# Evaluación de las pruebas diagnósticas

**C. Carnero-Pardo**

La sostenibilidad es uno de los mayores desafíos a los que se enfrentan los servicios de salud públicos; resulta difícil hacer compatibles unos recursos limitados con unas demandas crecientes y unos servicios cada vez más caros; se explica así que una de las metas más universalmente perseguidas por estas organizaciones sea la eficiencia a todos los niveles, y fundamentalmente asegurar la eficiencia de las nuevas ofertas que se incorporan a la cartera de servicios. Quizás sea esta una de las explicaciones del auge y desarrollo de una corriente como la medicina basada en la evidencia cuyos postulados y procedimientos, no sólo han calado en la práctica clínica, sino también en la salud pública, la gestión o en la política sanitaria [1].

En línea con esta situación, desde hace mucho tiempo, la incorporación a la práctica clínica de un nuevo tratamiento farmacológico o intervención terapéutica se precede de la documentación fehaciente de su eficacia a través de un cuidado y detallado proceso en el que se suceden estudios de complejidad creciente y que pretenden constatar más allá de toda duda, no sólo su eficacia, sino también su tolerabilidad [2]. Incluso, cada día más, las agencias reguladoras no sólo exigen el rigor metodológico derivado de este proceso, sino también estudios de corte económico, que trasciendan de la eficacia de la intervención y orienten sobre su eficiencia.

No ha ocurrido igual, en cambio, con las pruebas diagnósticas, un campo muy dinámico y activo debido al gran desarrollo tecnológico habido en las últimas décadas. La incorporación de nuevas pruebas diagnósticas a la práctica clínica no ha tenido este razonable control, a pesar de que suponen una parte importante del gasto sanitario. Recientes estudios demuestran que, incluso en las revistas más prestigiosas, los trabajos referidos a pruebas diagnósticas adolecen de importantes déficit metodológicos con sesgos que tienden a sobreestimar la validez de las mismas [3]; quizás por ello, múltiples técnicas diagnósticas, en muchos casos muy costosas, se han sumado a la práctica habitual o se recomienda su incorporación a ella, a pesar de que las evidencias disponibles sobre su eficiencia y eficacia sean más que dudosas o al menos cuestionables [4].

Dos iniciativas muy recientes tratan de corregir esta situación; por un lado, varios autores muy prestigiosos e influyentes [5-7] abogan porque la validación de las pruebas diagnósticas sea un proceso conceptual y estructuralmente paralelo al proceso de evaluación de fármacos, en el que estudios consecutivos de complejidad creciente controlen de forma progresiva los distintos sesgos que pueden afectar a esta validación y procuren objetivos más pragmáticos y aplicados; aunque no coinciden en la denominación de las distintas fases (Tabla I), éstas básicamente consistirían en:

– *Estudio preliminar o exploratorio.* Se trataría de un estudio transversal o caso-control con muestra de conveniencia que tendría por objeto documentar que los resultados de la prueba diagnóstica son distintos en sujetos sin y con el proceso a diagnosticar. En el estudio no se incluirían casos dudosos o complicados. Se trata de un estudio inicial que se suele hacer antes de abordar otros estudios más costosos. Dos ejemplos de este tipo de estudios se han publicado recientemente en *Revista de Neurología*; se trata del estudio preliminar del EUROTEST [8] y del test de las fotos [9], en los que se describen dos nuevos instrumentos de cribado y despistaje de deterioro cognitivo y demencia diseñados específicamente para poder aplicarse en analfabetos.
– *Estudio de validación retrospectiva.* Se trataría de un estudio de diseño transversal con el mismo objetivo que el estudio preliminar; pero, en este caso, la muestra incluye una representación adecuada del espectro del proceso a diagnosticar, con inclusión de casos dudosos y distintos estadios del proceso y, además, en la misma proporción que aparece en las condiciones en las que teóricamente se va aplicar el test. Este diseño puede evaluar la capacidad discriminativa, pero no predictiva del test. Un estudio de este tipo publicado en *Revista de Neurología* hace algunos años, evaluaba la validez del test de fluencia verbal semántica para la identificación de demencia en pacientes neurológicos [10].
– *Estudio de validación prospectiva.* Este estudio se lleva a cabo en sujetos todavía sin diagnosticar en los que se plantea el diagnóstico en cuestión, constituyendo pues una cohorte que, tras la realización del test a validar, se sigue hasta alcanzar el diagnóstico definitivo por otros medios; permite, pues, evaluar adecuadamente la capacidad predictiva-diagnóstica del test. Este carácter prospectivo y el hecho de aplicar el test antes de realizar el diagnóstico, permite además el control de los principales sesgos, pudiendo asegurar tanto la evaluación independiente y ciega del test a validar y del test diagnóstico con el que se compara, como que todos los sujetos, independientemente de sus resultados, se sometan al proceso diagnóstico completo. Un excelente ejemplo de este tipo de estudio se publica en este número de *Revista de Neurología* [11]; Pérez-Martínez et al, del Hospital de la Cruz Roja de Madrid, llevan a cabo un esmerado y riguroso estudio para evaluar la validez de una versión española del MIS [12], un test que evalúa el recuerdo libre y facilitado de cuatro palabras y que ha mostrado una gran validez para identificar sujetos con demencia, y en especial, enfermedad de Alzheimer. Los resultados de esta versión

