En la búsqueda de la mejor evidencia disponible, es necesario otorgar un valor jerárquico a la evidencia disponible, con el objetivo de tener una guía para decidir si aplicar o no una determinada intervención, tratamiento o procedimiento; a partir de la cual se pueda emitir una recomendación basada en la solidez de la evidencia que la respalda.

La primera jerarquización de la evidencia fue formulada por la *Canadian Task Force on the Periodic Health Examination* en 1979. Ésta, se desarrolló para la evaluación de medidas preventivas. Fue adaptada posteriormente por la U.S. *Preventive Services Task Force* (USPSTF) en 1984, quienes organizaron los niveles de evidencia (NE) y grados de recomendación (GR) para sujetos asintomáticos, indicando cuáles procedimientos eran los más adecuados y cuáles debían ser evitados.

En general, las clasificaciones se basan en los diseños de los estudios de donde proviene la evidencia, asumiendo que algunos de ellos están sujetos a más sesgos que otros; y por ende, justifican más débilmente las decisiones clínicas. Por otro lado, el análisis constante de la evidencia disponible desde la perspectiva de los diferentes escenarios clínicos, permite establecer GR para el ejercicio de procedimientos diagnósticos, terapéuticos, preventivos, económicos en salud, etc.; e indican la forma de valorar la evidencia disponible en función de etiología, daño, morbilidad y complicaciones; pronóstico, historia natural y curso clínico, etc. de una enfermedad o evento de interés[6].

Ha existido una proliferación de propuestas y clasificaciones que jerarquizan la evidencia; y junto con ello, sus respectivos GR. Ambos, pueden confundir a quienes se dedican a generar la evidencia a través del reporte de resultados por medio de artículos primarios y revisiones sistemáticas; así como de la realización de documentos de resumen de la evidencia (*overviews*), elaboración de guías clínicas, evaluaciones de tecnología sanitaria, etc.

Esto es debido a las diferencias y a la complejidad de las escalas de valoración existentes. Adicionalmente, el lenguaje utilizado para expresar la información no es de uso rutinario en la práctica cotidiana[6], lo que determina aún más confusión e incertidumbre respecto de cuál aplicar.

El objetivo de este artículo es actualizar la información y describir las clasificaciones más utilizadas para valorar la evidencia en el ámbito de la salud, analizando sus principales diferencias y aplicaciones para que el usuario pueda elegir la que mejor se adapte a sus necesidades y tomar de este modo decisiones sanitarias basando su práctica en la mejor evidencia disponible.

Clasificación de la evidencia y tipos de estudio

En las dos últimas décadas se ha producido un incremento significativo de la investigación clínica basada en la evidencia, como pilar fundamental en la toma de decisiones para los cuidados en salud. Sin embargo, no todos los conocimientos provenientes de artículos científicos tienen el mismo impacto o valor sobre la toma de decisiones; por tal razón, se debe aplicar un método riguroso para compilar la evidencia científica en torno a una pregunta; analizar de forma crítica los artículos científicos de los que disponemos para responder a la interrogante en cuestión, valorando la validez interna (metodología empleada y riesgo de sesgos), el impacto de los resultados y la validez externa del artículo (posible reproducibilidad de los resultados en la población que nos interesa). Previo a ello, el o los artículos seleccionados deben ser asignados a algún escenario clínico (tratamiento, prevención, etiología, daño, pronóstico e historia natural, diagnóstico diferencial, prevalencia, tamizaje, estudios económicos y análisis de decisión), a propósito de lo cual se elegirá la guía de usuario o pauta de lectura crítica correspondiente con la que se evaluará cada artículo.

Una vez identificado el escenario en el que corresponde catalogar al artículo (en ocasiones, puede ser asignado a más de uno), se aplica según el tipo de diseño del estudio en cuestión, la propuesta de NE y GR. Por lo anteriormente expuesto, es fundamental identificar el escenario y los diseños de investigación para poder valorar la evidencia utilizando alguna de las clasificaciones existentes.

Realizamos una búsqueda sistemática de la literatura en las bases de datos PubMed y MEDLINE y en los buscadores Google, Yahoo e Ixquick, utilizando los términos: medicina basada en la evidencia, evidencia, niveles de evidencia, grados de recomendación, fuerza de recomendación; en idiomas inglés y castellano. Obtuvimos una gran cantidad de información referente a niveles de evidencia y grados de recomendación, para finalmente resumir la información de 11 de las propuestas más utilizadas en la actualidad (CTFPHC, Sackett, USPSTF, CEBM, GRADE, SIGN, NICE, NHMRC, PCCRP, ADA y ACCF/AHA), que son las que se describen a continuación.

Grados de recomendación	Interpretación
A	Existe buena evidencia para recomendar la intervención clínica de prevención
B	Existe evidencia moderada para recomendar la intervención clínica de prevención
C	La evidencia disponible es contradictoria 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 evidencia moderada para NO recomendar la intervención clínica de prevención
E	Existe buena evidencia para NO recomendar la intervención clínica de prevención
I	Existe evidencia insuficiente (cualitativa y cuantitativamente) para hacer una recomendación; sin embargo, otros factores podrían influenciar en la decisión

Figura 1. Grados de recomendación para las intervenciones de prevención (CTFPHC).

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

Figura 2. Niveles de evidencia e interpretación de los tipos de estudio para intervenciones de prevención (CTFPHC).

Canadian Task Force on Preventive Health Care

La Canadian Task Force on Preventive Health Care (CTFPHC), fue elaborada por la Public Health Agency of Canada (PHAC) para desarrollar guías de práctica clínica que respaldasen las acciones de salud preventiva[7].

En sus inicios, este grupo hizo énfasis en el tipo de diseño utilizado y la calidad de los estudios publicados, basándose en los siguientes elementos: Un orden para los GR 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 la maniobra o acción determinada (se recomienda no hacer); la letra C, indica que la evidencia es "contradictoria"; y la letra I, indica insuficiencia en calidad y cantidad de evidencia disponible para establecer una recomendación (Figura 1)[8]. Y, NE clasificados según el diseño de los estudios de I a III, disminuyendo en calidad según se acrecienta numéricamente, para lo cual son clasificados según validez interna o calidad metodológica del estudio (Figuras 2 y 3).

Adicionalmente, CTFPHC se apoya en el sistema GRADE (*Grading of Recommendations Assessment, Development and Evaluation*)[9,10], para evaluar la calidad de la evidencia y realizar recomendaciones en el ámbito de la prevención.

Clasificación de Sackett

Esta sistematización propuesta por el epidemiólogo David L. Sackett, jerarquiza la evidencia en niveles que van de 1 a 5; siendo el nivel 1 la "mejor evidencia" y el nivel 5 la "peor, la más mala o la menos buena" (Figura 4)[11].

Las recomendaciones en apoyo de una intervención pueden ser generadas en base a estos cinco NE. De este modo, estudios nivel 1 conllevan a un GR A: resultados apoyados por estudios; nivel 2, reciben un GR B y las recomendaciones C se asignan a los resultados apoyados por estudios nivel 3, 4 ó 5. Así, el nivel indica el grado de certeza, generado por la fuerza de la evidencia. Grado A: Las conclusiones se generan a partir de la evidencia más fuerte de la investigación y por tanto son los más definitivos. Grado B: Las conclusiones se basan en pruebas más débiles y sólo son orientativas. Grado C: Las conclusiones se basan en pruebas débiles, por lo que son las menos fiables.

