BOL PEDIATR 2002; 42: 120-130

Revisión

Valoración crítica de documentos científicos. Aplicabilidad de los resultados de la valoración a nuestra práctica clínica

C. OCHOA SANGRADOR

Servicio de Pediatría. Hospital Virgen de la Concha. Zamora.

RESUMEN

La medicina basada en la evidencia nos ofrece herramientas de gran utilidad para poder resolver problemas clínicos mediante el análisis eficiente de la literatura científica. Herramientas metodológicas que, si adquirimos y ejercitamos, nos ayudarán a valorar cualquier evidencia científica y a integrarla con nuestros conocimientos y experiencia clínica, para poder decidir sobre su aplicabilidad e idoneidad en un paciente concreto.

En esta exposición repasaremos los principios generales de la valoración crítica de la literatura científica. También revisaremos los principales criterios a considerar en la valoración de la validez y aplicabilidad de los estudios de evaluación de intervenciones sanitarias y de pruebas diagnósticas.

Finalizaremos presentando algunas de las iniciativas que han permitido desarrollar y difundir la incorporación de la valoración crítica de documentos científicos al ejercicio de la medicina: el programa CASP y los Archivos de Temas Valorados Críticamente.

ABSTRACT

The evidence based medicine offers us very useful tools to solve clinical problems by means of the efficient analysis of scientific literature. If we acquire and exercise this methodological tools, we will be able to value any scientific evidence and integrate it with our knowledge and clinical experience, and also we will be able to decide on its applicability and suitability in a concrete patient.

In this article we will review the general principles of the critical appraisal of scientific literature. Also we will review the main criteria to consider in the analysis of the validity and applicability of the studies of evaluation of health interventions and diagnostic tests.

We will finalize presenting some of the initiatives that have allowed the incorporation of the critical appraisal of scientific documents at the clinical practice: the CASP program and the Critically Appraissed Topics Banks.

INTRODUCCIÓN

En nuestro ejercicio profesional nos enfrentamos a menudo a situaciones en las que se nos plantean dudas sobre distintos aspectos de la práctica clínica: etiología, diagnóstico, tratamiento, etc. Tradicionalmente hemos intentado solucionar nuestras dudas a través de consultas a libros de texto, consultas a revistas o preguntando a colegas. Pero este abordaje clásico presenta importantes limitaciones[1]. Como alternativa, la Medicina Basada en la Evidencia (MBE) propone un método estructurado para resolver dudas clínicas que comprende cuatro pasos[2]:

Correspondencia: Dr. Carlos Ochoa Sangrador. Jardines Eduardo Barrón 1 bis 3º. 49016 Zamora.
Correo electrónico: cochoas@meditex.es
Recibido: Junio 2002. *Aceptado:* Junio 2002

1. Convertir nuestra duda en una pregunta clínica estructurada.
2. Realizar una búsqueda bibliográfica para encontrar artículos que puedan responderla.
3. Realizar una valoración crítica de los documentos recuperados, analizando la validez y relevancia de los resultados.
4. Y por último, integrar la valoración realizada con nuestra experiencia clínica, considerando la aplicabilidad de los resultados en nuestros pacientes y actuar.

Una vez elaborada la pregunta clínica y encontrado el artículo en el que se trata de responder a la misma, el siguiente paso será realizar una valoración crítica del documento encontrado. Esta fase va a resultar fundamental ya que a menudo la calidad de los artículos científicos es deficiente, no se ajustan al problema clínico que se trata de resolver, tienen errores metodológicos que comprometen los resultados o éstos son presentados de forma que limitan su correcta interpretación.

El objetivo de la valoración crítica es analizar la validez y aplicabilidad de las evidencias publicadas, para lo que se requieren ciertos conocimientos y habilidades que es preciso adquirir y ejercitar. Comenzaremos esta exposición repasando los principales criterios a considerar en la valoración crítica de la literatura, y finalizaremos presentando algunas de las iniciativas que han contribuido a desarrollarla y difundirla.

Criterios a considerar en la valoración crítica

La valoración tendrá que comenzar necesariamente juzgando si el estudio trata de contestar a nuestra **pregunta clínica**. Esta cuestión debería haberse resuelto en las fases de búsqueda y selección del artículo, sin embargo debe ser comprobada antes de dedicar nuestro tiempo a un análisis más detallado[3]. Para ello, nos fijaremos en el tipo de población estudiada, el tipo de intervención terapéutica o diagnóstica evaluada, el diseño del estudio y los criterios empleados para la medición de los resultados.

Para contestar a una pregunta sobre eficacia o efectividad de una intervención terapéutica o preventiva el **diseño** más apropiado es el ensayo clínico aleatorizado (ECA). Por sus requisitos metodológicos, el ECA es el tipo de diseño que mayor grado de evidencia nos aporta. Sin embargo, existen distintos aspectos de la práctica clínica, que también generan preguntas (etiología, pronóstico, diagnóstico), que a menudo no pueden ser respondidas mediante ensayos clínicos, por lo que tendrán que utilizarse otros tipos de diseño: estudios de cohortes, estudios de casos y controles, series de casos, etc.

La valoración crítica va a ser diferente en función del tipo de pregunta clínica y del diseño elegido en el estudio evaluado. Existen **criterios de valoración comunes** a cualquier estudio entre los que merece la pena destacar: adecuación del diseño y de la población estudiada a la pregunta de investigación, tamaño muestral suficiente, homogeneidad de los grupos comparados al inicio del estudio, seguimiento uniforme y completo de los sujetos de estudio, medición y análisis apropiados de los resultados e interpretación adecuada de los mismos.

En cuanto a los **criterios de valoración específicos**, éstos han sido excepcionalmente expuestos, en forma de guías de interpretación, por el *Evidence-Based Medicine Working Group*[4-32].

