Camaño-Puig, R. (2019). "Evidencia científica, metodología y clasificaciones: niveles y recomendaciones", en Moreno-Castro, C. y Cano-Orón, L. (eds.) *Terapias Complementarias en la esfera pública.* Madrd: Dextra Editorial, págs. 259-293.

9. Evidencia científica, metodología y clasificaciones: niveles y recomendaciones

Ramón Camaño-Puig

Resumen

En este capítulo se recoge cómo se ha construido la imagen social de la medicina basada en la evidencia. La revisión sistemática de la evidencia disponible de cualquier intervención clínica se ha configurado como un procedimiento de investigación científica en sí mismo. Además, cuando resulta posible combinar los resultados de múltiples estudios (metanálisis), este método puede incluso aumentar la potencia y la precisión de las estimaciones de efectividad, lo que en las últimas décadas ha incidido en que se produzca un incremento significativo de la investigación clínica basada en la evidencia, un pilar fundamental en la toma de decisiones para los cuidados de la salud.

9.1. Introducción

Las revisiones sistemáticas permiten evaluar si múltiples estudios son consistentes entre sí y valorar si los hallazgos pueden generalizarse o si varían de unos grupos de población a otros. Por otra parte, cuando no se encuentran estudios que den respuesta a una pregunta concreta, en todo caso pueden suponer una ayuda a la hora de identificar áreas críticas y cuestiones que requieren de ulterior investigación.

En el campo de la salud se publican anualmente más de dos millones de artículos. Un volumen de información que resulta difícil de manejar y hace necesario facilitar el acceso de los profesionales de las distintas especialidades a resúmenes que faciliten el paso a los contenidos (Cochrane 1972) y que, de ese modo, les ayuden a mejorar la toma de decisiones.

En definitiva, se trata de evaluar el conocimiento y los argumentos científicos disponibles para orientar la toma decisión, tanto en guías para la práctica clínica como en las políticas y estrategias preventivas de salud pública. Sin embargo, no todos los conocimientos provenientes de artículos científicos tienen el mismo valor e impacto en la toma de decisiones; razón por la que debe aplicarse un método riguroso a la hora de compilar la evidencia científica en torno a una determinada pregunta. Así, es menester 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), así como el impacto de los resultados y la validez externa del artículo (posible reproducibilidad de los resultados en la población que nos interesa).

Podríamos decir que las teorías científicas son consistentes interna y externamente, y ahorrativas a la hora de proponer entidades o explicaciones y, por tanto, cumplen el principio de parsimonia. Son útiles, describen y explican fenómenos observables, y se pueden verificar y falsar empíricamente. Por consiguiente, son contrastables y están basadas en repetidos experimentos controlados. Son corregibles en función de los nuevos elementos observacionales que puedan aparecer, lo que les proporciona dinamismo; se hacen cambios en función del descubrimiento de nuevos datos. Las teorías científicas proceden a explicar lo que explicaban teorías previas y tienden a englobar todo el conocimiento, adquiriendo un carácter progresivo en la construcción del conocimiento, esto es, unas sobre otras. Una teoría puede carecer de alguno de los puntos anteriores, por ejemplo, en casos donde los experimentos no son estrictamente necesarios y se pueden llevar a cabo las consideraciones correspondientes con observaciones controladas y repetidas. Dicho esto, si una teoría falla en un número importante de los puntos considerados más arriba, se puede descartar con certeza, pues puede afirmarse que no es científica. Adicionalmente, se pueden realizar valoraciones en términos de calidad y validez de aquellos procedimientos que, cumpliendo todos los aspectos anteriormente enunciados, fundamentalmente se aplican para discernir si aquello que se plantea tiene la adecuada utilidad y es posible que sea llevado a cabo sin riesgos y de modo eficiente.

Realizar procedimientos de revisión sistemática mejora la reproducibilidad y validez. Además, estos incluyen una crítica científica de la información recuperada y separan la opinión de la evidencia, lo que redunda en una mejora de la objetividad. En esta línea, es de vital importancia hablar de conceptos tales como fiabilidad y validez. En términos coloquiales, ambos conceptos tienen un significado similar, lo que induce a cierta confusión cuando hablamos de ciencia. Por tanto, es necesario proceder a clarificarlos: cuando hablamos de «fiabilidad» en términos científicos, siempre hacemos referencia a la precisión de los instrumentos observacionales utilizados, de manera que, llegado el caso, se puede calcular la consistencia y estabilidad de las medidas tomadas. A mayor fiabilidad de un instrumento menor número de errores aleatorios e impredecibles, puesto que los pertinentes controles que se establecen tienden a excluirlos. Cuando hablamos de «validez», nos referimos a si el instrumento observacional mide correctamente el constructo que pretende medir. De este modo, pese a que son dos propiedades íntimamente relacionadas, hacen referencia a aspectos claramente diferenciados. La fiabilidad es una característica del instrumento, la validez se refiere a las generalizaciones que se hacen a partir de los resultados obtenidos a través del instrumento. Podría decirse, que la validez indica si el instrumento mide realmente el constructo que se está observando, mientras la fiabilidad se refiere a si lo mide de forma correcta, sin errores. Se establece una relación entre ambos conceptos de manera tal que la fiabilidad de un instrumento observacional influye en su validez; cuanto más fiable sea, mayor será también su validez y esta última, a su vez, nos informa indirectamente sobre la fiabilidad.

A lo largo de este escrito se va a realizar un análisis de las metodologías, de las búsquedas bibliográficas y de los procedimientos utilizados en la evidencia científica. Posteriormente, se llevará a cabo un análisis comparativo de los procedimientos utilizados en las ciencias de la salud que, de manera cada vez más habitual, se están basando en la evidencia científica en comparación con los métodos utilizados en las ciencias sociales. Para ello utilizaremos a modo de ejemplo la perspectiva de la gestión. Por último, analizaremos cuál es el impacto que comporta la utilización de estas metodologías describiendo las clasificaciones más utilizadas para valorar la evidencia en el ámbito de la salud, analizando sus principales diferencias y aplicaciones a la hora de que los posibles usuarios puedan elegir la que mejor se adapte a sus necesidades y, de ese modo, poder tomar decisiones en el terreno de la salud basándose en la mejor evidencia disponible.

Uno de los primeros en plantear la Evidencia Basada fue el epidemiólogo inglés, Archie Cochrane, quien generó la idea de base para el desarrollo de la actual Colaboración Cochrane; idea que no llegó a ver realizada. En 1972 publicó el libro *Effectiveness and Efficiency: Random Reflections on Health Services*[1], en el que planteaba la necesidad de recopilar y revisar de manera crítica los ensayos clínicos con el objetivo de que las decisiones médicas se fundamentaran en pruebas empíricas fiables. En 1992 se publicó en el *Journal of American Medical Association* un artículo firmado por el Evidence-Based Medicine Working Group (EBMWG) titulado «Evidence-based medicine. A new approach to teaching the practice of medicine» en el que se llamaba la atención sobre la necesidad de un «nuevo paradigma». Con este nuevo paradigma se pretendía cambiar la orientación del desarrollo de conocimiento basado en la intuición, en la experiencia clínica no sistemática y/o en el razonamiento fisiopatológico poniendo un mayor empeño en el análisis de las pruebas aportadas por la investigación.

Hasta ese momento, la toma de decisiones en el ámbito sanitario se había producido, casi de manera exclusiva, basándose en la experiencia de la persona que debía adoptarlas, esto es, recurriendo a soluciones conocidas para otros procesos relativamente similares donde se confiaba de manera específica en los conocimientos aprendidos durante el periodo formativo y durante la experiencia empírica. A este respecto conviene mencionar que, según diversas estimaciones, solo de un 15% a un 20% de las prácticas sanitarias se basan en una efectividad probada (Smith, 1991). En consecuencia, puede que estas revisiones no proporcionen la correspondiente información para que el investigador pueda extraer el sentido preciso de lo que allí se está diciendo o, en otras palabras, pueden inducir a cierto sesgo y falta de rigor en las interpretaciones. A esto hay que sumar el hecho de que se está haciendo más y más difícil extraerle sentido a una cada vez más abundante masa de estudios.

