ARTÍCULOS ORIGINALES/ORIGINAL ARTICLES

PREDICCIÓN EN LA EVALUACIÓN NEUROPSICOLÓGICA CLÍNICA: UNA APROXIMACIÓN CUANTITATIVA.

PREDICTION IN CLINICAL NEUROPSYCHOLOGICAL ASSESSMENT: A QUANTITATIVE APPROACH.

Alonso Ortega G.[1], Walter Lips C.[2]

Resumen

En primer lugar, esta revisión discute la exactitud predictiva de las aproximaciones clínica y estadística en el contexto de la neuropsicología aplicada. Asimismo, provee evidencia que permite comparar ambos métodos predictivos, destacando los beneficios de la predicción estadística en contextos clínicos. En la segunda sección, se describen diferentes estadísticos de eficiencia diagnóstica, con ejemplos que sustentan su uso. Finalmente, en la discusión se resumen ambos enfoques dando especial énfasis a algunos aspectos clínicos, pragmáticos y éticos que podrían traer potenciales efectos en la vida cotidiana de los pacientes.
Palabras Claves: Predicción clínica, predicción estadística, neuropsicología clínica.

Summary

This review firstly discusses the predictive accuracy of both, the clinical and actuarial approaches in the context of applied neuropsychology. It also provides evidence that allows to compare both predictives methods, highlihting the benefits of using actuarial predictive procedures in clinical settings. In the second section, different Diagnostic Efficiency Statistics are presented and explained. Some examples are given. Finally, the discussion summarizes both approaches with an especial emphasis on some clinical, pragmatic and ethical aspects that may have some potencial effects in patients' everyday life.
Keywords: Clinical prediction, Actuarial Prediction, Clinical Neuropsychology.

1 Psicólogo. Académico, Escuela de Psicología, Facultad de Medicina, Universidad de Valparaíso, Chile. Departamento de Psicología Fisiológica, Universität Bielefeld, Alemania.
2 Médico Psiquiatra. Centro de Salud Mental y Psiquiatría Comunitaria (COSAM), Concón, Chile. Académico, Escuela de Psicología, Facultad de Medicina, Universidad de Valparaíso, Chile.
Recibido: Enero 2012.
Aceptado: Abril 1012.

Introducción

El origen de la neuropsicología moderna se remonta a los siglos XIX y XX, asociado frecuentemente a los trabajos de Broca, Vigotsky y Luria(1, 2). No obstante, la Asociación Americana de Psicología (APA) ha reconocido a la neuropsicología clínica como una subespecialidad de la psicología, hace menos de dos décadas[3]. Una definición reciente de la Asociación Americana de Neuropsicología Clínica (AACN) la entiende como "una ciencia aplicada que examina el impacto del funcionamiento cerebral normal y anormal sobre un amplio rango de funciones cognitivas, emocionales y conductuales (p.211)"(3). A partir de tal definición, el neuropsicólogo es concebido básicamente como un profesional psicólogo con conocimientos específicos en neurociencias y las bases neurológicas del comportamiento.

El prefijo "neuro" supone conocimientos especiales de las relaciones cerebro-conducta(4) pero, en tanto "psicólogo", implica manejo en aquellas competencias propias de la profesión. Entre ellas, cobra especial relevancia el entrenamiento en la utilización de instrumentos de medición psicológica. Su uso apropiado puede contribuir al establecimiento de hipótesis que guíen las estrategias terapéuticas a seguir. No obstante, en la práctica clínica, rara vez contamos con un único test que constituya el estándar de referencia (i.e., gold standard) para obtener un diagnóstico de manera inequívoca. Quien realiza una evaluación cuenta además con una gran cantidad de información que, contextualizada al paciente y su entorno, le permitirá emitir juicios clínicos que forman parte importante del proceso diagnóstico.