Para mostrar el fundamento y la metodología de la valoración crítica de la literatura, abordaremos dos de los tipos de estudio habitualmente más considerados: los estudios sobre eficacia de intervenciones sanitarias y los estudios de evaluación de pruebas diagnósticas. Para ello, repasaremos los criterios de valoración específicos de ambos tipos de estudio (Tablas I y II), siguiendo el esquema propuesto por el *Evidence Based Medicine Working Group*[3]:

- ¿Son válidos sus resultados?
- ¿Cuáles son los resultados?
- ¿Son aplicables en tu medio?

Evaluación de una intervención sanitaria

¿Son válidos sus resultados?

Dado que una intervención sanitaria, ya sea terapéutica o preventiva, es un factor de estudio susceptible de ser controlado por el investigador, el diseño que proporciona las mejores pruebas sobre su eficacia es el ECA.

La característica metodológica clave del ECA es la **asignación aleatoria** de los sujetos a los grupos de estudio ya que, al intervenir solamente el azar, tiende a asegurar que se produce una distribución equilibrada de todas las variables, tanto conocidas como desconocidas, entre los grupos. Esta tendencia es mayor cuanto más elevado es el número

TABLA I. CRITERIOS DE EVALUACIÓN DE UNA INTERVENCIÓN SANITARIA

Validez de los resultados
Asignación aleatoria
Enmascaramiento de la asignación
Seguimiento completo
Análisis por intención de tratar / por protocolo
Diseño ciego
Homogeneidad inicial de los grupos de estudio
Manejo homogéneo de los grupos de estudio
Magnitud de los resultados
Medidas de efecto (RRR, RAR, RR, OR, diferencias de medias)
Control de variables de confusión
Intervalos de confianza
Aplicabilidad de los resultados
Adecuación del ámbito de estudio y criterios de inclusión-exclusión
Relevancia clínica de las variables principal y secundarias
Repercusión clínica de los resultados (NNT)
Adecuación del análisis de subgrupos
Valoración de perjuicios y costes

RRR reducción relativa del riesgo. RAR reducción absoluta del riesgo. RR riesgo relativo. OR odds ratio. NNT número necesario a tratar

TABLA II. CRITERIOS DE EVALUACIÓN DE UNA PRUEBA DIAGNÓSTICA

Validez de los resultados
Patrón de referencia válido
Comparación ciega de la prueba diagnóstica y el patrón de referencia
Espectro de pacientes adecuado
Descripción adecuada de los métodos
Magnitud de los resultados
Presentación correcta de resultados:
Cocientes de probabilidades
Análisis de probabilidades preprueba y postprueba
Curvas ROC
Intervalos de confianza
Aplicabilidad de los resultados
Validez externa de la reproducibilidad e interpretación
Adecuación del espectro de pacientes
Repercusión sobre el manejo diagnóstico / terapéutico del paciente
Beneficio sobre el paciente del resultado de la prueba

de sujetos. Para que ni el sujeto ni el investigador puedan influir en la decisión de qué intervención recibirá cada uno de los participantes, es igualmente importante que se aplique alguna técnica de enmascaramiento en el proceso de asignación.

Otro aspecto importante a considerar es el **seguimiento completo** de los grupos de estudio. Todos los sujetos incluidos en un ECA deberían ser tenidos en cuenta hasta su conclusión. La pérdida en el seguimiento de un número elevado de sujetos puede modificar los resultados finales, ya que su salida del estudio puede estar relacionada con una evolución diferente, que en ocasiones habrá sido más desfavorable y en otras más favorable que la del resto de la muestra. El lector tendrá que juzgar según las circunstancias concretas del estudio si las pérdidas serán cualitativa o cuantitativamente importantes como para invalidar los resultados. En ocasiones podremos estimar la repercusión de las pérdidas realizando un análisis de sensibilidad, en el que se prueba a asignar a los individuos perdidos una evolución favorable o desfavorable, observando si se producen cambios en los resultados.

Al igual que en la práctica clínica, en los ECA muchos sujetos incumplen la intervención prescrita o no se les puede aplicar por diversas circunstancias. En esta situación puede constituir un error excluir a esos sujetos del análisis, ya que las razones por las que un tratamiento no se realiza tienen implicaciones pronósticas, y la exclusión introduciría un sesgo en el estudio. En el **análisis por intención de tratar**, los datos se analizan considerando a cada sujeto como si hubiera recibido la intervención que le fue asignada inicialmente, y no la que realmente recibió (análisis por protocolo). Esta estrategia, controvertida, pretende controlar posibles sesgos, tanto conocidos como desconocidos, aunque no siempre es apropiada[33].

El conocimiento por parte de los pacientes del grupo de tratamiento al que pertenecen tiende a modificar la opinión sobre su eficacia. Igualmente ocurre con los investigadores, en los que puede inducir un comportamiento diferenciado en el manejo o evaluación de la respuesta. La mejor manera de evitar posibles sesgos es mantener, en la medida de lo posible, el **ciego respecto del tratamiento** tanto del paciente (simple ciego) como de los investigadores (doble ciego). El cumplimiento de este requisito debe estar expresamente justificado en la metodología del estudio, utilizando adecuadas técnicas de enmascaramiento.

TABLA III. EFICACIA DE DEXAMETASONA (DESDE EL 7º DÍA DE VIDA; INTRAVENOSA; 0,25 MG/KG/12 HORAS 2 DÍAS CADA 10 DÍAS) PARA LA PREVENCIÓN DE ENFERMEDAD PULMONAR CRÓNICA EN NEONATOS DE MUY BAJO PESO CON VENTILACIÓN ASISTIDA (RESULTADOS A LOS 36 SEMANAS DE VIDA)[34]. MEDIDAS DE EFECTO CON SUS INTERVALOS DE CONFIANZA AL 95% (IC 95%).