Desde el punto de vista científico, la generalización que se venía realizando a partir de experiencias individuales no sistemáticas procedentes de un limitado número de casos en los que los pacientes eran sometidos a unos procedimientos de utilidad incierta e incluso perjudicial era improcedente, sesgada e inaceptable. Aunque hasta este momento hemos utilizado la forma verbal del pretérito, desafortunadamente, en la actualidad, este enfoque de resolución de casos todavía está vigente y nos conduce a

[1] Traducción al castellano: Cochrane, A. L. (2000). *Efectividad y Eficiencia: Reflexiones al azar sobre los servicios sanitarios*. Barcelona: Centro Cochrane Iberoamericano.

preguntarnos sobre cómo proceder para poder distinguir lo que es eficaz de lo que no lo es, qué acciones reportan más beneficios y causan una menor acción iatrogénica. La evidencia basada es la idea que facilita la respuesta a estas cuestiones.

Un modelo paradigmático de incorporación de esta forma de pensar es la del Gobierno británico que, desde los años ochenta, ha puesto el énfasis en que tanto la práctica como las políticas se establezcan a partir del desarrollo de una evidencia basada más rigurosa. Para ello, planteó la iniciativa de las tres «es», Economía, Eficiencia y Efectividad[2], procediendo a la producción de guías detalladas y manuales sobre las mejores prácticas en numerosas disciplinas, todo ello orientado a la provisión de los mejores servicios públicos (David, Nutley and Smith, 2000).

Las actuaciones para fundamentar la práctica en la mejor evidencia disponible se han aplicado mayoritariamente en el ámbito de la salud, aunque en los últimos años también se han extendido a otras disciplinas. En este contexto, la preocupación por la efectividad tiene que ver con la validez y conveniencia de los métodos utilizados por los profesionales en su trabajo diario para la obtención de sus objetivos, así como por el conjunto de las diferentes entidades y agencias que proveen estos servicios. De esta manera, la preocupación por la efectividad en la provisión de servicios ha sido uno de los puntos de atención de los diferentes gobiernos ingleses, de manera que Tony Blair anunció en mayo de 1997 que «lo que cuenta es lo que funciona»[3].

El desarrollo y expansión de la Colaboración Cochrane coincidió con el desarrollo de internet, lo que permitió que las bases de datos, hasta entonces publicadas en CD, pudieran estar disponibles en la red a partir de 1998. Posteriormente, algunos miembros de diferentes departamentos de la Universidad canadiense de McMaster pertenecientes al EBMWG, en especial David Sackett, continuaron desarrollando esta idea y la presentaron más elaborada en 1997 en el libro *Evidence Based Medicine. How to Practice and Teach EBM*.

Más tarde, la Colaboración Cochrane ha extendido su actividad consiguiendo la expansión de esta información, en el caso de España, concretamente en el año 2003. Para dinamizar y extender las actividades de la Colaboración Cochrane se han ido constituyendo una serie de centros que

[2] En inglés: Economy, Efficiency and Effectiveness)

[3] En inglés: «what counts is what works». Tony Blair, Labour Party Manifesto for the 1997 General Election.

cubren diferentes áreas geográficas y culturales. A fecha de hoy hay 12 centros activos; en lo que a nuestro entorno se refiere, está el Centro Cochrane Iberoamericano, que se constituyó en 2000 a modo de ampliación de las funciones del Spanish Cochrane Centre nacido en 1997.

La estrategia ha consistido, no solamente en incrementar la calidad de la investigación que se viene realizando, sino también en proceder a realizar revisiones sistemáticas de la investigación ya realizada por medio de evaluaciones de la mejor evidencia disponible que presentan y ubican los hallazgos de manera que sean relevantes para los que toman las decisiones.

Las revisiones sistemáticas difieren de las tradicionales búsquedas bibliográficas en que las primeras se realizan siguiendo criterios de transparencia y replicabilidad científica con el objeto de minimizar el sesgo a través de la búsqueda bibliográfica exhaustiva de estudios publicados y sin publicar, y llevando a cabo un seguimiento de las distintas evaluaciones realizadas, de los procedimientos y de las conclusiones (Cook, Mulrow y Haynes, 1997). El proceso de revisiones sistemáticas y su procedimiento asociado al metanálisis se ha desarrollado a lo largo de los últimos veinte años siendo reconocido como una pieza clave en las actividades basadas en la evidencia.

En el año 2000, en la ciudad de Filadelfia (EE. UU.), en el ámbito de las ciencias sociales, se puso en marcha la Colaboración Campbell, que por aquel entonces reunió alrededor de unos 150 científicos relacionados con las ciencias sociales, y cuya pretensión era el desarrollo de una estructura similar a la de la Colaboración Cochrane en lo relativo a las ciencias de la salud con el objetivo de ayudar a los investigadores a tomar decisiones informadas acerca de los efectos de las intervenciones en los campos sociales, conductuales y educativos (Davies y Boruch, 2001).

En el ámbito de las ciencias sociales existen dificultades que lastran su desarrollo. Por una parte, se da una baja aceptación de las metodologías de investigación apropiadas que deberían utilizarse en la evaluación de la evidencia basada, y por otra, se da una menor aceptación de cómo la evidencia basada debe de informar las políticas y la práctica (Macdonald, 1999). Ambas cuestiones quedan reflejadas en el trabajo de Davies, Nutley y Smith (2000), quienes plantean que los diferentes puntos de partida a nivel epistemológico y ontológico, y las diferentes tradiciones profesionales afectan a los métodos y el entusiasmo con que los profesionales se embarcan en los procedimientos de evidencia. A esto añaden de manera más o menos optimista que observan la existencia de un vasto potencial en términos de investigación de la evidencia y apuntan que este potencial

puede ser más influyente de lo que ha sido hasta ahora. Un análisis pormenorizado de cada una de las ciencias ha llevado a autores tales como Whitley (1984) a identificar, por ejemplo, la situación científica de la gestión como una ciencia con orientación social y considerarla una ciencia fragmentada con un bajo nivel de control sobre estándares significativos, lo que supone que la significación de los problemas y las formas de resolverlos son «inestables, sometidos a disputas, y evaluados por diversos y difusos estándares» (1984: 343). Algo que puede conducir a una divergencia entre los investigadores y el resto de actores del proceso de investigación y, por tanto, a una proliferación de teorías y practicas irrelevantes (Anderson, Herriot y Hodgkinson, 2001). De acuerdo con estas posiciones, resulta necesario considerar los trabajos de Gibbons *et al*. (1994), relacionados con los modos de producción del conocimiento, como trabajos verdaderamente influyentes en el debate, ya que, según su consideración del modo de producción del conocimiento, se produce un constante flujo de ideas y aportaciones entre la práctica y la teoría, en el seno del cual, «se produce conocimiento en el contexto de aplicación»; lo que se ha convertido en el centro del debate acerca del futuro de la gestión.

Visto lo visto, si procedemos a realizar la comparación entre la investigación y la aplicación de la evidencia en el ámbito de la gestión, y la investigación y la aplicación de la evidencia en el ámbito de la salud, vemos que en este último se ha utilizado tradicionalmente un enfoque de aleatorización y de estudios de doble ciego, que suelen considerarse como los métodos más rigurosos para probar las nuevas intervenciones antes de generalizar sus usos.

Por otra parte, la investigación en el ámbito de la gestión es bastante reciente, y está mucho menos desarrollada en términos de programas y actuaciones que en el ámbito clínico, así como desde el punto de vista de los contenidos e hipótesis de investigación. Esta situación hace bastante difícil que se pueda realizar la agregación de datos para llevar a cabo investigaciones de síntesis, tales como aplicación de la metodología del metanálisis. La heterogeneidad de estudios imposibilita amalgamar los resultados obtenidos y, por tanto, evaluar la efectividad de las intervenciones. Esta situación da lugar a la pregunta de hasta qué punto los procesos de revisión en el ámbito de las ciencias de la salud, que están procediendo a transformarse en ciencias basadas en la evidencia, podrían proporcionar ideas al ámbito de las ciencias sociales y, entre ellas, al campo de la gestión para proceder a crear revisiones que sean relevantes y rigurosas.