*Aceptado: 28.04.05.*

*Sección de Neurología. Hospital Torrecárdenas. Almería, España.*

*Correspondencia: Dr. Cristóbal Carnero Pardo. Prof. Agustín Escribano, 10, 5.º B-1. E-18004 Granada. E-mail: ccarnerop@supercable.es*

*Conflicto de intereses. C. Carnero Pardo es el creador y titular del registro del EUROTEST y del test de las fotos.*

**Tabla I.** Tipos de estudios para validación de pruebas diagnósticas.

| Objetivos | Diseño | Sacket et al [5] | Pepe [6] | Gluud et al [7] |
|---|---|---|---|---|
| Exploratorio/preliminar | Caso-control frente a transversal | Fase I | Fase I | Fase IIa |
| Validación retrospectiva | Transversal | Fase II | Fases II y III | Fase IIb |
| Validación prospectiva | Cohorte | Fase III | Fase IV | Fase IIc |
| Estudio de impacto | Ensayo clínico controlado | Fase IV | Fase V | Fase III |
| Normalización | Transversal | | | Fase I |
| Evaluación de la fiabilidad | Test-retest evaluadores múltiples | | | |

española son similares a los encontrados en la versión original, configurándose pues como un buen instrumento, no sólo por su validez, sino también por su facilidad y rapidez; su inconveniente es que no puede aplicarse a analfabetos y que evalúa sólo memoria.

– *Evaluación del impacto y utilidad.* Al igual que en el caso de los estudios con medicamentos, la utilidad real debe evaluarse mediante un ensayo clínico controlado; en él, el test a evaluar se compara con el procedimiento diagnóstico habitual en sujetos en los que se plantea el diagnóstico en cuestión; la clave diferencial es que la asignación a un procedimiento diagnóstico u otro se realiza aleatoriamente y los resultados se evalúan en términos de resultados en salud (calidad de vida, costes, etc.). Escasean los estudios de este tipo [13], no sólo por su dificultad y complejidad, sino también porque no siempre es posible llevarlos a cabo –diagnóstico de procesos benignos sin tratamiento–; en otras ocasiones los resultados son obvios y en otras, en fin, los resultados tan sólo se pueden valorar después de un control clínico muy prolongado –test para el diagnóstico en fases iniciales de procesos de larga duración–.

El proceso de validación debe incluir otros estudios complementarios que evalúen los resultados de las pruebas diagnósticas en sujetos normales y la posible influencia en los mismos de variables como edad, sexo, etc. (estudios de normalización) u otras características del instrumento como la fiabilidad (test-retest o interobservador).

La segunda iniciativa parte de un grupo de trabajo de la Colaboración Cochrane y tiene como fin mejorar la calidad de los artículos que versan sobre evaluación de pruebas diagnósticas. Para ello, este grupo recomienda que los artículos que se publiquen reúnan unos STARD (*Standards for Reporting of Diagnostic Accuracy*), y proponen un listado de 25 elementos de información referidos a elementos del diseño del estudio, del test en evaluación y de sus resultados, que resultan esenciales para que el lector identifique posibles sesgos del estudio (atentados a la validez interna) y pueda valorar en su justa medida la aplicabilidad y posibilidad de generalizar los resultados (evaluación de la validez externa) (Tabla II) [14]. Este listado también proporciona un método para medir la calidad de los distintos estudios y, de hecho, se ha utilizado con este fin en algunas revisiones sistemáticas [4]. La iniciativa también recomienda a