	Enfermedad pulmonar crónica		Proporción eventos
	Sí	No	
Dexametasona n=39	9	30	P_i= 9 / 39= 0,23
Control (placebo) n=39	18	21	P_c= 18 / 39= 0,46

Reducción relativa del riesgo

$$RRR = \frac{P_c - P_i}{Pc} = \frac{0,46 - 0,23}{0,46} = 0,50$$ (IC 95% = 0,06-0,94)

Reducción absoluta del riesgo

$$RAR = P_c - P_i = 0,46 - 0,23 = 0,23$$ (IC 95% = 0,03-0,44)

Riesgo relativo

$$RR = \frac{P_i}{Pc} = \frac{0,23}{0,46} = 0,50$$ (IC 95% = 0,26-0,97)

Odds ratio

$$OR = \frac{9 / 30}{18 / 21} = \frac{9 \times 21}{18 \times 30} = 0,35$$ (IC 95% = 0,13-0,93)

Número necesario a tratar

$$NNT = \frac{1}{RAR} = \frac{1}{0,23} = 4$$ (IC 95% = 2-39)

Tal y como mencionamos anteriormente, la asignación aleatoria de un ECA pretende garantizar la **comparabilidad de los grupos al inicio del estudio**, de modo que lo ideal sería que sólo se diferenciaran en la intervención que van a recibir. No obstante, resulta fundamental realizar una comprobación de que los grupos obtenidos han sido finalmente homogéneos, especialmente si las muestras son pequeñas. Diferencias clínicamente importantes podrían comprometer la validez del estudio. Cuando esto ocurre los investigadores pueden recurrir a técnicas estadísticas que comprueben la influencia de dichas diferencias.

Asimismo, es importante comprobar que el **manejo** y **seguimiento** de los distintos grupos ha sido **homogéneo** a lo largo del estudio. Cualquier otra intervención aplicada a los sujetos de estudio y que pudiera influir en los resultados debería ser equiparable en ambos grupos o controlada en el análisis.

La valoración secuencial de estos criterios nos permitirá juzgar si el estudio es válido y por lo tanto si merece la pena seguir adelante examinando la magnitud de los resultados y su aplicabilidad.

¿Cuáles son los resultados?

Con frecuencia los ECA utilizan como criterio de medición de los resultados de una intervención la presencia o ausencia de un evento adverso o favorable. En el caso de una intervención beneficiosa, como la prevención de enfermedad pulmonar crónica con dexametasona en neonatos de muy bajo peso[34], su efecto debería reflejarse en la existencia de una menor proporción de eventos en el grupo de intervención que en el grupo control (Tabla III). La forma más simple de estimar ese efecto es calcular la diferencia absoluta entre ambas proporciones, lo que se conoce como reducción absoluta del riesgo (**RAR**).

Sin embargo, la medida de efecto más comúnmente referida en las publicaciones es la reducción relativa del riesgo (**RRR**), que ajusta la RAR a una escala relativa y que expresa la proporción de reducción riesgo respecto al riesgo en el grupo control (Tabla III). Otras medidas muy utilizadas son el riesgo relativo (RR) y la odds ratio (OR).

Todas estas medidas vendrán referidas o podrán ser calculadas a partir de los resultados contenidos en el trabajo examinado. Pero a pesar de que son las mejores medidas disponibles, nunca podremos estar absolutamente segu-

ros de que reflejan la verdadera reducción del riesgo, ya que son estimaciones puntuales obtenidas a partir de muestras. No obstante, es previsible que el verdadero valor se encuentre en las proximidades de esta estimación, tanto más próximo cuanto más grande sea la muestra (menor error aleatorio, más precisión), pudiéndose concretar esa proximidad en lo que se conoce como **intervalos de confianza**.

Habitualmente los investigadores utilizan el intervalo de confianza al 95% que quiere decir que dentro de ese intervalo se encontraría la verdadera reducción del riesgo en el 95% de los casos (en 95 de cada 100 estudios similares). El intervalo de confianza contiene información de interpretación más intuitiva que el clásicamente referido nivel de significación estadística ($p < 0,05$), ya que el lector puede juzgar cuáles son las estimaciones más favorables y desfavorables de la reducción del riesgo (límites superior e inferior del intervalo)[35]. No obstante, tanto el cálculo de los intervalos de confianza como del nivel de significación no deben sustraer al lector de la consideración de la significación clínica de los resultados obtenidos.

No todos los ECA utilizan como criterio de medición variables dicotómicas. En ocasiones el efecto se refiere a una variable cuantitativa: índice de oxigenación, presión arterial pulmonar, escala de Apgar, frecuencia cardiaca, etc.[36,37]. En estos casos, la medida del efecto de la intervención podrá estimarse calculando la diferencia entre los valores medios de dichas variables en los distintos grupos de estudio (**diferencia de medias o medianas**). Estas estimaciones puntuales y sus respectivos intervalos de confianza permitirán una interpretación directa y con sentido clínico de la dimensión de los resultados.

Una vez determinadas la magnitud y la precisión del efecto del tratamiento podemos proceder a valorar si los resultados del estudio son aplicables en la práctica clínica a nuestros pacientes.

¿Son aplicables en tu medio?

A la hora de juzgar si los resultados del estudio son aplicables a nuestro paciente es preciso tener en cuenta en qué manera nuestro entorno de trabajo se parece al **ámbito** en el que los pacientes han sido reclutados, si nuestro paciente cumple o no los **criterios de inclusión** y **exclusión** del estudio y si tiene condiciones de gravedad o comorbilidad diferentes que puedan interferir en el resultado.