9.2. Proceso para la realización de una revisión sistemática y metanálisis

Las revisiones sistemáticas de la literatura biomédica definen todos los procedimientos detalladamente, establecen criterios claros para delimitar cómo se recuperará la información que se va a evaluar, cómo se diseñan los estudios que se realizarán, en qué poblaciones, en qué ámbitos geográficos, que términos clave se utilizarán para la búsqueda o qué bases documentales se utilizarán.

El primer paso en este proceso consiste en formular la pregunta a la que se pretende dar respuesta y las razones para ello, que como en cualquier proceso de investigación deberían establecerse de la forma más clara y concisa posible definiendo los criterios de elección de los estudios para incluirlos. En el caso sanitario, la pregunta debe ser específica y estar estructurada de acuerdo al acrónimo PICO; Esto es, se debe delimitar el grupo de población (P) al que se refiere, cuál es la intervención (I) o procedimiento de interés del que se trata, con qué otro procedimiento alternativo de control (C) se compara y cuáles son los resultados (O, de *outcome* en inglés) en términos de salud a los que se presta atención para valorarlo, incluyendo los detalles del diseño del estudio.

Es necesario proceder a identificar y asignar el escenario clínico, ya sea en materia de prevención, etiología, diagnóstico diferencial, pronóstico, historia natural, prevalencia, daño, tratamiento, tamizaje, estudios económicos o análisis de decisión entre otros. Una vez identificado el escenario en el que corresponde catalogar el artículo (en ocasiones, puede ser asignado a más de uno), según el tipo de diseño del estudio en cuestión, se aplica la propuesta de niveles de evidencia y grados de recomendación utilizando alguna de las clasificaciones existentes o pautas de lectura crítica correspondiente con la que se evaluará cada capítulo (algunas de ellas serán objeto de análisis más adelante en este escrito). Es habitual que los estudios que se analicen no sean homogéneos, lo que exige que en su momento se planteen hipótesis de exploración y explicación de la heterogeneidad de los estudios.

El paso siguiente es clave en el proceso de revisiones sistemáticas. Se trata de la exploración y recuperación de la evidencia mediante una búsqueda exhaustiva, objetiva y reproducible de los trabajos originales sobre el tema que se genere de acuerdo a unos criterios prefijados y que se evaluará críticamente en pasos subsiguientes. Por tanto, será una búsqueda que habrá de llevase a cabo siguiendo detalladamente un protocolo. Se trata de una

búsqueda sistematizada en la que se utilizan los criterios establecidos previamente en la pregunta. Es imprescindible acudir a bases de datos de bibliografía que permitan definir búsquedas específicas mediante términos clave combinados con operadores booleanos. También deben permitir acotar el ámbito temporal de interés, el tipo de estudios, la población, etc.

Según el tema de interés, será necesario consultar una o más bases de datos bibliográficas, siendo las base de datos de consulta más habitual en ciencias de la salud *Medline*, la base de datos de la National Library of Medicine iniciada en 1960, a la que se puede acceder para realizar búsquedas de diferentes maneras, aunque las más habituales son a través de OVID o PubMed, disponible desde 1996 (www.ncbi.nlm.nih.gov/pubmed). Otras bases de datos bibliográficas útiles que se consultan con frecuencia son: Embase (www.embase.com/); PsycINFO (www.apa.org/pubs/databases/psycinfo/index.aspx); Web of Science (http://science.thomsonreuters.com/); o LILACS (http://lilacs.bvsalud), que es una parte de la Biblioteca Virtual de Salud (BVS) y una iniciativa impulsada por la Biblioteca Regional de Medicina (BIREME) perteneciente al Centro Latinoamericano y del Caribe de Información sobre Ciencias de la Salud, un centro especializado de la Organización Panamericana de la Salud.

Del mismo modo, es necesario consultar bases de datos específicas de revisiones sistemáticas de la evidencia científica como la biblioteca Cochrane, siendo frecuente que haya que acudir a fuentes en las que se pueden recuperar informes y documentos técnicos de diferentes organismos, instituciones y administraciones públicas, e incluso a tesis doctorales, trabajos no publicados o publicados en revistas no indexadas, etc., denominados en este caso literatura gris. Habitualmente, este tipo de documentos no están incluidos en las bases de datos bibliográficas de literatura biomédica. Desde 1997, para la identificación de este tipo de informes y documentos se pueden consultar bases de datos como OpenGrey (http://www.opengrey.eu/) –una base de datos multidisciplinar europea que compila documentos de los principales centros de documentación e información europeos adheridos a la European Association for Grey Literature Exploitation (EAGLE) y que cubre temas de ciencia, tecnología, biomedicina, economia, ciencias sociales y humanidades, clasificados de acuerdo al System for Information on Grey Literature in Europe (SIGLE)–. La exhaustividad y el rigor de la búsqueda bibliográfica determinará en gran medida la calidad y validez final de las revisiones sistemáticas y los metanálisis.

Una vez comenzada la búsqueda, resulta necesario seleccionar los títulos y resúmenes de los artículos para determinar su inclusión en el seguimiento

de la lista de criterios de inclusión y exclusión establecidos, y que deberá de ser lo más objetiva posible y aplicarse rigurosamente de forma ciega e independiente por varios evaluadores en aras de evitar el denominado sesgo de selección. En muchos casos, se tiende a utilizar la calidad metodológica de los trabajos como un criterio de inclusión, siendo aconsejable utilizar este criterio como una variable más a tener en cuenta junto con otras como son el diseño de los trabajos, el tamaño de la muestra, lo exhaustivo de la información que presentan, las intervenciones y las respuestas estudiadas, etc.

Una vez realizada la selección, hay que obtener el texto completo de los estudios que han cumplido los criterios de inclusión para su completa revisión mediante la utilización de criterios explícitos de evaluación de la calidad. Esto proporciona la transparencia necesaria a la revisión. Obtenidos los textos, se lleva a cabo la revisión crítica comenzando por evaluar la posibilidad de sesgos mediante el análisis de los resultados utilizando métodos validados que permitan valorar apropiadamente la medida del efecto. Los estudios han de priorizarse en función de su relevancia clínica, lo que proporciona un enfoque práctico en términos de aplicación. Llegado este punto, es necesario resumir los datos antes de proceder a su presentación.

En caso de que la realización de un metanálisis sea posible, resulta necesario generar un conjunto de estimaciones e intervalos de confianza buscando explicaciones para los aspectos relativos a la heterogeneidad de los estudios y valorar la confianza y las estimaciones obtenidas por los estudios mediante análisis de sensibilidad, que permita estudiar la influencia individual de cada estudio en el resultado del metanálisis. Esto es, que permita determinar si los resultados pueden verse sesgados por estudios con escasa calidad metodológica, trabajos no publicados o que no cumplan estrictamente los criterios de selección, etc. Ello consiste en replicar el metanálisis quitando uno de los estudios incluidos en cada paso para ver si se obtienen o no resultados similares de forma global. Esto permite evaluar su aplicabilidad en diferentes contextos clínicos (riesgo basal). Por último, hay que mantener las revisiones permanentemente actualizadas incorporando la nueva evidencia que vaya surgiendo.

9.3. Análisis crítico de las evidencias

Las primeras interpretaciones y consideraciones respecto a la aplicación de metodologías de búsqueda de la evidencia se llevaron a cabo en los años 90. Primero se centraron en el diseño de los estudios, para luego evolucionar y

acabar llegando al primer principio de la práctica clínica en evidencia que sostiene que existe una jerarquía de pruebas. Todo ello, desde la perspectiva, por ejemplo, que considera que los ensayos clínicos controlados aleatorios son el estándar oro de la investigación, ya que incorporan salvaguardas metodológicas que reducen el riesgo de sesgo en comparación con los estudios observacionales. Tales planteamientos incitan a pensar en una estructura conceptual jerarquizada que induce la idea de una pirámide, clara expresión de la representación de una jerarquía. Una representación que devino en natural para los profesionales de la salud que se basarían en la evidencia y que se familiarizaron rápidamente con esta pirámide al leer la literatura, a la hora de aplicar pruebas o transmitir esta idea a los estudiantes.