La principal desventaja es que al no existir sub categorías en algunas situaciones es difícil entregar un GR; por ejemplo, en el caso de un ensayo clínico (EC) con una muestra pequeña y riesgo de sesgo moderado, el GR ¿sería A o B? La otra desventaja que presenta es que dependen fundamentalmente de diseños clásicos y robustos, y no considera estudios menos habituales o rigurosos (por ejemplo estudios de corte transversal, poblacionales, etc.). Sin embargo, esta clasificación fue pionera y ha servido de base para el desarrollo de clasificaciones más completas, como la propuesta del Centre for Evidence-Based Medicine (CEBM) y otras.

Validez interna	Interpretación
Buena	Un estudio (incluido RS y meta-análisis) que cumple los criterios específicos de un estudio bien diseñado.
Moderada	Un estudio (incluido RS y meta-análisis) que no cumple (o no está claro que cumpla) al menos uno de los criterios específicos de un estudio bien diseñado, aunque no tenga defectos metodológicos graves.
Insuficiente	Un estudio (incluido RS y meta-análisis) que tiene en su diseño al menos un defecto metodológico grave, o que no cumple (o no está claro que cumpla), al menos uno de los criterios específicos de un estudio bien diseñado. O, que no tenga defectos metodológicos graves, pero que acumule defectos menores que hagan que los resultados del estudio no permitan plantear recomendaciones.

Figura 3. Validez interna e interpretación de los tipos de estudio para intervenciones de prevención (CTFPHC).

U.S. Preventive Services Task Force

El U.S. Preventive Services Task Force (USPSTF) es un grupo independiente de expertos en prevención y MBE, creado en 1984 en EE.UU. Es el encargado de valorar de forma rigurosa la investigación clínica con el fin de evaluar los méritos de las medidas preventivas, incluidas las pruebas de detección, servicios de asesoramiento, vacunas y medicamentos preventivos[12].

Generaron una jerarquización, estableciendo la fuerza de sus recomendaciones a partir de la calidad de la evidencia y del beneficio neto, el que fue definido como beneficio menos daño del servicio preventivo, evaluado tal como se aplica en la atención primaria a la población general.

El USPSTF asigna un nivel de certeza para evaluar el beneficio neto de un servicio preventivo basado en la naturaleza de la evidencia total disponible para sustentar el GR (Figura 5).

GR	NE	Terapia, prevención, etiología y daño	Pronóstico	Diagnóstico	Estudios económicos
A	**1a**	RS de EC con AA	RS con homogeneidad y Meta-análisis de estudios de cohortes concurrentes	RS de estudios de diagnóstico nivel 1	RS de estudios económicos nivel 1
	1b	EC con AA e intervalo de confianza estrecho	Estudio individual de cohortes concurrente con seguimiento superior del 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 de estudios de cohortes	RS de estudios de cohortes históricas	RS de estudios de diagnósticos de nivel mayor que 1	RS de estudios económicos de nivel mayor que 1
	2b	Estudios de cohortes individuales. EC de baja calidad	Estudio individual de cohortes históricas	Comparación independiente y 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, con análisis de sensibilidad
C	**4**	Series de casos. Estudios de cohortes y de 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 investigación económica
AA: Asignación aleatoria.					

Figura 4. Clasificación de los niveles de evidencia según Sackett.

Niveles de certeza	Descripción
Alta	La evidencia disponible incluye resultados consistentes de estudios bien diseñados, bien conducidos en poblaciones representativas de atención primaria. Estos estudios evalúan los efectos del servicio de prevención en la salud. Esta conclusión es por lo tanto poco probable que sea fuertemente afectada por los resultados de futuros estudios
Moderada	La evidencia disponible no es suficiente para determinar los efectos de la acción preventiva, pero la confianza en la estimación se ve limitada por factores tales como: • Número, tamaño o calidad de los estudios individuales • Inconsistencia de los resultados entre los estudios individuales • Generalización limitada de los resultados en la práctica habitual la atención primaria • Falta de coherencia en la cadena de la evidencia existente A medida que más información se encuentre disponible, la magnitud o la dirección del efecto observado podría cambiar, y este cambio puede ser lo suficientemente importante como para alterar la conclusión
Baja	La evidencia disponible es insuficiente para evaluar los efectos sobre los resultados de salud, debido a: • Limitado número o tamaño de los estudios • Defectos importantes en el diseño del estudio o los métodos • Inconsistencia de los resultados entre los estudios individuales • Lagunas en la cadena de la evidencia existente • Hallazgos no generalizables para la práctica habitual de la atención primaria • Falta de información sobre los resultados importantes de salud Más información puede permitir la estimación de los efectos sobre los resultados de salud

Figura 5. Descripción de los niveles de evidencia para exámenes periódicos de salud (USPSTF).

Recomendación	Interpretación	Sugerencia para la práctica
A	Se recomienda la acción preventiva. Existe alta certeza que el beneficio neto es substancial	Ofrecer o proporcionar este servicio
B	Se recomienda la acción preventiva. Hay una alta certeza de que el beneficio neto es moderado o existe moderada certeza de que el beneficio neto es de moderado a sustancial	Ofrecer o proporcionar este servicio
C	Se recomienda selectivamente el ofrecimiento o la prestación de este servicio a los pacientes individuales basadas en criterios profesionales y las preferencias del paciente. Hay por lo menos moderada certeza que el beneficio neto es pequeño	Ofrecer o proporcionar este servicio para los pacientes seleccionados en función de las circunstancias individuales
D	NO se recomienda la acción preventiva. Hay certeza moderada o alta que el servicio no tiene ningún beneficio neto o que los daños son mayores que los beneficios	Desalentar el uso de este servicio
I	Se concluye que la evidencia actual es insuficiente para evaluar el equilibrio entre los beneficios y los daños de la acción preventiva. La evidencia es deficiente, de mala calidad, o es contradictoria, y el balance de riesgos y beneficios no se puede determinar	Lea la sección de consideraciones clínicas de las recomendaciones de la USPSTF. Si el servicio es ofrecido, los pacientes deben comprender la incertidumbre que existe sobre el equilibrio entre beneficios y daños

Figura 6. Grados de recomendación para exámenes periódicos de salud (USPSTF).

A los GR se le asigna una letra (A, B, C, D, o I). Se apoyan en grados de certeza, que se definen como la probabilidad que el beneficio neto de un servicio preventivo que ha sido evaluado por la USPSTF sea correcto. El grado A, sugiere recomendar la acción ya que existe un alto grado de certeza que el beneficio neto es sustancial; el extremo opuesto es el grado I, que sugiere que no hay suficiente evidencia para evaluar el beneficio neto de una acción y por lo tanto, no se puede recomendar.

El USPSTF ha actualizado sus definiciones de las calificaciones que asigna a las recomendaciones y ahora incluye una columna de "sugerencias para la práctica" asociados con cada grado (Figura 6)[13].

Centre for Evidence-Based Medicine, Oxford

La propuesta del Centre for Evidence-Based Medicine de Oxford (CEBM) 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[14]. Es una innovación complementaria a otras iniciativas. Tiene la ventaja que gradúa la evidencia de acuerdo al mejor diseño para cada escenario clínico, otorgándole intencionalidad, agregando las revisiones sistemáticas (RS) en los distintos ámbitos. Por ejemplo, al tratarse de un escenario de pronóstico, 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 (Figura 7).

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[6].