En ocasiones los autores del trabajo presentan en los resultados información diferenciada de subgrupos de pacientes, en los que las medidas de efecto son más o menos favorables. Este hecho requiere especial precaución cuando los resultados globales han resultado menos favorables. Podríamos considerar los resultados del **análisis de subgrupos** si la diferencia del efecto es grande, no atribuible al azar, el análisis ha sido planificado *a priori* y resulta concordante con los resultados de otros estudios[38].

Para poder planificar la aplicabilidad del estudio es preciso valorar la importancia clínica del parámetro empleado en la medición de los resultados. Existen algunos parámetros (ejemplo: supervivencia) que no ofrecen dudas en cuanto a su repercusión clínica, sin embargo otros (ejemplos: escalas subjetivas de valoración de síntomas, indicación de hospitalización), cuyo **significado clínico** es más problemático[25].

Otro aspecto importante a la hora de juzgar la significación clínica de un resultado, es considerar también los posibles **efectos adversos** no deseados ligados a la intervención aplicada. Volviendo al ejemplo del tratamiento con dexametasona aplicado para reducir el riesgo de enfermedad pulmonar crónica en neonatos[34], los autores refieren que el grupo tratado precisó con más frecuencia la administración de insulina por hiperglucemias. Por lo tanto, en la decisión final sobre la aplicabilidad de los resultados tendrán que valorarse la magnitud e importancia tanto de los beneficios como de los perjuicios.

La magnitud del efecto esperado en un paciente puede estimarse a partir de las medidas de RAR y RRR. Sin embargo, existe otra medida que ofrece una información más intuitiva de los posibles beneficios o perjuicios: el número necesario a tratar (NNT)[39]. El NNT es el número de sujetos que se necesitaría tratar con una intervención específica para producir, o evitar, un evento determinado. El equivalente al NNT para la aparición de efectos adversos no deseados se denomina número necesario a perjudicar (NNP). El NNT y su intervalo de confianza puede calcularse fácilmente a partir de los inversos de la RAR y los límites de su intervalo de confianza (Tabla III).

Con la intervención evaluada en nuestro ejemplo (Tabla III) la reducción absoluta del riesgo es 23% (RAR=0,23) por lo que tendremos que tratar 4 sujetos (NNT=4) para beneficiar a uno. Si queremos determinar en un paciente con-

creto la magnitud del efecto esperado necesitamos estimar su riesgo basal (sin intervención), según su gravedad y comorbilidad, y a partir de él calcular (aplicando la reducción relativa del riesgo) su RAR y su NNT. De igual manera podemos proceder para calcular su NNP. Ambas medidas nos ilustrarán sobre la magnitud de los beneficios y perjuicios a la hora de tomar una decisión.

En este punto, debemos señalar que la consideración de beneficios y perjuicios no puede guiarse exclusivamente por medidas cuantitativas de impacto. También deben tenerse en cuenta la repercusión clínica de los mismos y por supuesto el **coste**, facilidad de aplicación, grado de cumplimiento y accesibilidad de la intervención considerada.

Evaluación de una prueba diagnóstica

¿Son válidos sus resultados?

Para poder valorar la validez de una prueba diagnóstica es preciso comparar sus resultados con los de un patrón de referencia en una serie de pacientes[40]. El patrón de referencia empleado tiene que contar con una validez contrastada o, al menos, aceptada por consenso. La utilización de un **patrón de referencia** defectuoso puede introducir sesgos en las estimaciones de validez de la prueba diagnóstica.

En relación con el patrón de referencia, resulta también importante considerar si es capaz de clasificar el estado de enfermedad en todas las observaciones. En el caso de que existan observaciones con un diagnóstico indeterminado, si éstas son excluidas del análisis, se producirán estimaciones sesgadas de las características operativas de la prueba diagnóstica. Este sesgo, conocido como **sesgo por exclusión de indeterminados**, ocasiona habitualmente sobrestimaciones de la sensibilidad y de la especificidad[41,42].

Si aceptamos la validez del patrón de referencia, la siguiente cuestión a tener en cuenta es si tanto la prueba diagnóstica como el patrón de referencia han sido realizados de forma independiente. Cuando no se realizan de forma independiente, puede existir un **sesgo de revisión** si el resultado de una prueba es susceptible de interpretación subjetiva y se ve influida por el conocimiento del diagnóstico o de las características clínicas del paciente. Para poder garantizar la validez de las estimaciones, deberían realizarse de forma ciega la prueba diagnóstica y el patrón de referencia.

Otro aspecto importante que condiciona la validez del estudio es la inclusión en el mismo de un **adecuado espectro de pacientes**, similar al que nos encontramos en nuestra práctica clínica. Para valorar esta cuestión es preciso que los criterios de selección y las características clínicas y epidemiológicas de la muestra analizada estén claramente presentados.

El diseño del estudio debe tratar de garantizar que en la muestra no se hayan excluido pacientes, en función del resultado de la prueba o de la existencia de mayor o menor riesgo de enfermedad. Incurriremos en un **sesgo de verificación diagnóstica**, cuando la probabilidad de que se les realice el patrón de referencia sea menor entre los sujetos con la prueba diagnóstica negativa y por lo tanto sea menos probable que éstos entren en el estudio[43,44].

El último criterio a valorar para juzgar la validez del estudio es si se describen los **métodos** con suficiente detalle como para permitir su reproducción. Esta descripción debe incluir todos los aspectos de la preparación de pacientes, realización de la prueba y su interpretación.

Si después de considerar todos estos aspectos, hemos decidido que el estudio es suficientemente válido, procederemos a examinar las propiedades de la prueba diagnóstica.

¿Cuáles son los resultados?

El punto de partida del proceso diagnóstico es habitualmente un paciente, con unas características de gravedad y comorbilidad concretas, que le confieren una probabilidad determinada de tener la entidad a diagnosticar (**probabilidad preprueba**). El objetivo de la realización de la prueba diagnóstica es, una vez conocido el resultado, modificar esa probabilidad hasta obtener una probabilidad postprueba. La magnitud y dirección de ese cambio va a depender de las características operativas de la prueba diagnóstica, pero en todo caso debemos tener en cuenta que el punto de partida, la probabilidad preprueba, va a resultar muy importante en ese proceso.