Hasta el momento se han descrito diferentes versiones de la pirámide de evidencias, tradicionalmente han tendido a centrarse en mostrar los estudios con características más endebles en la base de la pirámide, normalmente series de casos, seguidos de los estudios de cohortes controlados en medio, a continuación, los ensayos de control aleatorios y, en la cúspide, las revisiones sistemáticas y metanálisis (figura 9.1).

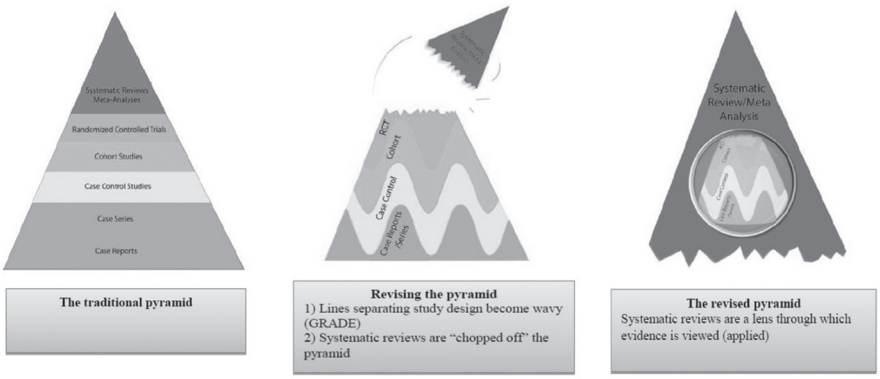

Figura 9.1. Evolución de la representación de la jerarquía de la evidencia científica (Tomado de Murad *et al.,* 2015).

En muchos casos, esta descripción es intuitiva y probablemente correcta. Representa de una manera clara la validez y jerarquía de los estudios en términos de riesgo de sesgo. A principios del siglo XXI, el grupo de trabajo Grading of Recommendations Assessment, Development and Eva-

luation (GRADE) presentó un marco en el que la certidumbre en cuestiones de evidencia se basaba en una combinación en la que se consideraba el diseño de los estudios y otros factores como, por ejemplo, la imprecisión o la incoherencia que desafiaban al concepto de pirámide (Guyatt *et al*., 2008). En 2014, la *User's Guide on systematic reviews and meta-analysis* procedió a cuestionar la ubicación de las revisiones sistemáticas en la parte superior de la pirámide y presentó un enfoque de dos pasos en el que se evaluaba, primero la credibilidad del proceso utilizado para generar una revisión sistemática y segundo, la evaluación de la certidumbre en evidencia basada en el enfoque de GRADE (Murad et al., 2014). En la figura 9.1, que muestra la evolución de la representación de la jerarquía de la evidencia científica, en la primera pirámide, encontramos la jerarquía de la evidencia científica en su momento inicial: en la base se sitúan los estudios que proporcionan evidencia de menor nivel; en el resto de niveles hacia el vértice se sitúan los diseños de estudios que proporcionan evidencia científica progresivamente de mejor calidad; y el vértice lo ocupan las revisiones sistemáticas y metanálisis. En la segunda pirámide de la figura 9.1 se pueden apreciar los cambios producidos y la evolución que se planteó, especialmente en lo relativo a dos conceptos:

1. Se eliminaron las líneas rígidas que separaban los diferentes niveles en los que se podía apreciar el diseño de los estudios que pasaron a ser representadas como líneas onduladas que suben y bajan para de esa manera reflejar el enfoque GRADE de valoración que incluye niveles; y
2. Se retiraron del ápice de la pirámide las revisiones sistemáticas y el metanálisis y fueron transformadas en una lente de observación a través de la cual se observan otros estudios en términos de aplicación y evaluación.

Tanto en la enseñanza como en la práctica clínica basadas en la evidencia se pueden realizar comparaciones entre las pirámides originales y las revisadas usando la imagen de la pirámide para explicar el proceso de acercamiento al enfoque GRADE y la *User's Guide to systematic reviews* y así elaborar consideraciones acerca de la evolución del pensamiento basado en la evidencia en el ámbito de la práctica clínica y alcanzar una moderna comprensión de la certeza en la evidencia (Murad, 2015).

Una parte crucial de las revisiones sistemáticas de la evidencia científica consiste en el análisis crítico de los estudios recopilados atendiendo al

diseño de los distintos estudios, a la metodología utilizada en todos sus apartados, a posibles sesgos y a los errores de medición, el tamaño de la población estudiada, cómo se seleccionó, si existe un buen grupo de comparación, si el análisis de los datos es adecuado y si hay control de posibles factores de confusión. La validez interna del estudio refleja en qué medida las conclusiones del estudio se corresponden con la realidad; la validez externa refleja el grado en que las conclusiones del estudio son aplicables al universo exterior al estudio.

9.4. Clasificaciones de evidencia y grados de recomendación

El grupo de Sackett, ubicado en la Universidad McMaster, desarrolló los principios de la enseñanza y práctica de la aplicación de metodologías de búsqueda de la evidencia que han mantenido su vigencia hasta el momento actual y pueden ser resumidos como «integrar la experiencia clínica individual con la mejor evidencia disponible a partir de la investigación sistemática» (EBWG, 1992; Manterola, 2002). A partir de ese momento, surgieron diferentes grupos de trabajo integrados por profesionales de la salud y científicos independientes que se han dedicado a discutir aspectos de la metodología y a elaborar protocolos y otras herramientas de soporte para la realización de revisiones sistemáticas, especialmente, para el análisis crítico y decisiones sobre la calidad y nivel de la evidencia. Se pueden mencionar, entre otros, la colaboración Cochrane (http://www.cochrane.org/), el grupo de trabajo GRADE (http://www.gradeworkinggroup.org/_es/index.htm), surgido ante el incremento en el número de sistemas de clasificación de la evidencia que se ha ido produciendo y que trata de construir y validar una clasificación sencilla que integre todos los aspectos importantes a tener en cuenta (Maro-Castillejo & Zulaica, 2007; Manterola, et. al, 2014) y la red *Critical Appraisal Skills Programme,* conocida en español por su acrónimo CASPe (http://www.redcaspe.org/), que desarrolla un Programa de Habilidades en Lectura Crítica en español y enseña a personas con diferentes perfiles (clínicos, investigadores, gestores y pacientes) pertenecientes a instituciones u organizaciones del sistema de salud mediante talleres presenciales y/o virtuales. Asimismo, enseña habilidades para la docencia de la lectura crítica en sus entrenamientos dirigidos a entrenadores.

En función del rigor científico del diseño de los estudios pueden construirse escalas de clasificación jerárquica de la evidencia a partir de las

cuales pueden establecerse recomendaciones respecto a la adopción de un determinado procedimiento o intervención sanitaria (Jovell y Navarro, 1995; Guyatt et al., 1995). En general, las clasificaciones se basan en el análisis de 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. Al haberse dado una proliferación de propuestas y clasificaciones que jerarquizan la evidencia y, por tanto, sus respectivos grados de recomendación, puede aparecer cierta confusión a la hora de generar la evidencia. Efecto que puede deberse, tanto al lenguaje que se utiliza para expresar la información –un lenguaje no habitual en la práctica (Manterola y Zabando, 2009)– como a la complejidad de las escalas de valoración y las diferencias entre ellas, lo que implica confusión e incertidumbre respecto a cuál de ellas aplicar.