El año 2009 se realizó una revisión de la tabla original con la idea de simplificarla para permitir una búsqueda heurística rápida y al mismo tiempo que permitiese jerarquizar la evidencia encontrada. Se adicionó un

GR	NE	Tratamiento, prevención, etiología y daño	Pronóstico e historia natural	Diagnóstico	Diagnóstico diferencial y prevalencia	Estudios económicos y de análisis de decisión
A	1a	RS con homogeneidad de EC con asignación aleatoria	RS de estudios de cohortes con homogeneidad (que incluya estudios con resultados comparables, en la misma dirección y validados en diferentes poblaciones)	RS de estudios de diagnóstico de alta calidad con homogeneidad (que incluya estudios con resultados comparables, en la misma dirección y en diferentes centros clínicos)	RS con homogeneidad de estudios de cohortes prospectivas	RS con homogeneidad de estudios económicos de alta calidad
	1b	EC individual con intervalo de confianza estrecho	Estudios de cohortes individuales, con un seguimiento mayor de 80% de las cohortes y validadas en una sola población	Estudios de cohortes que validen la calidad de una prueba específica, con estándar de referencia adecuado o a partir de algoritmos de estimación del pronóstico o de categorización del diagnóstico o probado en un centro clínico	Estudios de cohortes prospectivas con buen seguimiento	Análisis basado en costes o alternativas clínicamente sensibles; RS de la evidencia. Incluye análisis de sensibilidad
	1c	Todos o ninguna	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 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 de estudios de cohortes con homogeneidad	RS de estudios de cohortes históricas o de grupos controles no tratados en EC con homogeneidad	RS de estudios de diagnósticos de nivel 2 con homogeneidad	RS con homogeneidad de estudios 2b y mejores	RS con homogeneidad de estudios económicos con nivel mayor a 2
	2b	Estudios de cohortes individuales con seguimiento inferior a 80%. EC de baja calidad	Estudio individual de cohortes históricas o seguimiento de controles no tratados en un EC o guía de práctica clínica no validada	Estudios exploratorios que a través de una regresión logística determinan factores significativos y validados con estándar de referencia adecuado (independiente de la prueba diagnóstica)	Estudio individual de cohortes históricas o de seguimiento insuficiente	Análisis basado en costes o alternativas clínicamente sensibles; limitado a revisión de la evidencia. Incluye análisis de sensibilidad
	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 de estudios con homogeneidad de estudios 3b y mejor calidad	RS de estudios con homogeneidad de estudios 3b y mejor calidad	RS de estudios con homogeneidad de estudios 3b y mejor calidad
	3b	Estudios de casos y controles individuales		Comparación enmascarada y objetiva de un espectro de pacientes que podría 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 aplicación de un estándar de referencia		Estudio no consecutivo de cohorte, o análisis muy limitado de la población basado en pocas alternativas o costes, datos de mala calidad, pero con análisis de sensibilidad que incorporan variaciones clínicamente sensibles
C	4	Series de casos, estudios de cohortes y de casos y controles de baja calidad	Series de casos y estudios de cohortes de pronóstico de baja calidad	Estudios de casos y controles con escasos o sin estándares de referencia independientes	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 explícita, ni basada en fisiología, ni en trabajo de investigación juicioso, ni en "principios fundamentales"

Figura 7. Niveles de evidencia de CEBM (2009).

Pregunta	Paso 1 (Nivel 1*)	Paso 2 (Nivel 2*)	Paso 3 (Nivel 3*)	Paso 4 (Nivel 4*)	Paso 5 (Nivel 5*)
¿Qué tan común es el problema?	Encuestas locales y actuales aleatorias de la muestra (o censos)	RS de encuestas que coincidan con las circunstancias locales	Muestra local no aleatoria	Serie de casos	N/A
¿Es preciso el test de monitoreo o test diagnóstico? ***(Diagnóstico)***	RS de estudios transversales con estándar de referencia aplicado de forma consistente y con enmascaramiento	Estudios individuales de corte transversal con estándar de referencia aplicado de forma consistente y con enmascaramiento	Estudios no consecutivos, estudios sin un estándar de referencia aplicado de forma consistente	Estudios caso control o estándar de referencia pobre o no independiente	Mecanismos basados en el razonamiento
¿Qué pasaría si no se agrega una terapia? ***(Pronóstico)***	RS de estudios de cohorte de inicio	Estudios de cohorte de inicio	Estudio de cohorte o el brazo control de un EC con AA *	Estudio de caso-control o estudios de cohorte pronóstica de pobre calidad	N/A
¿Esta intervención ayuda? ***(beneficios del tratamiento)***	RS de EC con AA o ensayos n-de-1	EC con AA o estudios observacionales con un efecto dramático	Estudio de cohorte, con seguimiento controlado sin AA *	Serie de casos, estudios caso-control o estudios históricos controlados	Mecanismos basados en el razonamiento
¿Cuáles son los daños comunes? ***(efectos nocivos del tratamiento)***	RS de EC con AA, RS de estudios de casos y controles anidados, ensayo n-de-1 con el paciente sobre el que está planteando la pregunta, o estudio observacional con un efecto dramático.	EC individual con AA o (excepcionalmente) estudios observacionales con un efecto dramático	Estudio de cohorte, con seguimiento controlado (post-comercialización) con un número suficiente para descartar un daño común. (Para los daños a largo plazo, la duración del seguimiento debe ser suficiente)	Series de casos, estudios de casos y controles, o estudios históricos controlados	Mecanismos basados en el razonamiento
¿Cuáles son los daños raros? ***(efectos nocivos del tratamiento)***	RS de EC con AA ó ensayo n-de-1	EC con AA o (excepcionalmente) estudios observacionales con un efecto dramático			
¿Vale la pena esta prueba para detección temprana? ***(tamizaje)***	RS de EC con AA	EC con AA	Estudios de cohortes con seguimiento controlado	Serie de casos, estudios de casos y controles o estudios históricos controlados	Mecanismos basados en el razonamiento
*El NE se podrá clasificar hacia abajo en base a la calidad del estudio, de las imprecisiones, del carácter indirecto de la evidencia, debido a la inconsistencia entre los estudios, o porque el tamaño del efecto absoluto es muy pequeño, y el nivel se podrán clasificar hacia arriba si hay un tamaño de efecto grande o muy grande. AA : Asignación aleatoria					

Figura 8. Niveles de evidencia de CEBM (2011).

glosario con la definición de los términos relevantes, en forma precisa y entendible. Además, se consideró que las pruebas de "*screening*" o tamizaje debían considerarse como una entrada independiente y que se debía resaltar la importancia de las RS.

En esta nueva versión[15], se distinguen las filas que representan la serie de pasos que se deberían seguir en búsqueda de la mejor evidencia. Hacia el extremo izquierdo se encuentra la evidencia más fuerte y hacia la derecha la más débil. Las columnas representan los tipos de preguntas que el clínico pudiera encontrarse: ¿Qué tan común es el problema?, ¿Es precisa la prueba diagnóstica?, ¿Qué pasaría si no se instaura el tratamiento?, ¿La intervención ayudará al paciente?, ¿Qué tan frecuente o infrecuente es la complicación?, ¿Vale la pena utilizar esta prueba diagnóstica para detección precoz? (Figura 8).

En primer lugar se debe identificar a cuál de las preguntas de la columna izquierda corresponde el artículo del que proviene la evidencia que se desea valorar. Posteriormente, se debe identificar en la fila hacia la derecha, el diseño del estudio. Este punto es el que indica el NE, el que se identifica en la fila superior. Si bien la tabla es de fácil manejo y permite una valoración rápida y práctica de la evidencia, presenta al mismo tiempo la desventaja que al no existir subcategorías en relación a la calidad de los estudios, puede sesgar el valor que se le otorga a la evidencia. Por ejemplo, si la evidencia encontrada para una pregunta de terapia proviene de una RS, probablemente se categorizará como NE 1, pero si al continuar con la búsqueda nos encontramos con un EC con doble enmascaramiento y alta calidad metodológica que presenta un resultado contrario a la RS, ¿a cuál le creeremos? No sólo el NE que representa un determinado diseño va a avalar la decisión ya que pueden existir RS con fallas metodológicas que limitan la validez del estudio, situaciones en que las subcategorías serían de gran ayuda; sin embargo, al incluir estas subcategorías la simpleza de esta herramienta se pierde. Por lo tanto, CEBM sugiere que los NE se interpreten con una dosis de sentido común y buen juicio, lo que es posible lograr realizando una sistemática y exhaustiva búsqueda de la literatura científica que permita obtener los artículos relevantes y realizar un acucioso análisis crítico de la literatura antes de valorar la evidencia y aplicarla en la práctica si así lo amerita[16].