**Tabla II.** Listado de elementos a incluir en estudios sobre evaluación de pruebas diagnósticas (iniciativa STARD).

| Sección | | Elemento | Descripción |
|---|---|---|---|
| Título/Resumen/ Palabras clave | | 1 | El artículo se identifica como un estudio sobre validez diagnóstica (VD) |
| Introducción | | 2 | Consta que el objetivo es estimar la VD o compararla con la de otro test |
| Métodos | Participantes | 3 | Describe la población estudiada y los criterios de selección y exclusión |
| | | 4 | Describe el criterio de reclutamiento: síntomas, otros tests, hacer el test actual |
| | | 5 | Describe el muestreo: consecutivo, aleatorio, no especificado |
| | | 6 | Describe la recogida de datos: prospectivo, retrospectivo |
| | Test | 7 | Describe el estándar de referencia y su justificación |
| | | 8 | Describe o referencia las especificaciones técnicas y cómo tomar las medidas |
| | | 9 | Describe las unidades de medida, puntos de corte o categorías de los resultados |
| | | 10 | Describe el número, entrenamiento y experiencia de los que realizan los tests |
| | | 11 | Describe si la aplicación de la prueba y del test de referencia se realizó de forma ciega |
| | Estadística | 12 | Describe los métodos para calcular o comparar la VD y el grado de incertidumbre |
| | | 13 | Describe si se ha calculado la fiabilidad y cómo se ha hecho |
| Resultados | Participantes | 14 | Informa cuándo se realizó el estudio (fecha de inicio y final del reclutamiento) |
| | | 15 | Informa de los caracteres demográficos y clínicos |
| | | 16 | Informa de las razones para las pérdidas (diagrama de flujo) |
| | Resultados | 17 | Informa del tiempo y circunstancias (tratamiento) entre el test y el estándar |
| | | 18 | Informa del grado de afectación de los enfermos y del diagnóstico de los no enfermos |
| | | 19 | En caso de valores discretos, hay tabla 2 × 2 con los resultados<br>En caso de valores continuos, se informa de la distribución<br>Se describen los resultados perdidos o indeterminados |
| | | 20 | Se describen los efectos adversos del test y del estándar |
| | Estimación | 21 | Se describe la VD y su incertidumbre estadística (p. ej., IC 95%) |
| | | 22 | Informa cómo se han manejado los perdidos, indeterminados o extremos |
| | | 23 | Informa de la variabilidad de la VD entre evaluadores, centros o subgrupos |
| | | 24 | Informa sobre la fiabilidad en el caso de que se haya evaluado |
| Discusión | | 25 | Discute la aplicabilidad clínica de los hallazgos del estudio |

los autores la elaboración de un diagrama de flujo que resuma y esquematice el diseño del estudio y el flujo de los sujetos participantes entre las distintas fases del mismo, por entender que una gráfica de este tipo puede transmitir de forma fácil y transparente una información esencial para el lector.

Esta iniciativa ha tenido una rápida y generalizada aceptación y se ha adoptado por las principales revistas biomédicas, entre ellas *Neurology*, de forma que los artículos sobre evaluación de pruebas diagnósticas que pretendan publicarse en éstas, deben contener información explícita relativa a estos elementos, así como un diagrama de flujo que debería ser la figura 1 del artículo.

En definitiva, es conveniente, pues, que los autores, revisores y lectores tengan muy presentes al valorar los artículos sobre evaluación de pruebas diagnósticas el tipo de estudio, el diseño y los objetivos del mismo, así como el control que se ha realizado de los posibles sesgos, con objeto de que las conclusiones y la valoración sean adecuadas y proporcionadas a los resultados y a las circunstancias en las que se han obtenido; por otro lado, sería recomendable la adopción de la iniciativa STARD por parte de los autores, revisores y el nuevo equipo editorial de *Revista de Neurología*, lo que es de esperar que redunde en mejorar la calidad de los artículos de este tipo que en adelante se publiquen en ésta, y en facilitar a los lectores la lectura crítica de los mismos.

## BIBLIOGRAFÍA


1. Muir-Gray JA. Evidence-based healthcare. How to make health policy and management decisions. 2 ed. Edinburgh: Churchill-Livingstone; 2001.
2. Du Souich P, Orme M, Erill S. The IUPHAR compendium of basic principles for pharmacological research in humans. Irvine: IUPHAR; 2004.
3. Lijmer J, Mol B, Heisterkamp S, Bonsel G, Prins M, Van der Meulen J, et al. Empirical evidence of design-related bias in studies of diagnostic test. JAMA 1999; 282: 1061-6.
4. Carnero-Pardo C. Revisión sistemática sobre la utilidad de la tomografía por emisión de positrones en el diagnóstico de la enfermedad de Alzheimer. Rev Neurol 2003; 37: 860-70.
5. Sackett DL, Haynes RB. Evidence base of clinical diagnosis: The architecture of diagnostic research. BMJ 2002; 324: 539-41.
6. Pepe MS. The statistical evaluation of medical test for classification and prediction. New York: Oxford University Press; 2003.
7. Gluud C, Gluud LL. Evidence based diagnostic. BMJ 2005; 330: 724-6.
8. Carnero-Pardo C, Montoro-Ríos M. Evaluación preliminar de un nuevo test de cribado de demencia (EUROTEST). Rev Neurol 2004; 38: 201-9.
9. Carnero-Pardo C, Montoro-Ríos M. El test de las fotos. Rev Neurol 2004; 39: 801-6.
10. Carnero C, Lendínez A. Utilidad del test de fluencia verbal semántica en el diagnóstico de demencia. Rev Neurol 1999; 29: 709-14.
11. Pérez-Martínez DA, Baztán JJ, González-Becerra M, Socorro A. Evaluación de la utilidad diagnóstica de una adaptación española del *Memory Impairment Screen* de Buschke para detectar demencia y deterioro cognitivo. Rev Neurol 2005; 40: 644-8.
12. Buschke H, Kuslansky G, Katz M, Stewart WF, Sliwinski MJ, Eckholdt HM, et al. Screening for dementia with the memory impairment screen. Neurology 1999; 52: 231-8.
13. Troughton R, Frampton C, Yandle T, Espiner E, Nicholls M, Richards A. Treatment of heart failure guided by plasma aminoterminal brain natriuretic peptide (N-BNP) concentrations. Lancet 2000; 355: 1126-30.
14. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326: 41-4.