El nivel de evidencia se clasificará según el grado de consistencia entre los resultados de los estudios analizados y el tipo de estudios de que se trate. Para ello se han propuesto distintos esquemas y baremos de clasificación del nivel de evidencia y del grado de recomendación teniendo en cuenta el número de estudios con resultados consistentes, contradictorios o neutros, su diseño, su calidad e incluso la evaluación económica de las intervenciones investigadas. A este respecto, aunque existen diferentes escalas de gradación de la calidad de la evidencia científica, encontramos una gran similitud entre ellas. Estas escalas evalúan la calidad de la evidencia de una forma elaborada, considerando el tipo de diseño de los estudios, los niveles de evidencia y los grados de recomendación para sujetos asintomáticos indicando los procedimientos más adecuados y aquellos que deberían evitarse. Los grados de recomendación se establecen a partir de la calidad de la evidencia y del beneficio neto (beneficios menos perjuicios) de la medida evaluada. Además, en ella se realizan análisis de coste-efectividad (Saha et al. 2001). A continuación pasaremos a describir algunas de estas escalas cronológicamente.

9.4.1. Canadian Task Force on Preventive Health Care

La primera de ellas fue formulada en 1979 por la *Canadian Task Force on the Periodic Health Examination* (CTFPHC, 1979) y tenía por objeto la evaluación de medidas preventivas. Se adoptó en 1984 por la *United States Preventive Services Task Force* (2017) y, posteriormente, fue publicada por Harris *et al*. (2001) en su tercera edición, que puede consultarse en la web

de la *Agency for Healthcare Research and Quality*. *La Canadian Task Force on Preventive Health Care* desarrolló para uso de la *Public Health Agency of Canada* (PHAC) los grados de recomendación para las intervenciones de prevención que se incluirían en las guías de práctica clínica elaboradas por este organismo con el fin de que respaldasen las acciones de salud preventiva (Birtwhistle, et al., 2012) poniendo el énfasis en los tipos de diseño utilizados y en la calidad de los estudios publicados. Así se elaboró un orden alfabético para los Grados de Recomendación (tabla 9.1), en donde las letras A y B indicaban que existía una buena evidencia para recomendar la intervención y, por tanto, se recomendaba llevarla a cabo; la letra C indicaba que la evidencia era contradictoria, lo cual no permitía realizar recomendaciones a favor o en contra, aunque podían tenerse en cuenta otros elementos a la hora de tomar decisiones; las letras D y E indicaban que no debían llevarse a cabo las intervenciones clínicas en razón de la evidencia moderada o en contra de realizar la recomendación o de no realizarla; y finalmente, la letra I, indicaba una evidencia insuficiente, tanto en calidad como en cantidad, para hacer una recomendación. Sin embargo, otros factores podrían afectar a la decisión (CTFPHC, 2003).

Tabla 9.1. *Grados de recomendación para las intervenciones de prevención (CTFPHC)*

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 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 cuantitativa) para hacer una recomendación; sin embargo, otros factores podrían influenciar la decisión.

La CTFPHC procedió a clasificar el diseño de los estudios según niveles de evidencia e interpretación de los tipos de estudio para intervenciones de prevención estableciendo una clasificación de I a III, en donde el nivel I representaba la mayor calidad y el III la menor (tabla 9.2).

Tabla 9.2. *Niveles de evidencia e interpretación de los tipos de estudio para intervenciones de prevención (CTFPHC)*

Niveles de evidencia	Interpretación
I	La evidencia existente surge a partir de ensayos clínicos CON asignación aleatoria.
II-1	La evidencia existente surge a partir de ensayos clínicos SIN asignación aleatoria.
II-2	La 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	La 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	La evidencia existente surge a partir de la opinión de expertos basada en la experiencia clínica; estudios descriptivos o informes de comités de expertos.

Todo ello para pasar a clasificar los estudios según su validez interna o su calidad metodológica (tabla 9.3) con posterioridad.

Tabla 9.3. *Validez interna e interpretación de los tipos de estudio para intervenciones de prevención (CTFPHC)*

Validez interna	Interpretación
Buena	Un estudio, incluidas revisiones sistemáticas y metanálisis, que cumple los criterios específicos de un estudio bien diseñado.
Moderada	Un estudio, incluidas revisiones sistemáticas y metanálisis, 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.

Continúa

Tabla 9.3. *Continuación*

Validez interna	Interpretación
Insuficiente	Un estudio, incluidas revisiones sistemáticas y metanálisis, que tiene al menos un defecto metodológico grave en su diseño 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.

Adicionalmente, CTFPHC se apoya en el sistema GRADE, para evaluar la calidad de la evidencia y realizar recomendaciones en el ámbito de la prevención (Guyatt et al., 2008).

9.4.2. Clasificación de Sackett

Esta sistematización propuesta por Sackett (1989) jerarquiza la evidencia, tal y como se puede apreciar en la tabla 9.4, en niveles que van de 1 a 5; siendo el nivel 1 la mejor evidencia y el nivel 5 la peor, o si se prefiere la menos buena.

Tabla 9.4. *Clasificación de los niveles de evidencia según Sackett*

GR	NE	Terapia, prevención, etiología y daño	Pronóstico	Diagnóstico	Estudios económicos
A	1a	RS de EC con AA	RS con homogeneidad y metanálisis de estudios de cohortes concurrentes	RS de estudios diagnósticos 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

Continúa

Tabla 9.4. *Continuación*

GR	NE	Terapia, prevención, etiología y daño	Pronóstico	Diagnóstico	Estudios económicos
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 nivel mayor que 1
	2b	Estudios de cohortes individuales. EC de baja calidad	Estudios individuales 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 costos. Incluye un análisis de sensibilidad
	3a	RS con homogeneidad de estudios de casos y controles			
	3b	Estudios 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 de 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 fisiología o en investigación económica

AA: Asignación aleatoria.

La clasificación de los niveles de evidencia según Sackett se basa en dos escalas, la de grados de recomendación y la de las recomendaciones en apoyo de una intervención que pueden ser generadas en base a cinco niveles de evidencia. En el grado de recomendación A, las conclusiones se generan a partir de las evidencias más potentes en la investigación (1a, 1b, 1c) y, por tanto, son más contundentes. En el grado B, las conclusiones se basan en pruebas más débiles y solo son orientativas. Van disminuyendo sucesivamente en función del poder de los diseños de investigación (2a, 2b, 3a, 3b). En el grado C, las conclusiones se basan en pruebas débiles, por lo que son las menos fiables, y así sucesivamente. Por tanto, el nivel indica el grado de certeza generado por la fuerza de la evidencia, de manera que nos encontramos con la desventaja, tal y como ocurre en otras clasificaciones, de que al no existir subcategorías, en algunas situaciones resulta difícil proponer un grado de recomendación. Esta clasificación ha sido precursora de las actuales y ha servido como base para el desarrollo de clasificaciones que han tratado de evitar sus problemas.

9.4.3. GRADE Working Group

Precisamente el amplio número de sistemas de clasificación de la evidencia dio lugar a la creación del grupo de trabajo GRADE, que se desarrolló sobre la base de la experiencia previa con otras herramientas existentes y cuyo fin era intentar obtener un sistema para construir y validar una clasificación sencilla que integrara todos los aspectos importantes a tener en cuenta (Maro-Castillejo y Zulaica, 2007; Manterola, Asenjo-Lobos y Otzen, 2014).

GRADE es un sistema para clasificar la calidad de la evidencia y la fuerza de la recomendación aplicable a una amplia gama de intervenciones y contextos que intenta ofrecer un sistema estructurado, transparente y explícito; «más razonable, confiable y ampliamente aplicable» (Canfield & Dahm, 2012). Su diferencia respecto de otros es que GRADE no valora la calidad de los estudios individuales; 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 equilibrio entre beneficios y riesgos, la calidad de la evidencia, la aplicación de esta en circunstancias específicas y la situación de riesgo basal, que son los puntos clave evaluados en cada artículo (tabla 9.5).