Consideremos el escenario diagnóstico más simple, en el que tanto el patrón de referencia como la prueba diagnóstica clasifican a los pacientes en dos grupos, en función de la presencia o ausencia de un síntoma, signo o enfermedad. En la Tabla IV presentamos, como ejemplo, los resultados de un estudio sobre la validez de la detección de leucocitos en orina como prueba diagnóstica de infección uri-

TABLA IV. TABLA DE CONTINGENCIA DE LA EVALUACIÓN DEL TEST DE LA ESTEARASA LEUCOCITARIA PARA EL DIAGNÓSTICO DE INFECCIÓN URINARIA (UROCULTIVO POSITIVO)[45]. CARACTERÍSTICAS OPERATIVAS DE LA PRUEBA CON SUS INTERVALOS DE CONFIANZA AL 95% (IC 95%)

		Urocultivo +	Urocultivo -	
Test de la estearasa	+	81 (a)	160 (b)	241
Leucocitaria	-	21 (c)	427 (d)	448
		102	587	689

Sensibilidad (Se) = a/(a+c) = 81/102 = 0,79 (IC 95%: 0,71-0,87). Especificidad (Es) = d/(b+d) = 427/587 = 0,72 (IC 95%: 0,69-0,76).
Valor predictivo positivo (VPP) = a/(a+b) = 81/241 = 0,33 (IC 95%: 0,27-0,39).
Valor predictivo negativo (VPN) = d/(c+d) = 427/448 = 0,95 (IC 95%: 0,93-0,97).
Cociente de probabilidades positivo = Se/(1-Es) = 2,91 (IC 95%: 2,47-3,44).
Cociente de probabilidades negativo = (1-Se)/Es = 0,28 (IC 95%: 0,19-0,42).
Probabilidad preprueba (Ppre) = (a+c)/(a+b+c+d) = 102/689 = 0,14. Odds preprueba = Ppre/(1-Ppre) = 0,14/(1-0,14) = 0,16

naria45. En este caso, el urocultivo, cuya positividad confirma la existencia de infección (para orinas obtenidas por técnica estéril), es el patrón de referencia.

Generalmente los resultados se expresan a partir de la proporción de aciertos de la prueba diagnóstica entre las poblaciones enferma (sensibilidad) y sana (especificidad). La sensibilidad es la probabilidad de que la prueba sea positiva si la condición de estudio está presente (patrón de referencia positivo), mientras que la especificidad es la probabilidad de que la prueba sea negativa si la condición está ausente. Sin embargo, la sensibilidad o la especificidad no nos facilitan el cálculo de la probabilidad postprueba. Para ello resulta más útil el empleo de los cocientes de probabilidades.

El **cociente de probabilidades** (CP) para un determinado resultado de una prueba diagnóstica está definido como la probabilidad de dicho resultado en presencia de enfermedad dividida por la probabilidad de dicho resultado en ausencia de enfermedad. Los CP resumen información de la sensibilidad y de la especificidad e indican la capacidad de la prueba para incrementar o disminuir la verosimilitud de un determinado diagnóstico. En la Tabla IV podemos observar los CP de los resultados positivo y negativo de la prueba de la estearasa leucocitaria en orina y su relación con la sensibilidad y especificidad. Utilizando los CP se pueden calcular las probabilidades postprueba (valores predictivos) a partir de la probabilidad preprueba de cada paciente individual.

Para poder operar con los CP en el cálculo de probabilidades, éstas deben transformarse en ventajas (*odds*). Las ventajas u *odds* se calculan dividiendo las probabilidades por sus complementarios (P / [1-P]). Los pasos a seguir en el cálculo de la **probabilidad postprueba** son: 1) transformar la probabilidad preprueba en *odds* preprueba, 2) multiplicar la *odds* preprueba por el CP del resultado encontrado para obtener la odds postprueba, 3) transformar la *odds* postprueba en probabilidad (probabilidad = odds / [1+odds]).

Si asumimos el CP positivo del estudio de Lohr y cols.[45] (2,91), podemos estimar la probabilidad postprueba de infección urinaria en un escenario diferente al suyo. Para un lactante sano (probabilidad preprueba estimada = 0,01) si la prueba de la estearasa leucocitaria es positiva la probabilidad postprueba de infección urinaria será 0,028 (2,8%), extremadamente baja. Este cálculo, aparentemente complejo, se simplifica mucho utilizando nomogramas desarrollados a tal efecto[46].

Una de las ventajas de los CP es que si la prueba tiene más de 2 resultados posibles, se puede calcular un CP para cada uno de ellos, permitiéndonos interpretar la contribución al diagnóstico de cada resultado. Otra de las ventajas radica en que los CP facilitan el cálculo de las modificaciones de probabilidad obtenidas al aplicar en serie varias pruebas diagnósticas, recurso frecuentemente empleado en la práctica clínica y en los estudios de análisis de decisión.

Cuando la prueba que se evalúa se mide en una escala cuantitativa, la sensibilidad y la especificidad varían en función del punto de corte que se utilice para separar los valores normales de los anormales. En estos casos, los resultados pueden representarse gráficamente como una **curva ROC** (iniciales del término inglés original *receiver operating characteristics*), que permite conocer las características de la prueba según diferentes puntos de corte y que puede utilizarse para elegir el más adecuado[47-50].

Al igual que en otros tipos de estudios, la valoración de la validez de las pruebas diagnósticas se hace sobre muestras, por lo que los resultados obtenidos son sólo estimaciones puntuales, sujetas a variabilidad aleatoria, y por lo

tanto deben proporcionarse con sus **intervalos de confianza**. Estos intervalos de confianza tendrán que ser aplicados en el cálculo de la probabilidad postprueba para poder juzgar la utilidad de la prueba diagnóstica.