No obstante que CEBM actualizó la propuesta, es menester señalar que la anterior (Figura 7), sigue siendo utilizando de forma masiva por los grupos de investigación, que ven en ésta una mejora de la clasificación de Sackett.

Grado de recomendación. Descripción.	Beneficio vs. Riesgo y cargas	Calidad metodológica que apoya la evidencia	Implicancias
1A. Recomendación fuerte, evidencia de alta calidad	Los beneficios superan claramente los riesgos y cargas, o viceversa.	EC sin importantes limitaciones o evidencia abrumadora de estudios observacionales.	Recomendación fuerte, puede aplicarse a la mayoría de los pacientes en la mayoría de circunstancias, sin reserva.
1B. Recomendación fuerte, evidencia de moderada calidad	Los beneficios superan claramente los riesgos y cargas, o viceversa.	EC con importantes limitaciones (resultados inconsistentes, defectos metodológicos, indirectos o imprecisos) o pruebas excepcionalmente fuertes a partir de estudios observacionales.	Recomendación fuerte, puede aplicarse a la mayoría de los pacientes en la mayoría de circunstancias, sin reserva
1C. Recomendación fuerte, evidencia de baja o muy baja calidad	Los beneficios superan claramente los riesgos y cargas, o viceversa.	Estudios observacionales o series de casos.	Recomendación fuerte, pero puede cambiar cuando se disponga de mayor evidencia de calidad.
2A. Recomendación débil, evidencia de alta calidad	Beneficios estrechamente equilibrados con los riesgos y la carga.	EC sin importantes limitaciones o evidencia abrumadora de estudios observacionales.	Recomendación débil, la mejor acción puede variar dependiendo de las circunstancias de los pacientes o de los valores de la sociedad.
2B. Recomendación débil, evidencia de moderada calidad	Beneficios estrechamente equilibrados con los riesgos y la carga.	EC con importantes limitaciones (resultados inconsistentes, defectos metodológicos, indirectos o imprecisos) o pruebas excepcionalmente fuertes a partir de estudios observacionales.	Recomendación débil, la mejor acción puede variar dependiendo de las circunstancias de los pacientes o de los valores de la sociedad.
2C. Recomendación débil, evidencia de baja o muy baja calidad	Incertidumbre en las estimaciones de beneficios, riesgos y cargas; los beneficios, riesgos, y la carga puede estar estrechamente equilibrado.	Estudios observacionales o series de casos.	Recomendaciones muy débiles, otras alternativas pueden ser igualmente razonables.

Figura 9. GRADE modificado: grados de recomendación.

Grade Working Group

GRADE (The Grading of Recommendations Assessment, Development and Evaluation), es un sistema para clasificar la calidad de la evidencia y fuerza de recomendación aplicable a una amplia gama de intervenciones y contextos. Fue elaborado en base a la experiencia previa con otras herramientas existentes para conseguir un sistema "más razonable, confiable y ampliamente aplicable"[17].

La principal diferencia de este sistema en relación a otros, es que GRADE no valora la calidad de un estudio individual; sino que le da un valor a la evidencia para una medida resultado en particular, a partir de varios estudios primarios.

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 ésta en circunstancias específicas y la situación de riesgo basal, que son los puntos claves evaluados en cada artículo (Figura 9).

Destaca en esta propuesta, la elaboración de una tabla de síntesis que se obtiene de forma sistemática y que se basa en la evaluación de la calidad de la evidencia según el tipo de diseño[18]: EC con asignación aleatoria: calidad alta; estudios observacionales: calidad baja; y cualquier otra evidencia: calidad muy baja.

El grado disminuye si es que en la calidad del estudio existe: una limitación importante (-1) o muy importante (-2); una inconsistencia importante (-1); incertidumbre respecto de si la evidencia es directa o indirecta: si es alguna (-1) y si es máxima (-2); información imprecisa o escasa (-1); y alta probabilidad de sesgo de información (-1).

Por otra parte, el grado aumenta si es que: la evidencia de la asociación es fuerte, con un RR > a 2 ó < 0,5 basado en evidencia consistente derivada de dos o más estudios observacionales, sin factores de confusión plausibles (+1); o cuando la evidencia de la asociación es muy fuerte, con un RR > a 5 ó < 0,2 basado en evidencia directa, sin amenazas importantes para la validez (+2); o cuando existe evidencia de un gradiente dosis respuesta (+1); o cuando todos los potenciales factores de confusión posibles se han podido controlar en su efecto (+1).

Las ventajas del sistema GRADE respecto de los otros sistemas de clasificación son: que utiliza definiciones explícitas y juicios secuenciales durante el proceso de clasificación; que proporciona una descripción detallada de los criterios para la calidad de la evidencia para los resultados individuales y para la calidad general de la evidencia; que pesa la importancia relativa de los resultados; que considera el equilibrio entre los beneficios de salud versus los daños, costos y gastos; y, que permite desarrollar perfiles de evidencia y resúmenes de los hallazgos[19].

Es una herramienta muy completa que incorpora el uso de un software de uso libre para extraer los datos y realizar la síntesis, por ende, requiere tiempo para su aplicación ya que el análisis de un estudio individual es muy exhaustivo. Ha sido incorporada por diversas instituciones para para evaluar la calidad de la evidencia disponible, realizar recomendaciones y generar guías de práctica clínica. Entre ellas, destacan organismos internacionales tales como World Health Organization, Cochrane Library, World Allergy Organization (WAO), Surviving Sepsis, UpToDate, etc; instituciones norteamericanas como CDC's Healthcare Infection Control Practices

Advisory Committee (HICPAC), CDC's Division of Viral Hepatitis, Infectious Diseases Society of America, CDC's Advisory Committee on Immunization Practices (ACIP), Agency for Healthcare Research and Quality (AHRQ), The University of Pennsylvania Health System Center for Evidence-based Practice, American College of Physicians, American Thoracic Society, American Gastroenterological Association (AGA), The American Society of Colon and Rectal Surgeons, The Eastern Association for the Surgery of Trauma (EAST), American College of Chest Physicians, American Society for Gastrointestinal Endoscopy (ASGE), The American Association for the Study of Liver Diseases, etc; organizaciones europeas como British Medical Journal (UK), Clinical Evidence (UK), National Institute for Clinical Excellence (NICE, UK), The Scottish Intercollegiate Guidelines Network (SIGN, UK), NHS Quality Improvement (UK), German Center for Evidence-based Nursing "sapere aude" (Alemania), Evidence-based Nursing Südtirol, Alto Adige (Italia), European Society of Thoracic Surgeons, European Respiratory Society, European Association for the Study of the Liver (EASL), Belgian Centre for Evidence-Based Medicine (CEBAM), etc; instituciones canadienses como Health Quality Ontario, The Canadian Agency for Drugs and Technologies in Health, Evidence-Based Tuberculosis Diagnosis, The Canadian Cardiovascular Society, La Société Canadienne de Cardiologie, Canadian Task Force on Preventive Health Care (CTFPHC), etc.[10]; a las que habría sumar alrededor de 50 instituciones más[10]. En Chile, el Ministerio de Salud, ha hecho interesantes esfuerzos por incorporar GRADE en la elaboración de guías de práctica clínica.