Tabla 9.5. *GRADE modificado: grados de recomendación*

Grado de recomendación-descripción	Beneficio vs. Riesgo y cargas	Calidad metodológica que apoya la evidencia	Implicaciones
1A Recomendación fuerte, evidencia de alta calidad	Los beneficios superan claramente los riesgos y cargas o viceversa	Ensayo clínico 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	Ensayo clínico 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 más evidencia
2A Recomendación débil, evidencia de alta calidad	Beneficios estrechamente equilibrados con los riesgos y la carga	Ensayo clínico 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

Continúa

Tabla 9.5. *Continuación*

Grado de recomendación-descripción	Beneficio vs. Riesgo y cargas	Calidad metodológica que apoya la evidencia	Implicaciones
2B Recomendación débil, evidencia de moderada calidad	Beneficios estrechamente equilibrados con los riesgos y la carga	Ensayo clínico 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, riesgo y la carga pueden estar estrechamente equilibrados	Estudios observacionales o series de casos	Recomendaciones muy débiles, otras alternativas pueden ser igualmente razonables

En esta propuesta destaca 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 que permite desarrollar perfiles de evidencia y resúmenes de los hallazgos (Guyatt et al., 2011). Las ventajas del sistema GRADE respecto de otros sistemas de clasificación son que este utiliza definiciones explícitas y juicios secuenciales durante el proceso de clasificación, proporcionando una descripción detallada de los criterios para la calidad de la evidencia, para los resultados individuales y para la calidad general de la evidencia al sopesar la importancia relativa de los resultados tomado en consideración el equilibrio entre los beneficios de salud versus los perjuicios, costos y gastos, lo que per-

mite desarrollar perfiles de evidencia y resúmenes de los hallazgos (Guyatt et al., 2006).

Es una clasificación muy completa que realiza la extracción de datos y la síntesis mediante un software de uso libre que requiere bastante tiempo a la hora de aplicarlo puesto que el análisis de los estudios es exhaustivo. Esta clasificación ha sido incorporada por diversas instituciones internacionales y nacionales de salud, de diferentes países, para evaluar la calidad de la evidencia disponible, realizar recomendaciones y generar guías de práctica clínica.

9.4.4. U.S. Preventive Services Task Force

El *Unites States Preventive Services Task Force* (USPSTF), creado en 1984, es un grupo independiente de expertos en prevención y metodologías de búsqueda de la evidencia. Entre sus actividades se encuentra la valoración de la investigación clínica que se realiza con el fin de evaluar las medidas preventivas, las pruebas de detección o *screening*, los servicios de asesoramiento, las vacunas y diferentes usos de los medicamentos (Harris et al., 2001). Este grupo procedió a generar una jerarquización que asignaba niveles de certeza para evaluar el beneficio neto de un servicio preventivo basándose en la naturaleza de la evidencia total disponible para sustentar el grado de recomendación (tabla 9.6). Estableció la fuerza de sus recomendaciones a partir de la calidad de la evidencia y del beneficio neto, que fue definido como beneficio menos daño del servicio preventivo evaluado tal como se aplicaba en la atención primaria a la población general.

Tabla 9.6. *Descripción de los niveles de evidencia para exámenes periódicos de salud (USPSTF)*

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 poco probable que sea fuertemente afectada por los resultados de futuros estudios.

Continúa

Tabla 9.6. *Continuación*

Niveles de certeza	Descripción
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 en 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 puede cambiar, y este cambio puede ser lo suficientemente importante como para alterar la conclusión.

Los grados de recomendación (A, B, C, D, o I) en la tabla 9.7 se apoyan en grados de certeza que se definen como la probabilidad de que el beneficio neto de un servicio preventivo que ha sido evaluado por la USPSTF sea correcto.

Tabla 9.7. *Grados de recomendación para exámenes periódicos de salud (USPSTF)*

Recomendación	Interpretación	Sugerencias para la práctica
A	Se recomienda la acción preventiva. Existe alta certeza de que el beneficio neto es substancial.	Ofrecer o proporcionar este servicio.
B	Se recomienda la acción preventiva. Hay alta certeza de que el beneficio neto es moderado o existe moderada certeza de que el beneficio neto es moderado a substancial.	Ofrecer o proporcionar este servicio.

Continúa

Tabla 9.7. *Continuación*

Recomendación	Interpretación	Sugerencias para la práctica
C	Se recomienda selectivamente el ofrecimiento o la prestación de este servicio a los pacientes individuales basándose en criterios profesionales y las preferencias del paciente. Hay por lo menos moderada certeza de que el beneficio neto es pequeño.	Ofrecer o proporcionar este servicio a los pacientes seleccionados en función de las circunstancias individuales.
D	NO se recomienda la acción preventiva. Hay certeza moderada o alta de que el servicio no tendrá ningún beneficio neto o que los daños serán mayores que los beneficios.	Desalentar el uso de este servicio.
I	Se concluye que la evidencia actual es insuficiente para evaluar la relación entre los beneficios y los perjuicios de la acción preventiva. La evidencia es deficiente, de mala calidad o contradictoria y no se puede determinar la relación riesgo-beneficio.	Léase 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.

El grado A sugiere recomendar la acción, ya que existe un alto grado de certeza de que el beneficio neto es substancial; 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 tal como se puede apreciar en la tabla 9.7; ahora incluye una columna de sugerencias para la práctica asociadas con cada grado.

9.4.5. Scottish Intercollegiate Guidelines Network

El Servicio Nacional de Salud (NHS) escocés desarrolló la *Scottish Intercollegiate Guidelines Network* (SIGN) utilizando una metodología explícita

que se basa en tres principios básicos: a) identificar y evaluar críticamente la evidencia disponible diseñadas como un vehículo para acelerar la traducción del nuevo conocimiento en acción; b) reducir la variabilidad de la práctica; y c) mejorar los resultados relevantes para los pacientes (SIGN, 2018). Anteriormente, se basaban en los niveles de evidencia desarrollados por la *American Healthcare Policy Research* (AHCPR) de los Estados Unidos. Dadas las limitaciones encontradas con este sistema, se desarrolló una nueva clasificación que se utiliza desde el año 2000 (Harbour y Miller, 2001) (tablas 9.8 y 9.9).

Tabla 9.8. *Niveles de evidencia (SIGN)*

Nivel de evidencia	Tipo de estudio
1++	Metanálisis de gran calidad, revisiones sistemáticas de ensayos clínicos aleatorizados o ensayos clínicos aleatorizados con muy bajo riesgo de sesgos.
1+	Metanálisis bien realizados, revisiones sistemáticas de ensayos clínicos aleatorizados o ensayos clínicos aleatorizados con bajo riesgo de sesgos.
1-	Metanálisis, revisiones sistemáticas de ensayos clínicos aleatorizados o ensayos clínicos aleatorizados con alto riesgo de sesgos.
2++	Revisiones sistemáticas 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-controles con alto riesgo de confusión, sesgos o azar y una significante probabilidad de que la relación no sea causal.
3	Estudios no analíticos (observaciones clínicas y series de casos).
4	Opiniones de expertos.

La propuesta del SIGN se originó estableciendo el foco de interés en el tratamiento y los procedimientos terapéuticos. Se diferencia de las anteriores por su particular énfasis en el análisis cuantitativo que aportan las

revisiones sistemáticas; y otorga además importancia a la reducción del error sistemático.

Tabla 9.9. *Grados de recomendación (SIGN) (9)*

Grado de recomendación	Nivel de evidencia
A	Al menos un metanálisis, revisión sistemática o ensayo clínico aleatorizado calificado como 1++ y directamente aplicable a la población objeto, o una revisión sistemática de ensayos clínicos aleatorizados o un cuerpo de evidencia consistente principalmente en estudios calificados como 1+ directamente aplicables a la población objeto y que demuestren globalmente consistencia de los resultados.
B	Un cuerpo de evidencia que incluya estudios calificados como 2++ 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 cuerpo de evidencia que incluya estudios calificados como 2+ directamente aplicables a la población objeto y que demuestren globalmente consistencia de los resultados, o extrapolación de estudios calificados como 2++.
D	Niveles de evidencia 3 o 4, o extrapolación de estudios calificados como 2+. Niveles de evidencia y grados de recomendación.