¿Son aplicables en tu medio?

A la hora de valorar la aplicabilidad de la prueba diagnóstica en nuestros pacientes tenemos que considerar si la prueba es suficientemente reproducible, independientemente del ámbito y condiciones de aplicación, y de la persona que la interprete. Por ello, resulta importante que los trabajos publicados incluyan información expresa sobre la **reproducibilidad** de la prueba y los criterios de interpretación de la misma.

Otro aspecto a tener en cuenta es si resulta razonable asumir que las características operativas de la prueba diagnóstica, estimadas en el estudio, van a ser similares en nuestros pacientes. Si el ámbito en el que se ha evaluado la prueba es similar al nuestro, nuestros pacientes cumplen los criterios de inclusión del estudio y no violan los de exclusión, parece sensato aceptar la aplicabilidad de los resultados. Si el **espectro de pacientes** incluido en el estudio es diferente al nuestro la decisión deberá ser tomada con cautela.

La utilidad clínica de la prueba depende también de la **repercusión** que tengan **sobre nuestra actitud diagnóstica** y **terapéutica**. En ocasiones la probabilidad preprueba de nuestro paciente será tan baja, que independientemente cuál sea el resultado de la prueba, la probabilidad postprueba será igualmente baja y por lo tanto no merecerá la pena llevarla a cabo (umbral diagnóstico). En el otro extremo, si la probabilidad preprueba es muy elevada, su resultado no va a modificar nuestra decisión de tratar (umbral terapéutico), por lo que a veces podremos obviarla. En la zona intermedia, donde más interés debería tener la prueba, el grado de información que nos aporta dependerá de la magnitud de los cocientes de probabilidades; valores cercanos a 1 resultarán poco útiles, mientras que valores lejanos modificarán de forma importante las probabilidades postprueba y su rendimiento diagnóstico.

Pero el criterio último de la utilidad de una prueba, al margen de que ésta ofrezca información diagnóstica no disponible previamente o de que modifique nuestro comportamiento clínico, es si el paciente obtiene algún beneficio. Existen escenarios en los que la prueba diagnóstica no resulta coste-efectiva, conlleva riesgos, o conduce a decisiones terapéuticas sin repercusión sobre el paciente.

Papel de los Talleres CASP y de los Archivos de Temas Valorados Críticamente (CAT banks) en el desarrollo de la valoración crítica

Existen varias iniciativas que han permitido desarrollar y difundir la incorporación de la valoración crítica de documentos científicos al ejercicio de la medicina. Las más reseñables son el programa CASP y los Archivos de Temas Valorados Críticamente (*CAT banks*).

Programa CASP

El *Critical Appraisal Skills Programme* (CASP), o Programa de Habilidades en Lectura Crítica, pretende dar respuesta a uno de los problemas que dificultan la práctica de la MBE: la falta de conocimiento para realizar una lectura crítica. Tiene por lo tanto como misión la enseñanza de habilidades para buscar de forma eficiente las pruebas científicas y para evaluar críticamente las pruebas encontradas.

Nació en Gran Bretaña, desde donde se ha ido extendiendo a otros países, formando actualmente una organización llamada CASP internacional[51] dentro de la cual tiene un papel muy activo el grupo español CASPe[52]. El programa CASP colabora con el *Centre for Evidence-Based Medicine*[53], Centro de la Medicina Basada en la Evidencia de la Universidad de Oxford, que enseña a los clínicos cómo tomar decisiones, basadas en la evidencia, sobre un paciente concreto.

La actividad pedagógica del programa CASP se desarrolla a través de talleres en los que se enseñan, en un entorno multidisciplinar y participativo, habilidades para hacer lectura crítica. En estos talleres se trabaja sobre un escenario concreto, que aborda una pregunta surgida de la práctica clínica, que se trata de contestar leyendo y criticando el mejor artículo publicado disponible.

La valoración crítica de los artículos se realiza siguiendo una sistemática, cuya estructurada básica responde a tres aspectos generales, que se concretan en las tres preguntas propuestas por el *Evidence Based Medicine Working Group*[3]: ¿son válidos sus resultados? ¿cuáles son los resultados? ¿son aplicables en tu medio?

Para abordar estas preguntas el grupo CASPe propone en su página web[52] unas prácticas y sencillas guías, adap-

tadas de las publicadas por el *Evidence Based Medicine Working Group* en la revista JAMA[4-7,9,10,17] que facilitan la valoración de: ensayos clínicos, revisiones sistemáticas, artículos sobre pruebas diagnósticas y sobre evaluación económica.

*Archivos de Temas Valorados Críticamente (*CAT banks*)*

Los archivos de Temas Valorados Críticamente (TVC), o *CAT banks* en inglés (siglas de *Critically Appraissed Topics*), nacieron en las Universidades de Mc Master y Oxford ante la necesidad de archivar y clasificar las respuestas a las preguntas que se generaban en la práctica clínica y que eran abordadas con metodología de la MBE.

Un TVC es un documento de extensión corta en el que se detalla una respuesta válida y relevante a una pregunta clínica. Su estructura es la siguiente[2,54]:

— Título claro e informativo del contenido del TVC.
— Especificación de la pregunta clínica a responder.
— La estrategia de búsqueda que se ha seguido para localizarlo y la base o bases de datos utilizadas.
— El artículo científico (válido y clínicamente importante) que mejor la responde.
— Un breve resumen de los resultados clínicamente importantes que contribuyen a responder la pregunta.
— Un apartado de comentarios en donde se realizan algunas puntualizaciones sobre el diseño del estudio y su aplicabilidad al medio del profesional que ha realizado la pregunta.
— En ocasiones, bibliografía auxiliar que sirva para complementar los comentarios.