NE	Interpretación
1++	Meta-análisis de alta calidad, RS de EC ó EC de alta calidad con muy poco riesgo de sesgo
1+	Meta-análisis bien realizados, RS de EC ó EC bien realizados con poco riesgo de sesgos
1-	Meta-análisis, RS de EC ó 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 bajo riesgo 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

Figura 10. Niveles de evidencia para estudios de tratamiento. Propuesta del SIGN.

Grado de recomendación	Interpretación
A	Al menos un meta-análisis, RS ó 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	Volumen de evidencia científica compuesta por estudios clasificados como 2 ++, directamente aplicable a la población blanco de la guía y que demuestran gran consistencia entre ellos; o evidencia científica extrapolada desde estudios clasificados como 1 ++ ó 1+
C	Volumen de evidencia científica compuesta por estudios clasificados como 2 + directamente aplicables a la población blanco 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+

Figura 11. Grados de recomendación para estudios de tratamiento. Propuesta del SIGN.

Scottish Intercollegiate Guidelines Network

El Scottish Intercollegiate Guidelines Network (SIGN) desarrolla guías de práctica clínica basadas en la evidencia, realizadas para el Servicio Nacional de Salud (NHS) de Escocia. Derivan de RS de la literatura científica y son diseñadas como un vehículo para acelerar la traducción del nuevo conocimiento en acción para cumplir con el objetivo de reducir la variabilidad de la práctica y mejorar los resultados relevantes para los pacientes[20].

La propuesta del SIGN, se originó teniendo como foco de interés la temática del tratamiento y los procedimientos terapéuticos. Se diferencia de las anteriores por su particular énfasis en el análisis cuantitativo que aportan las RS; y otorga además importancia a la reducción del error sistemático o sesgo (Figuras 10 y 11)[21].

Como fortaleza, es relevante destacar que considera la calidad metodológica de los estudios que componen las RS, situación de sumo interés, dada la alta producción anual de RS.

Como debilidad podemos señalar que no considera en la elaboración de las recomendaciones la realidad científica y tecnológica del momento, pues éstas se crean con una rigidez que puede ser peligrosa para quienes usan con ortodoxia las recomendaciones para la implementación de políticas de salud. 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 las implicancias económicas de las medidas recomendadas; situación que puede limitar su utilización en la práctica clínica latinoamericana.

Antiguamente, el SIGN se basaba en los NE desarrolladas por la Agencia de E.U.A. para el Cuidado de la Salud y la Investigación (AHCPR). Dadas las limitaciones encontradas con este sistema, desarrollaron una nueva clasificación que utilizan desde el año 2000[22].

Este sistema de clasificación pretende dar mayor peso a la calidad de la evidencia que respalda cada recomendación, y hacer hincapié en que el cuerpo de la evidencia debe ser considerado en su conjunto, y no depender de un sólo estudio para apoyar a cada recomendación. También pretende dar más peso a las recomendaciones respaldadas por estudios observacionales de gran calidad donde los EC no se pueden realizar por razones prácticas o éticas.

En 2009, el SIGN tomó la decisión de implementar el enfoque GRADE en su directriz metodológica, metodología que se encuentra actualmente en desarrollo[21].

National Institute for Health and Clinical Excellence

La iniciativa National Institute for Health and Clinical Excellence (NICE) nace del National Health Service del Reino Unido (NHS) y actualmente abarca la valoración de la evidencia en diferentes escenarios clínicos: tratamiento, diagnóstico, pronóstico y estudios de costo efectividad. Además incluye el tópico de la experiencia del paciente para informar preguntas de revisión (estudios cualitativos y encuestas en estudios transversales) y guías clínicas[23].

Desde el año 2009 utiliza la pauta GRADE para evaluar la calidad de la evidencia en terapia y procedimientos terapéuticos. No obstante ello, NICE difiere de GRADE en dos puntos: integra una revisión de la calidad de los estudios de coste-efectividad; y no utiliza etiquetas como "resumen global" para la calidad de la evidencia o la fuerza de una recomendación, sino que utiliza la redacción de recomendaciones a fin de reflejar la fuerza de la recomendación.

Para valorar la calidad de la evidencia en estudios de diagnóstico, utiliza la herramienta QUADAS-2; y la calidad de la evidencia en estudios de pronóstico se realiza mediante la aplicación de una lista de chequeo *ad-hoc*. Por otra parte, para evaluar la calidad de las guías clínicas utiliza el instrumento AGREE II.

La jerarquización de la evidencia y los GR para estudios de pruebas diagnósticas se pueden apreciar en las Figuras 12 y 13.

Finalmente cabe señalar, que NICE presenta la información a través de un resumen de la calidad de la evidencia encontrada, el que es elaborado tras la aplicación de la evaluación sistemática de la información, utilizando para ello las herramientas ya mencionadas y la ulterior redacción de recomendaciones[24].

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.

Figura 12. Niveles de evidencia para estudios de pruebas diagnósticas. Propuesta de la NICE.

Grados de Recomendación	Interpretación
A	Estudios de pruebas diagnósticas con un nivel de evidencia Ia o Ib
B	Estudios de pruebas diagnósticas con un nivel de evidencia II
C	Estudios de pruebas diagnósticas con un nivel de evidencia III
D	Estudios de pruebas diagnósticas con un nivel de evidencia IV

Figura 13. Grados de recomendación para estudios de pruebas diagnósticas. Propuesta de la NICE.

National Health and Medical Research Council

El National Health and Medical Research Council (NHMRC) se ha utilizado en Australia desde 1999. Es una tabla de jerarquía de la evidencia, creada con el objetivo de valorar la evidencia en las guías de práctica clínica y evaluación de tecnologías sanitarias (Figura 14).

Los manuales proponen valorar la evidencia en tres dimensiones: la fuerza, el tamaño del efecto y la relevancia clínica. En este esquema, la solidez de la evidencia determina el nivel, la calidad de la evidencia y su precisión estadística. Con el uso de esta forma de jerarquización, se ha observado que a menudo la evidencia obtenida no es susceptible de ser sometida a meta-análisis y por tanto su valoración se relaciona sólo a los estudios individuales.

La tabla actual (basada en la jerarquización del CEBM), es más amplia que la inicial y está estructurada de forma distinta. Contiene cinco columnas para cada una de las áreas de investigación (intervención, precisión diagnóstica, pronóstico, etiología y tamizaje); y una columna ubicada al extremo izquierdo de las anteriores presenta los NE.

Al valorar la evidencia se sugiere indicar siempre el área de investigación, por ejemplo, NE II Intervención, NE IV prueba diagnóstica.

Es muy completa pero poco práctica para el uso cotidiano. Va acompañada de un glosario y un formulario tipo para registrar cada uno de los pasos de la valoración hasta la ulterior recomendación[25].

Practicing Chiropractors' Committee on Radiology Protocols

El Practicing Chiropractors' Committee on Radiology Protocols (PCCRP) desarrolla guías clínicas, para lo cual utiliza una clasificación de NE adaptada de la clasificación utilizada por el United States Department of Health and Human Services[26]. Ésta, no se limita sólo a EC; sino que considera todos los tipos de estudios que pueden aportar evidencia para la práctica.

Además, el PCCRP utiliza la propuesta de GR presentada por Phillips y cols.[27], con una leve modificación para ajustarse a los NE no clínicos (Figura 15).