Este sistema de clasificación otorga un mayor peso a la calidad de la evidencia que respalda cada recomendación y hace hincapié en que la evidencia debe ser considerada en su conjunto y que no dependa de un solo estudio para apoyar cada recomendación. También reconoce que hay aspectos en los que se pueden realizar, ya sea por razones prácticas o éticas, ensayos clínicos. Para ello está previsto conceder más relevancia a las recomendaciones respaldadas por estudios observacionales de gran calidad. En 2009, SIGN tomó la decisión de implementar el enfoque GRADE como directriz, metodología que se encuentra actualmente en desarrollo (SIGN, 2011).

9.4.6. Centre for Evidence-Based Medicine de Oxford

El Centre for Evidence-Based Medicine de Oxford (CEBM) procede a valorar la evidencia según el área temática o escenario clínico y según el tipo de estudio que involucra al problema clínico en cuestión (CEBM, 2009). Plantea una acción complementaria de otras iniciativas. Gradúa la evidencia de acuerdo con el mejor diseño para cada escenario clínico, otorgándole la intencionalidad y agregando las revisiones sistemáticas en los distintos ámbitos. Su ventaja se basa en que asegura la utilización del conocimiento que concierne a cada escenario por su alto grado de especialización. Adicionalmente, contribuye a aclarar cómo afecta la ausencia de rigurosidad metodológica al diseño de los estudios, disminuyendo su valoración en el grado de evidencia y en la fuerza de la recomendación (tabla 9.10).

Tabla 9.10. *Grados de recomendación 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

9.4.7. National Health and Medical Research Council

El National Health and Medical Research Council (NHMRC) 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 que se ha venido utilizando en Australia desde 1999 (tabla 9.11).

La tabla actual, que está basada en la jerarquización del CEBM, es más amplia que la inicial y está estructurada de forma distinta, por lo que a menudo la evidencia obtenida no es susceptible de ser sometida a metanálisis y, por tanto, su valoración se relaciona solo con los estudios individuales. Contiene 5 columnas para cada una de las áreas de investigación (intervención, precisión diagnóstica, pronóstico, etiología y tamizaje) y la

Tabla 9.11. *Niveles de evidencia según NHMRC*

Nivel	Intervención	Precisión diagnóstica	Pronóstico	Etiología	Tamizaje
I	RS estudios de nivel II	RS estudios de nivel II	RS estudios de nivel II	RS estudios de nivel II	RS estudios de 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 no 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 consecutivos con una presentación clínica definida	Todo o ninguno	Todo o ninguno	EC pseudoaleatorizado controlado (ej. Asignación alternada o algún otro método)

Continúa

Tabla 9.11. *Continuación*

Nivel	Intervención	Precisión diagnóstica	Pronóstico	Etiología	Tamizaje
III-2	Estudios comparativos con controles concurrentes: EC experimental sin AA Estudios de cohortes Estudios de casos y controles Series temporales interrumpidas con un grupo control	Comparación con un estándar de referencia que no cumple con el criterio para un 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: EC 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 de un solo brazo Series temporales interrumpidas sin grupo control paralelo	Estudios de casos y controles de diagnósticos	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-texts	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 serie de casos	Serie de casos

AA: asignación aleatoria. PD: prueba diagnóstica.

columna ubicada en el extremo izquierdo presenta los niveles de evidencia. En su aplicación, se sugiere que, al valorar la evidencia, se debería indicar el área de investigación. Se considera 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 (Merlin, et al. 2009).

9.4.8. Agencia de Evaluación de Tecnología Médica

En lo que se refiere a nuestro país, hay que destacar el esquema de gradación propuesto por la Agència d'Avaluació de Tecnologia Mèdica (AATM) de la Generalitat de Catalunya (Manterola, 2009). Esta clasificación tiene en cuenta además del diseño de los estudios una valoración específica de su calidad (tabla 9.12).

Tabla 9.12. *Niveles de calidad de la evidencia científica (AATM)*

Nivel	Fuerza de la evidencia	Tipo de diseño	Condiciones de rigurosidad científica
I	Adecuada	Metanálisis de ECA	Análisis de datos individuales de los pacientes No heterogeneidad Diferentes técnicas de análisis Metarregresión Metanálisis Calidad de los estudios
II	Adecuada	ECA de muestra grande	Evaluación del poder estadístico Multicéntrico Calidad del estudio
III	Buena a regular	ECA de muestra pequeña	Evaluación del poder estadístico Calidad del estudio
IV	Buena a regular	Ensayo prospectivo controlado no aleatorizado	Controles coincidentes en el tiempo Multicéntrico Calidad del estudio
V	Regular	Ensayo retrospectivo controlado no aleatorizado	Controles históricos Calidad del estudio

Continúa

Tabla 9.12. *Continuación*

Nivel	Fuerza de la evidencia	Tipo de diseño	Condiciones de rigurosidad científica
VI	Regular	Estudios de cohorte	Multicéntrico Apareamiento Calidad del estudio
VII	Regular	Estudios de casos y controles	Multicéntrico Calidad del estudio
VIII	Pobre	Series clínicas no controladas Estudios descriptivos: Vigilancia epidemiológica Encuestas Registros Bases de datos Comités de expertos Conferencias de consenso	Multicéntrico
IX	Pobre	Anécdotas o casos únicos	

ECA: ensayo controlado aleatorizado.

Esta clasificación es una prueba más de la amplia difusión del uso de las prácticas de valoración de la evidencia, lo que es un indicativo de la necesidad de proceder a interiorizar los procesos de sistematización de búsqueda de evidencia y a la generalización de los resúmenes especializados, tal y como propuso Cochrane (1972), como pasos fundamentales para la obtención de la mejor evidencia disponible. Sin dejar de tomar en consideración no solo las recomendaciones aquí planteadas respecto a las características metodológicas y poblacionales, sino también culturales, económicas, tecnológicas, ambientales, etc.; otorgando el valor necesario, tanto a la validez y como a la fiabilidad, sin importar que aunque hay elementos que parecen funcionar en otros lugares, tendrían que ser siempre considerados en nuestro contexto.

Se ha procedido aquí a presentar cronológicamente una mínima expresión de las múltiples propuestas de clasificaciones para valorar la evidencia que existen y que están en uso en la actualidad, aproximadamente un

centenar, desde la más antigua y más simple formulada por Sackett (1978), origen de muchas de las actuales, a otras que han pretendido ser exhaustivas, y considerar todas las posibilidades en el análisis de la investigación y la evidencia.

En la actualidad, todavía quedan profesionales clínicos que ven en la variedad de opciones existentes para la evaluación de la evidencia más problemas que ayuda para desarrollo de su actividad profesional. Una cuestión que debería resolverse con celeridad, porque en estos momentos solo una parte de la atención sanitaria está basada en la evidencia, pero no queda muy lejos el tiempo en que toda actuación sanitaria estará basada en evidencia disponible. Adicionalmente, es necesario hacer mención al lenguaje especializado que se utiliza, dado que ha generado situaciones de confusión e incomprensibilidad; y es un factor de retraso en su utilización.

En la práctica clínica resulta esencial la aplicación de la evidencia en torno a los diferentes problemas clínicos, ya que nos permite ejercer el juicio clínico basándonos en el nivel de confianza que nos proporcionan los resultados que aparecen en las investigaciones publicadas y nos facilita la evaluación de los beneficios y los riesgos permitiendo el desarrollo de guías de práctica clínica y la realización de las correspondientes evaluaciones de tecnologías sanitarias. Este es un elemento esencial que ha determinado la utilización de las propuestas de evaluación descritas anteriormente, tanto en instituciones sanitarias como en la docencia universitaria, por lo que es relevante conocer y comprender mejor la metodología a través de la cual se efectúan estos procesos. Un conjunto de clasificaciones amplias con una propuesta distinta para diferentes escenarios (CEBM, NHMRC, etc.), así como de clasificaciones específicas que se centran en un escenario puntual (CTFPHC, SIGN, etc.) y otras como GRADE, que ha sido incorporada para evaluar la calidad de la evidencia disponible, para realizar recomendaciones y generar guías de práctica clínica basadas en la evidencia por alrededor de 90 instituciones de prestigio nacionales e internacionales (Guyatt, et al., 2008).