Un TVC no ha de ser interpretado como la única respuesta existente a una pregunta, puesto que no siempre es el producto de una búsqueda bibliográfica exhaustiva; por lo tanto, está abierto a las oportunas modificaciones que vayan apareciendo sobre el tema. La mayoría de los TVC resumen las evidencias de una única investigación, generalmente un ensayo clínico en el que se evalúa una intervención terapéutica o preventiva, pero también existen TVC sobre revisiones sistemáticas, estudios de evaluación de pruebas diagnósticas, pronóstico de enfermedades, etiología, etc.

Los TVC son confeccionados habitualmente por médicos a título individual, aunque su valor se multiplica cuando son elaborados por grupos de trabajo, en reuniones científicas, sesiones de equipos clínicos o por clubes de revistas.

La elaboración de los TVC tiene mayor valor formativo que la mera lectura de aquellos que han sido realizados por otros. Por esto, aunque se han creado varios bancos de TVC promovidos por instituciones universitarias y asociaciones científicas, accesibles en línea a través de internet[55-60], éstos tienen únicamente un valor referencial y como punto de inicio para actualizar conceptos acerca de temas clínicos.

Conclusión

En esta exposición hemos repasado los principios generales de la valoración crítica de la literatura científica. También hemos revisado los principales criterios a considerar en la valoración de la validez y aplicabilidad de los estudios de evaluación de intervenciones sanitarias y de pruebas diagnósticas.

Hemos podido ver cómo la MBE nos ofrece herramientas de gran utilidad para poder resolver problemas clínicos mediante el análisis eficiente de la literatura científica. Herramientas metodológicas que, si adquirimos y ejercitamos, nos ayudarán a valorar cualquier evidencia científica y a integrarla con nuestros conocimientos y experiencia clínica, para poder decidir sobre su aplicabilidad e idoneidad en un paciente concreto.

La valoración crítica resultará sin duda beneficiosa para nuestros pacientes, pero requiere cierto aprendizaje y sobre todo un cambio de actitud que permita superar las dificultades iniciales. Experiencias como la del programa CASP o la de los archivos de Temas Valorados Críticamente nos pueden ayudar a iniciarnos en este camino.

BIBLIOGRAFÍA

1. Evidence-Based Medicine Working Group. Evidence-Based Medicine. A new approach to teaching the practice of medicine. *JAMA* 1992; **268**: 2420-5.
2. Sackett DI, Richardson WS, Rosenberg W, Haynes RB. Medicina Basada en la Evidencia. Cómo ejercer y enseñar la MBE. Ed. Churchill Livinnstone, Madrid, 1997.
3. Oxman AD, Sackett DL, Guyatt GH. User's guides to the medical literature. I. How to get started. *JAMA* 1993; **270**: 2093-5.
4. Guyatt GH, Sackett DL, Cook DJ. User's guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993; **270**: 2598-601.

5. Guyatt GH, Sackett DL, Cook DJ. User's guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* 1994; **271**: 59-63.

6. Jaeschke R, Guyatt G, Sackett DL. User's guides to the medical literature. III. How to use an article about a diagnostic test. B. Are the results of the study valid? *JAMA* 1994; **271**: 389-91.

7. Jaeschke R, Gordon H, Guyatt G, Sackett DL. User's guides to the medical literature. III. How to use an article about a diagnostic test. B. what are the results and will they help me in caring for my patients? *JAMA* 1994; **271**: 703-7.

8. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. User's guides to the medical literature. IV. How to use an article about harm. *JAMA* 1994; **27**: 1615-9.

9. Laupacis A, Wells G, Richardson S, Tugwell P. User's guides to the medical literature. V. How to use an article about prognosis. *JAMA* 1994; **272**: 234-7.

10. Oxman AD, Cook DJ, Guyatt GH. User's guides to the medical literature. VI How to use an overview. *JAMA* 1994; **272**: 1367-71.

11. Richardson WS , Detsky AS. User's guides to the medical literature. VII. How to use a Clinical Decision Analysis. A. Are the results of the study valid? *JAMA* 1995; **273**: 1292-5.

12. Richardson WS, Detsky AS. User's guides to the medical literature. VII. How to use a Clinical Decision Analysis. B. What are the results and will they help me in caring for my patients? *JAMA* 1995; **273**: 1610-3.

13. Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt GH. User's guides to the medical literature. VIII Clinical practice guidelines (A). Are the recommendations valid? *JAMA* 1995; 274: 570-4.

14. Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt GH. User's guides to the medical literature. VIII Clinical practice guidelines (B). What are the recommendations and will they help you in caring for your patient? *JAMA* 1995; **274**: 1630-2.

15. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. User's guides to the medical literature. IX. A method for grading health care recomendations. *JAMA* 1995; **274**: 1800-4.

16. Naylor CD , Guyatt GH. User's guides to the medical literature. X. How to use an article reporting variations in the outcomes of health services. *JAMA* 1996; **275**: 554-8.

17. Naylor CD, Guyatt GH. User's guides to the medical literature. XI. How to use an article about a clinical utilization review. *JAMA* 1996; **275**: 1435-9.

18. Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. User's guides to the medical literature. XII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. *JAMA* 1997; **277**: 1232-7.

19. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D. User's guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 1997; **277**: 1552-7.

20. O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF. User's guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA* 1997; **277**: 1802-6.

21. Dans AL, Dans LF, Guyatt GH, Richardson S. User's guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. *JAMA* 1998; **279**: 545-9.

22. Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J. User's guides to the medical literature: XV. How to use an article about disease probability for differential diagnosis. Evidence-Based Medicine Working Group. *JAMA* 1999; **281**: 1214-9.