De este modo, agrupa la evidencia en cinco niveles según el diseño de los estudios:

Nivel I. EC controlados con asignación aleatoria, incluyendo estudios cuasi-aleatorios como también con asignación alternada.

Nivel	Intervención	Precisión diagnóstica	Pronóstico	Etiología	Tamizaje
I	RS de estudios nivel II	RS de estudios nivel II	RS de estudios nivel II	RS de estudios nivel II	RS de estudios nivel II
II	EC controlado, con AA	Estudios de precisión de PD con una comparación enmascarada e independiente con un estándar de referencia válido, entre sujetos consecutivos con una presentación clínica definida	Estudios de cohortes prospectivas	Estudios de cohortes prospectivas	EC controlado con AA
III-1	EC pseudoaleatorizado controlado (ej. asignación alternada o algún otro método)	Estudios de precisión de PD con una comparación enmascarada e independiente con un estándar de referencia válido, entre sujetos no consecutivas con una presentación clínica definida	Todo o ninguno	Todo o ninguno	EC controlado pseudoaleatorizado (por ejemplo, asignación alternada o algún otro método)
III-2	Estudios comparativo con controles concurrentes: • EC experimental sin AA • Estudios de cohortes • Estudios casos y controles • Series temporales interrumpidas con un grupo control	Comparación con un estándar de referencia que no cumple con el criterio requerido para el NE II y III-1	Análisis de los factores pronósticos entre los sujetos de un solo brazo de un EC controlado con AA	Estudios de cohortes retrospectivas	Estudios comparativos con controles concurrentes: • Ensayo experimental sin AA • Estudios de cohortes • Estudios de casos y controles
III-3	Estudios comparativos s/controles concurrentes: • Estudios con controles históricos • EC Dos o más estudios de un solo brazo • EC Series temporales interrumpidas sin grupo control paralelo	Estudios de casos y controles de diagnóstico	Estudios de cohortes retrospectivas	Estudios de casos y controles	Estudios comparativos sin controles concurrentes: • Estudios con controles históricos • EC Dos o más estudios de un solo brazo
IV	Series de casos, ya sea con resultados post-test o pre-test/post-test	Estudios de rendimiento diagnóstico sin estándar de referencia	Series de casos, o estudios de cohortes de sujetos en diferentes etapas de la enfermedad	Estudios de corte transversal o series de casos	Series de casos
AA: Asinación aleatoria. PD: Prueba diagnóstica.					

Figura 14. Niveles de evidencia según NHMRC.

Tipo de estudios	Grados A - D	Grados a - d
Nivel clínico I	A= Estudios nivel I consistentes ó RS ó meta análisis	
Nivel clínico II	B= Estudios nivel II consistentes ó un sólo estudio nivel I	
Nivel clínico III	B= Estudios nivel III consistentes	
Nivel clínico IV	C= Estudios nivel IV consistentes ó extrapolaciones de nivel II o III	
Opinión de experto V	D= Evidencia nivel V ó estudios inconsistentes de nivel I – IV	
Estudio poblacional		A = Estudios clase I consistentes B = Un solo estudio clase I o estudios consistentes clase II y III C = Estudios consistentes clase IV D = Evidencia no concluyente
Ciencias básicas, biomecánicos, estudios de validez. Encuestas profesionales		A = Estudios consistentes B = Un solo estudio positivo D = Estudios no concluyente
Estudios de confiabilidad		A = Estudios clase I consistentes B = Un solo estudio clase I o estudios consistentes clases II C = Un solo estudio clase II D = Evidencia no concluyente

Figura 15. Grados de recomendación PCCRP.

Nivel II. EC sin asignación aleatoria (estudio prospectivo, previamente planificado, con criterios de elegibilidad y medidas de resultado pre-determinados).

Nivel III. Estudios observacionales. Incluye estudios de cohortes prospectivos y retrospectivos, estudios de casos y controles; y la investigación de servicios de salud que incluya ajustes por posibles variables de confusión.

Nivel IV. Estudios observacionales sin grupo control (estudios de cohortes sin controles y series de casos).

Y, nivel V, que corresponde a la opinión de expertos.

NE	Descripción
A	Clara evidencia de EC controlados con asignación aleatoria, generalizables bien realizados, que están adecuadamente conducidos, incluyendo: • Evidencia de un EC multicéntrico bien conducido • Evidencia de un meta-análisis que incorpora las calificaciones de calidad en el análisis Evidencia no experimental convincente, es decir, la norma "todo o nada" desarrollada por el CEBM. Evidencia de apoyo de EC controlados bien realizados que están adecuadamente conducidos, incluyendo: • Evidencia de un EC bien realizado en una o más instituciones • Evidencia de una RS que incorpora las calificaciones de calidad en el análisis
B	Evidencia de apoyo de estudios de cohortes bien realizados, que incluyen: • Evidencia de un estudio de cohortes prospectivo bien realizado o registro • Evidencia de un meta-análisis de estudios de cohortes bien realizado Evidencia de apoyo de un estudio de casos y controles bien realizado
C	Evidencia de apoyo de estudios pobremente controlados o no controlados, que incluyen: • Evidencia de los EC con asignación aleatoria con defectos metodológicos menores, que pudieran invalidar los resultados • Evidencia de estudios observacionales con un alto potencial de el sesgo (como una serie de casos con comparación con controles históricos) • Evidencia de serie de casos o reporte de casos
E	Consenso de expertos o experiencia clínica

Figura 16. Niveles de evidencia de la ADA.

American Diabetes Association

La American Diabetes Association (ADA), ha participado activamente en el desarrollo y difusión de las normas de atención, directrices y documentos relacionados a la diabetes mellitus, aportando un sistema de clasificación para calificar la calidad de la evidencia científica que apoya sus recomendaciones disciplinarias (Figura 16)[28].

Estos niveles agrupados por letras, van de la A a la E y en base a esta valoración se realiza la recomendación.

Es una propuesta simple, en la que el NE ofrece directamente un GR, por ejemplo: un nivel A aporta un GR A.

Las recomendaciones se asocian a evidencia proveniente de estudios clínicos (niveles A-D). Por otra parte, un GR E se considera una categoría separada en las que no existe evidencia proveniente de EC; o, en situaciones en las que los que los EC pueden ser poco prácticos; o, en aquellas circunstancias en las que hay evidencia contradictoria.

The American College of Cardiology Foundation/ American Heart Association Task Force on Practice Guidelines

El American College of Cardiology Foundation / American Heart *Association (*ACCF/AHA) Task Force es el encargado de elaborar, actualizar y revisar las directrices prácticas para las enfermedades cardiovasculares y sus respectivos procedimientos.

En el análisis de los datos y la elaboración de recomendaciones de las guías de práctica clínica, el comité de redacción utiliza metodologías basadas en la evidencia desarrolladas por el *Task Force*. La "clase de la recomendación" (COR) es una estimación del tamaño del efecto de un tratamiento; teniendo en cuenta los riesgos versus los beneficios, evidencia y grado de acuerdo que un determinado tratamiento o procedimiento es o no útil/efectivo, o que en algunos casos puede causar daño[29].

El esquema para COR y NE se resume en la Figura 17, que también ofrece frases sugeridas para escribir recomendaciones para cada COR.

Un aporte a esta metodología es la separación de las recomendaciones; por ejemplo: Clase III para delimitar si la recomendación es considerada "ningún beneficio", o se asocia con "daño" al paciente.

Por otro lado, dado el número cada vez mayor de estudios de comparación de efectividad, se sugieren verbos y frases de comparación para la redacción de recomendaciones para la eficacia comparativa de un tratamiento o estrategia frente a otro (pero sólo para COR I y IIa, NE A o B).