El siguiente paso consiste en darle un valor a esta evidencia, para lo cual se debe elegir la clasificación que más se ajuste a nuestras necesidades y que nos permita discriminar entre un nivel y otro. Hemos de hacer notar que algunas de ellas son específicas para cierto tipo de escenarios, lo que nos puede facilitar el ajustar la elección para efectuar las recomendaciones más adecuadas al entorno asistencial y poblacional.

Bibliografía

Anderson, N., Herriot, P. & Hodgkinson, G. P. (2001). «The practitioner–researcher divide in industrial, work and organizational psychology: Where are we now and where do we go from here?», *Journal of Occupational and Organizational Psychology,* 74(4): 391-411.

Birtwhistle R, Pottie K, Shaw E, Dickinson J A, Brauer P, Fortin M, Bell M, Singh H, Tonelli M, Connor GS, Lewin G, Joffres M, Parkin P. (2012) «Canadian Task Force on Preventive Health Care: we're back!», *Can Fam Physician*, 58(1): 13-5.

Canadian Task Force on the Periodic Health Examination (1979). «The periodic health examination», *Can Med Assoc J*, 121(9):1193-1254.

Canadian Task Force on Preventive Health Care (2003). «New grades for recommendations from the Canadian Task Force on Preventive Health Care», *Can Med Assoc J*, 169 (3): 207-8.

Canfield SE, Dahm P. (2012). «Rating the quality of evidence and the strength of recommendations using GRADE», *World J Urol*, 29 (3): 311-317.

Centre for Evidence-based Medicine (CEBM)-Levels of Evidence (March 2009). Disponible en http://www.cebm.net/index.aspx?o=1025. [15/01/2018].

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 [15/01/2018].

Ciapponi A. (2015). **«**QUADAS-2: instrumento para la evaluación de la calidad de estudios de precisión diagnóstica», *Evidencia- Actualización en la Práctica Ambulatoria*, 18(1): 22-26.

Cochrane AL. (1972). *Effectiveness & Efficiency: Random Reflections on Health Services*. London: Royal Society of Medicine Press.

Cook DJ, Mulrow CD & Haynes RB. (1997). «Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions», *Annals of Internal Medecine,* 126(5): 376-380.

Consorcio AGREE (2009). «Instrumento AGREE II: instrumento para la evaluación de guías de práctica clínica», Disponible en: http://www.guiasalud.es/contenidos/documentos/Guias_Practica_Clinica/Spanish-AGREE-II.pdf [10/06/2018]

Davies H, Nutley S & Smith P. (2000). «What Works? The Role of Evidence in Public Sector Policy and Practice», Public Money and Management, 19(1): 3-5.

Davies P & Boruch R. (2001). Editorial: The Campbell Collaboration. «Does for public policy what Cochrane does for health», *British Medical Journal,* 323:294-295.

Evidence Based Working Group (1992). Evidence-based medicine. «A new approach to teaching the practice of medicine», *Journal of the American Medical Association*, 268(17): 2420-2425.

Gibbons M, Limoges C, Nowotny H, Schwartzman P, Scott P & Trow M. (1994). *The New Production of Knowledge: the Dynamics of Science and Research in Contemporary Societies*. Sage: London.

Guyatt G, Gutterman D, Baumann M H, Addrizzo-Harris D, Hylek E M, Phillips B, Raskob G, Lewis SZ & Schünemann H. (2006). «Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians Task Force», *Chest*, 129 (1): 174-181.

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P & Schünemann HJ. (2011). «GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables», *J Clin Epidemiol*, 64 (4): 383-394.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ, GRADE Working Group. (2008). «GRADE: An emerging consensus on rating quality of evidence and strength of recommendations», *Br Med J*, 336(7650):924-926.

Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ & Cook RJ. (1995). «Users› Guides to the Medical Literature: IX. A method for grading health care recommendations», *JAMA*, 274 (22):1800-1804.

Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, Atkins D Methods Work Group, Third US Preventive Services Task Force (2001). «Current methods of the US Preventive Services Task Force: a review of the process», *Am J Prev Med*, 20(3 Suppl): 21-35.

Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, Moschetti I, Phillips B, Thorton H, Goddard O & Hodgkinsonet. (2011) «Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document)», Disponible en http://www.cebm.net/index.aspx?o=5653 [16/04/2018].

Jovell AJ & Navarro-Rubio MD. (1995). «Evaluación de la evidencia científica», *Med Clin (Barc)*, 105(19):740-743.

Macdonald G. (1999). «Evidence-Based social care: Wheels off the Runaway?», *Public Money & Mangement*, 19 (1): 25-32.

Manterola C, Asenjo-Lobos C & Otzen T. (2014). «Jerarquización de la evidencia. Niveles de evidencia y grados de recomendación de uso actual», *Rev. chil. Infectol*, 31(6): 705-708.

Manterola C & Zavando D. (Grupo MINCIR) (2009). «Cómo interpretar los "Niveles de Evidencia" en los diferentes escenarios clínicos», *Rev Chil Cir*, 61(6): 582-595.

Manterola C. (2002). «Medicina basada en la evidencia. Conceptos generales y razones para aplicación en cirugía», *Rev Chil Cir*, 54(5): 550-554.

Manterola C. (2009). «Medicina basada en la evidencia o medicina basada en pruebas. Generalidades acerca de su aplicación en la práctica clínica cotidiana», *Rev Med Clin Condes*, 20: 125-30.

Maro-Castillejo M & Viana Zulaica C. (2007) «Calidad de la evidencia y grado de recomendación», *Guías Clínicas*, 7, Supl 1: 65-82.

Merlin T, Weston A, Tooher R. (2009). «Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'», *BMC Med Res Methodol*, 9: 34.

Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Deveraux PJ, Prasad K, Neuman I, Carrasco-Labra A, Agoritsas T, Hatala R, Meade MO, Wyer P, Cook DJ & Guyatt G. (2014). «How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature», *JAMA,* 312(2):171–9.

Murad MH, Alsawas M, Asi N, Alahdab F. (2015). «The New Evidence Pyramid**»,** *International Society of Health Care for Evidence-based Health Care*, 21st edition; 8-9. Disponible en: http://ebm.mcmaster.ca/documents/ebhc_newsletter_21-Oct2015.pdf [10/08/2015].

NICE. The guidelines manual: consultation on the 2012 update. Disponible en http://www.nice.org.uk/aboutnice/howwework/developingniceclinicalguidelines/GuidelinesManualConsultation2012.jsp [21/04/2018].

Sackett DL. (1989). «Rules of evidence and clinical recommendations on the use of antithrombotic agents», *Chest*, 95(2 Suppl):2S-4S.

Sackett RL, Richardson WS, Rosemberg W & Haynes RB. (1997). *Evidence-Based Medicine. How to practice and teach EBM*. Londres: Churchill Livingstone.

Saha S, Hoerger TJ, Pignone MP, Teutsch SM, Helfand M, Mandelblatt JS, for the Cost Work Group of the Third U.S. Preventive Services Task Force. (2001). «The art and science of incorporating cost effectiveness into evidence-based recommendations for clinical preventive services», *Am J Prev Med*, 20(3S): 36- 43.

Scottish Intercollegiate Guidelines Network (2015). SIGN 50. A guideline developer's handbook. Disponible en: https://www.sign.ac.uk/assets/sign50_2015.pdf [14/03/2018].

SIGN (2014) Methodological principles. Disponible en http://www.sign.ac.uk/methodology/index.html [18/02/2018].

Smith R. (1991). Where is the wisdom…? The poverty of medical evidence. *British Medical Journal*, 303(6806):798-9.

United States Preventive Services Task Force (2017). About the USPSTF. Disponible en: https://www.uspreventiveservicestaskforce.org/Page/Name/about-the-uspstf. [12/01/2018]

Whitley R. (1984). «The scientific status of management research as a practically-oriented social science», *Journal of Management Studies,* 21(4): 369–390.