23. Guyatt GH, Sinclair J, Cook DJ, Glasziou P. User's guides to the medical literature: XVI. How to use a treatment recommendation. Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group. *JAMA* 1999; **281**: 1836-43.

24. Barratt A, Irwig L, Glasziou P, Cumming RG, Raffle A, Hicks N, et al. User's guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. *JAMA* 1999; **281**: 2029-34.

25. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. User's guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. *JAMA* 1999; **282**: 771-8.

26. McAlister FA, Laupacis A, Wells GA, Sackett DL. User's Guides to the Medical Literature: XIX. Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect. *JAMA* 1999; **282**: 1371-7.

27. Hunt DL, Jaeschke R, McKibbon KA. User's guides to the medical literature: XXI. Using electronic health information resources in evidence-based practice. Evidence-Based Medicine Working Group. *JAMA* 2000; **283**: 1875-9.

28. McAlister FA, Straus SE, Guyatt GH, Haynes RB. User's guides to the medical literature: XX. Integrating research evidence with the care of the individual patient. Evidence-Based Medicine Working Group. *JAMA* 2000; **283**: 2829-36.

29. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. User's guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA* 2000; **284**: 79-84.

30. Giacomini MK, Cook DJ. User's guides to the medical literature: XXIII. Qualitative research in health care A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 2000; **284**: 357-62.

31. Giacomini MK, Cook DJ. User's guides to the medical literature: XXIII. Qualitative research in health care B. What are the results and how do they help me care for my patients? Evidence-Based Medicine Working Group. *JAMA* 2000; **284**: 478-82.

32. Richardson WS, Wilson MC, Williams JW Jr, Moyer VA, Naylor CD. User's guides to the medical literature: XXIV. How to use an article on the clinical manifestations of disease. Evidence-Based Medicine Working Group. *JAMA* 2000; **284**: 869-75.

33. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996; **313**: 36-9.

34. Brozanski BS, Jones JG, Gilmore CH, Balsan MJ, Vázquez RL, Israel BA, et al. Effect of pulse dexamethasone therapy on the incidence and severity of chronic lung disease in the very low birthweight infant. *J Pediatr* 1995; **126**: 769-76.

35. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; **292**: 46-50.

36. Subhedar NV, Shaw NJ. Changes in oxygenation and pulmonary haemodynamics in preterm infants treated with inhaled nitric oxide. *Arch Dis Child Fetal Neonatal Ed* 1997; **77**: 191-7.

37. Saugstad OD, Roowelt T, Aalen O. Resuscitation of asphyxiated newborn infants with room air or oxygen: an international controlled trial: the Resair 2 study. *Pediatrics* 1998; **102** (1): el.

38. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992; **116**: 78-84.

39. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988; **318**: 1728-33.

40. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Second Edition. Clinical epidemiology, a basic science for clinical medicine.(1991) Second Edition.Boston / Toronto: Little, Brown and Company.

41. Feinstein AR. Diagnostic and spectral markers. En: Feinstein AR, editor. Clinical Epidemiology. The architecture of clinical research. Filadelfia: WB Saunders, 1985. p. 597-631.

42. Ochoa Sangrador C, Brezmes Valdivieso MF. Efectividad de los test diagnósticos. *An Esp Pediatr* 1995; **42** (6): 473-5.

43. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; **39**: 207-15.

44. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992; **45**: 581-6.

45. Lohr JA, Portilla MG, Geuder TG, Dunn ML, Dudley SM. Making a presumptive diagnosis of urinary tract infection by using a urinalysis performed in an on-site laboratory. *J Pediatr* 1993; **122**: 22-5.

46. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med* 1975; **293**: 257.

47. Moise A, Clément B, Ducimetière P, Bourassa MG. Comparison of Receiver Operating Curves Derived from the Same Population: A Bootstrapping Approach. *Comput Biomed Re*s 1985; **18**: 125-31.

48. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; **39**: 561-77.

49. Guangqin MA, Hall WJ. Confidence Bands for Receiver Operating Characteristics Curves. *Med Decis Making* 1993; **13**: 191-7.

50. Centor RM, Keightley GE. Receiver operating characteristic (ROC) curve area analysis using the ROC ANALYZER. SCAMC Proc 1989; 222-226.

51. Critical Appraisal Skills Programme. CASP internacional network. [en línea] [fecha de acceso 16 de marzo de 2002] URL disponible en: http: / / www.caspinternational.org.uk /

52. Programa de habilidades en lectura crítica - España. CASPe [en línea] [fecha de acceso 27 de junio de 2002] URL disponible en: http: / / www.redcaspe.org

53. Centre for Evidence-Based Medicine [en línea] [fecha de acceso 16 de marzo de 2002] URL disponible en: http: / / cebm.jr2.ox.ac.uk /

54. Álvarez S. Dp / Doc. Formación continuada orientada a problemas [en línea] [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / usuarios.bitmailer.com / rafabravo / DpDoc.html y http: / / www.aepap.org / pedev / dpdoc.htm

55. CAT Bank [en línea] Michigan University [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / www.ped.med.umich.edu / ebm / cat.htm

56. CAT Bank [en línea] Washington University [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / depts.washington.edu / pedebm /

57. CAT Bank [en línea] Rochester University [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / www.urmc.rochester.edu / medicine / res / CATS / ped.html

58. CAT Bank [en línea] North Carolina University [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / www.med.unc.edu / medicine / edursrc / !catlist.htm

59. The PedsCCM Evidence- Based Journal Club [en línea] Web Pediatric Critical Care Medicine [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / pedsccm.wustl.edu / EBJournal_Club.html

60. Archivo de Temas Valorados Críticamente [en línea] web de la AEPap [fecha de acceso 16 de marzo de 2002]. URL disponible en: http: / / www.aepap.org / pedev / pedev-4.htm