Lo interesante de este enfoque es que cualquier combinación de COR y NE es posible. Por ejemplo, una recomendación puede tener Clase I, incluso si se basa exclusivamente en la opinión de expertos y no hay estudios de investigación que se hayan llevado a cabo. Del mismo modo, una Clase IIa o IIb se le puede asignar un Nivel A, si hay varios EC controlados con asignación aleatoria que llegan a conclusiones divergentes.

Hacen énfasis en que la asignación de un NE B o C no debe interpretarse en términos que la recomendación es débil. Muchas de las preguntas clínicas importantes tratadas en las guías clínicas no se prestan a la experimentación o aún no han sido abordadas por las investigaciones de alta calidad. A pesar que los EC controlados con asignación aleatoria pueden no estar disponibles, la pregunta clínica puede ser tan relevante que sería negligente no incluirla en la guía.

Discusión

El paso fundamental para la aplicación clínica de la evidencia en torno a un problema clínico es darle un valor a la evidencia disponible que nos permita ejercer

	Clase I **Beneficio>>>Riesgo** Procedimiento/tratamiento DEBERIA ser realizado o administrado	Clase IIa **Beneficio>>Riesgo** Se requiere de estudios adicionales con objetivos focalizados. ES RAZONABLE realizar un procedimiento o administrar un tratamiento	Clase IIb **Beneficio ± Riesgo** Se requiere de estudios adicionales con objetivos amplios; datos de registros adicionales podrían ser de ayuda. PODRIA CONSIDERARSE el procedimiento o tratamiento	Clase III **Riesgo ± Beneficio** Procedimiento/tratamiento NO debería ser realizado o administrado PUESTO QUE NO ES ÚTIL Y PUEDE SER PERJUDICIAL
Nivel A **Múltiples poblaciones evaluadas*** **Datos derivados de múltiples EC con AA ó RS**	- Recomendación: El tratamiento es útil o efectivo - Suficiente evidencia de múltiples EC ó RS	- Recomendación: a favor del tratamiento o procedimiento siendo útil o efectivo - Alguna evidencia conflictiva de múltiples EC ó RS	- Recomendación de la utilidad/ eficacia menos establecida - Mayor evidencia contradictoria de múltiples EC ó RS	- Recomendación: El procedimiento o tratamiento no es útil o efectivo y podría ser dañino - Evidencia suficiente de múltiples EC ó RS
Nivel B **Poblaciones limitadas evaluadas*** **Datos derivados de EC simples o estudios sin AA**	- Recomendación: El procedimiento o tratamiento es útil o efectivo - Evidencia de un EC simple ó estudios sin AA	- Recomendación a favor del tratamiento o procedimiento siendo útil o efectivo - Alguna evidencia contradictoria proveniente de un solo EC ó de estudios sin **con** AA	- Recomendación de la utilidad/ eficacia menos establecida - Mayor evidencia contradictoria de un solo EC ó estudios sin AA	- Recomendación: El procedimiento o tratamiento no es útil o efectivo y podría ser dañino - Evidencia de un solo EC ó de estudios sin AA
Nivel C **Poblaciones muy limitadas evaluadas*** **Sólo consensos de opiniones de expertos, estudios de casos o de cuidados estándar**	- Recomendación: El procedimiento o tratamiento es útil o efectivo - Sólo opinión de expertos, estudios de casos o de cuidados estándar	- Recomendación: a favor del tratamiento o procedimiento siendo útil o efectivo - Sólo opinión de expertos contradictoria, Estudios de casos o de cuidados estándar	- Recomendación de la utilidad/ eficacia menos establecida - Sólo opinión de expertos divergente, Estudios de casos o de cuidados estándar	- Recomendación: El procedimiento o tratamiento no es útil o efectivo y podría ser dañino - Sólo opinión de expertos, Estudios de casos o de cuidados estándar
AA : Asignación aleatoria.				

Figura 17. Niveles de evidencia del ACCF/AHA Task Force on Practice Guidelines.

el juicio clínico en base al grado de confiabilidad que los resultados de las investigaciones científicas nos entregan y que permita evaluar beneficios vs riesgos, a la luz de la evidencia actual.

Existen múltiples propuestas de clasificaciones para valorar la evidencia (se estima que superan el centenar); desde la inicial y más simple (Sackett), que sirvió de base para las actuales; hasta otras que han querido considerar todas las posibles situaciones que se pueden encontrar, haciéndolas inmanejables debido a su extensión. Entre otras tantas, se pueden mencionar las de la Agency for Healthcare Research and Quality[30], de la Agència d'Avaluació de Tecnologia Mèdica (AATM) de la Generalitat de Catalunya[31], de la Canadian Asthma Consensus Group[32]; la de NANDA-I[33], del Servicio Andaluz de Salud[34], SORT[35], etc.

Junto a ello, el lenguaje epidemiológico empleado, ha llevado consigo confusión e incomprensión por parte de los médicos clínicos, quienes ven en esta gran variedad de opciones más conflictos que ayuda al desarrollo de la práctica profesional. En este artículo se presentan las clasificaciones más relevantes 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 general o particular.

Sin embargo, lo interesante de todo esto, es que el desarrollo creciente de las guías clínicas y la realización de evaluaciones de tecnologías sanitarias, ha determinado la necesidad de utilización de cada una de las iniciativas antes señaladas, tanto en universidades como establecimientos sanitarios; razón por la que es relevante conocerlas, de modo tal de comprender mejor la metodología a través de la cual se efectúan estos procesos.

La sistematización de la búsqueda de la evidencia es el paso fundamental para la obtención de la mejor evidencia disponible. Sin embargo, al momento de acercarnos a ella siempre se ha de considerar la existencia de características poblacionales, culturales, económicas, tecnológicas, ambientales, etc.; es decir, darle relevancia al concepto de validez externa "sin importar" lo que "parece resultar" en otras latitudes, sin pruebas previas en nuestra realidad.

El paso siguiente corresponde a darle un valor a esta evidencia, para lo cual se debe elegir la clasificación que más se ajuste a nuestra necesidad y que permita discriminar entre un nivel y otro, para realizar las recomendaciones más adecuadas al entorno asistencial y poblacional. Es así como se dispone de clasificaciones amplias, con una propuesta distinta para diferentes escenarios (CEBM, NHMRC, etc.), como de clasificaciones más específicas, que apuntan a un escenario puntual (CTFPHC, SIGN, etc.) y otras como GRADE que ha sido incorporada para evaluar la calidad de la evidencia disponible, realizar recomendaciones y generar guías de práctica clínica basadas en la evidencia, por alrededor de 90 prestigiosas instituciones nacionales e internacionales[10].

Finalmente, nos parece oportuno señalar, que este manuscrito aporta el resumen de la información referente a los 11 sistemas, propuestas y clasificaciones de jerarquización de la evidencia y GR de mayor uso en la actualidad

Clasificación	Tratamiento	Prevención	Etiología	Daño	Pronóstico	Diagnóstico	Prevalencia	Tamizaje	Económicos
CTFPHC		X							
SACKETT	X	X	X	X	X	X			X
USPSTF						X			
CEBM	X	X	X	X	X	X	X		X
GRADE	X								
SIGN	X								
NICE	X				X	X			X
NHMRC	X		X		X	X		X	
PCCRP	X								
ADA	X								
ACCF/AHA	X								

Figura 18. Clasificaciones de niveles de evidencia y grados de recomendación analizadas. Se puede observar que algunas de ellas son específicas para cierto tipo de escenarios.

(Figura 18); actualización de la información publicada en 2009[6], documento en el que se plasmó la información relacionada a las seis propuestas de jerarquización más utilizadas en ese momento.

Resumen

Existen múltiples propuestas y clasificaciones que jerarquizan la evidencia, que pueden confundir a quienes se dedican a generar la evidencia tanto en evaluaciones de tecnología sanitaria, elaboración de guías clínicas, etc. El objetivo de este artículo es actualizar la información y describir las clasificaciones más utilizadas para valorar la evidencia en el ámbito de la salud, analizando sus principales diferencias y aplicaciones para que el usuario pueda elegir la que mejor se adapte a sus necesidades y tomar de este modo decisiones sanitarias basando su práctica en la mejor evidencia disponible. Se realizó una búsqueda sistemática de la literatura en las bases de datos PubMed y MEDLINE y en los buscadores Google, Yahoo e Ixquick. Se obtuvo una gran cantidad de información referente a niveles de evidencia y grados de recomendación, para finalmente resumir la información de 11 de las propuestas más utilizadas en la actualidad (CTFPHC, Sackett, USPSTF, CEBM, GRADE, SIGN, NICE, NHMRC, PCCRP, ADA y ACCF/AHA), entre las que destaca la del GRADE WORKING GROUP, incorporada por alrededor de 90 organizaciones nacionales e internacionales, tales como la World Health Organization, The Cochrane Library, American College of Physicians, American Thoracic Society, UpToDate, etc. y a nivel local por el Ministerio de Salud, para generar guías de práctica clínica.

Referencias bibliográficas

1.- Sackett D L, Rosenberg W M, Gray J A, Haynes R B, Richardson W S. Evidence based medicine: what it is and what it isn't. Br Med J 1996; 312: 71-2.
2.- Sackett D L, Richardson W S, Rosenberg W H R. Evidence-based medicine: how to practice and teach EBM. (Churchill Livingstone, 2000).
3.- Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992; 268: 2420-5.
4.- 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: 125-30.
5.- Manterola C. Medicina basada en la evidencia. Conceptos generales y razones para aplicación en cirugía. Rev Chil Cir 2002; 55: 550-4.
6.- Manterola C, Zavando D. (Grupo MINCIR). Cómo interpretar los "Niveles de Evidencia" en los diferentes escenarios clínicos. Rev Chil Cir 2009; 61: 582-95.
7.- Birtwhistle R, Pottie K, Shaw E, Dickinson J A, Brauer P, Fortin M, et al. Canadian Task Force on Preventive Health Care: we're back! Can Fam Physician 2012; 58: 13-5.
8.- Canadian Task Force on Preventive Health Care. New grades for recommendations from the Canadian Task Force on Preventive Health Care. Can Med Assoc J 2003; 169: 207-8.
9.- Canadian Task Force on Preventive Health Care. Putting prevention into practice. Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group. Disponible en http://canadiantaskforce.ca/methods/grade. (acceso el 10 de marzo de 2014).
10.- Guyatt G H, Oxman A D, Vist G E, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Br Med J 2008; 336: 924-6.
11.- Sackett D L. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1989; 95 (2 Suppl): 2S-4S.
12.- Harris R P, Helfand M, Woolf S H, Lohr K N, Mulrow C D, Teutsch S M, et al; Methods Work Group, Third US Preventive Services Task Force. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001; 20 (3 Suppl): 21-35.
13.- U.S. Preventive Services Task Force. Grade Definitions. Disponible en http://www.uspreventiveservicestaskforce.org/uspstf/grades.htm. (acceso el 13 de febrero de 2014).
14.- Centre for Evidence-based Medicine (CEBM)-Levels of Evidence (March 2009). Disponible en http://www.cebm.net/index.aspx?o=1025. (acceso el 15 de febrero de 2014).

15.- Centre for Evidence-based Medicine (CEBM)-Levels of Evidence (2011). Disponible en http://www.cebm.net/mod_product/design/files/CEBM-Levels-of-Evidence-2.1.pdf. (acceso el 16 de febrero de 2014).
16.- Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, et al. Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document). (2011). Disponible en http://www.cebm.net/index.aspx?o=5653. (acceso el 16 de febrero de 2014).
17.- Canfield S E, Dahm P. Rating the quality of evidence and the strength of recommendations using GRADE. World J Urol 2011; 29: 311-7.
18.- Guyatt G, Oxman A D, Akl E A, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011; 64: 383-94.
19.- Guyatt G, Gutterman D, Baumann M H, Addrizzo-Harris D, Hylek E M, Phillips B, et al. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians Task Force. Chest 2006; 129: 174-81.
20.- SIGN. Methodological principles. Disponible en http://www.sign.ac.uk/methodology/index.html. (acceso el 18 de febrero de 2014).
21.- SIGN. SIGN 50 a guideline developer´s handbook. (Scottish Intercollegiate Guidelines Network, 2011). Disponible en http://www.sign.ac.uk/guidelines/fulltext/50/index.html. (acceso el 18 de febrero de 2014).
22.- Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. Br Med J 2001; 323: 334-6.
23.- National Institute for Health and Care Excellence. NICE web page. Disponible en http://www.nice.org.uk. (acceso el 21 de febrero de 2014).
24.- NICE. The guidelines manual: consultation on the 2012 update. Disponible en http://www.nice.org.uk/aboutnice/howwework/developingniceclinicalguidelines/GuidelinesManualConsultation2012.jsp. (acceso el 21 de febrero de 2014).
25.- Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian "levels of evidence". BMC Med Res Methodol 2009; 9: 34.
26.- PCCRP Document Section I. PCCRP web page. Disponible en http://www.pccrp.org. (acceso el 21 de febrero de 2014).
27.- Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B D M: The Oxford Centre for Evidence-based Medicine Levels of Evidence. Levels of Evidence and Grades of Recommendation (2001). Disponible en http://www.cebm.net/levels_of_evidence.asp. (acceso el 21 de febrero de 2014).
28.- American Diabetes Association. Introduction. ADA evidence-grading system for clinical practice recommendations. Diabetes Care 2008;31:S1-S2. Disponible en http://care.diabetesjournals.org/content/31/Supplement_1/S1.full. (acceso el 22 de febrero de 2014).
29.- Greenland P, Alpert J S, Beller G A, Benjamin E J, Budoff M J, Fayad Z A, et al; American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 2010; 122: 2748-64.
30.- Agency for Healthcare Research and Quality (AHRQ). EPC Evidence Reports [Internet]. Rockville: AHRQ. Disponible en http://www.ahrq.gov/clinic/epcindex.htm#methodology. (acceso el 19 de marzo de 2014).
31.- Jovell A J, Navarro-Rubio M D. Evaluación de la evidencia científica. Med Clin (Barc) 1995; 105: 740-3.
32.- Canadian Asthma Consensus Group. Disponible en http://www.lung.ca/cts-sct/pdf/Adult_Asthma_Consensus.pdf. (acceso el 19 de marzo de 2014).
33.- NANDA-I. Disponible en http://eldiagnosticoenfermero.blogspot.com.es/2010/12/niveles-de-evidencia-de-la.html. (acceso el 19 de marzo de 2014).
34.- Niveles de evidencia. Servicio Andaluz de Salud. Disponible en http://www.juntadeandalucia.es/servicioandaluzdesalud/hrs3/index.php?id=manual_procedimientos_niveles. (acceso el 19 de marzo de 2014).
35.- Ebell M H, Siwek J, Weiss B D, Woolf S H, Susman J, Ewigman B, et al. Strength of Recommendation Taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. Am Fam Physician 2004; 69: 548-56.