Sin embargo, a menudo tales juicios se ven afectados por la manera en que el clínico maneja e interpreta la información. Por ello, el proceso diagnóstico no está exento de la influencia de sesgos. Tales inclinaciones pueden provenir de simples errores de percepción o interpretación de la información, así como de la formación académica o de la experiencia clínica personal. Desconocer la existencia de sesgos en los procesos de toma de decisiones sería prácticamente desconocer la naturaleza de nuestro sistema cognitivo. Al respecto, Paul E. Meehl argumenta que "es absurdo, así como arrogante, pretender que el adquirir el grado académico de doctor nos inmuniza de alguna manera de cometer errores de muestreo, percepción, memorización, retención, recuperación e inferencia, que se suponen propios de la mente humana" (p.728)(5).

Considerando lo anterior, no resulta ilógico preguntarse: ¿qué tan acertadas resultan ser las predicciones basadas en juicios clínicos? Este tema ha sido abordado con anterioridad y, hasta la fecha, aún genera controversia. Sin embargo, más que alimentar el debate, el objetivo de la presente revisión es proveer evidencia que respalda el uso de métodos cuantitativos formales para mejorar la precisión y exactitud al establecer hipótesis diagnósticas. Por ello, en primer lugar, abordaremos el rol predictivo de las aproximaciones clínica vs. estadística, dentro del proceso diagnóstico. En segundo lugar, propondremos el uso de algunos algoritmos cuantitativos (i.e., estadísticos) que pueden contribuir a mejorar la capacidad predictiva en la práctica clínica cotidiana. Finalmente, presentaremos algunas consideraciones generales respecto de la adopción de una aproximación cuantitativa como complemento a los procedimientos diagnósticos.

Predicción clínica vs. estadística

Millis plantea la pregunta diagnóstica fundamental que todo clínico quisiera poder contestar: ¿Dado un resultado positivo en un test, cuál es la probabilidad de que el paciente padezca la condición en cuestión?(6) Concluir su presencia luego del resultado positivo en un test podría, ciertamente, simplificar las cosas. No obstante, tal manera de razonar es a menudo equívoca. El resultado positivo de un test constituye evidencia a favor de una hipótesis diagnóstica solamente si es integra-

do con otras piezas de información(6). Por este motivo, el clínico analiza, procesa y complementa información proveniente de diversas fuentes (e.g., observación clínica, historia del paciente, exámenes físicos y de laboratorio, resultados de tests psicológicos, etc.) para luego emitir un juicio clínico que representa su apreciación respecto de una o más hipótesis diagnósticas. Este procedimiento es conocido también como predicción clínica(7) y, tal como lo define Meehl(8), se basa principalmente en "juicios clínicos informales"[1](sic) para arribar a conclusiones diagnósticas.

Entonces, podríamos reformular la pregunta inicial de la siguiente manera: ¿Dado un juicio clínico que apoya cierta hipótesis diagnóstica, cuál es la probabilidad de que el paciente padezca la condición en cuestión? Nuevamente, no podemos dar respuesta a tal interrogante. Al menos no de manera precisa. Al igual que en el caso del resultado positivo de un test, concluir la presencia de una condición únicamente sobre la base de un juicio clínico puede resultar también, a veces, equívoco. Más aún si consideramos la existencia de datos que sugieren que la predicción clínica es poco precisa en comparación con la capacidad predictiva de otros métodos existentes (i.e., estadísticos)(7, 11).

Ya a mediados de la década de los 50, Meehl estableció la distinción entre "predicción clínica vs. estadística"(7). En su libro homónimo, Meehl entiende la predicción estadística como aquella en que la información es sistematizada y cuantificada, de modo tal que no se requiera un juicio profesional posterior para arribar a una conclusión diagnóstica(8). Ello es posible gracias al uso de procedimientos algorítmicos, objetivos y formales (e.g., ecuaciones o fórmulas) que proveen al clínico de índices que poseen interpretaciones estándar y que, por tanto, no requieren lecturas adicionales. De este modo, la utilización de una aproximación predictiva formal y objetiva, aporta mayor precisión y exactitud a los juicios clínicos empleados al elaborar un diagnóstico. Aun cuando ambas aproximaciones predictivas (i.e., clínica y estadística) buscan lo mismo, la diferencia básica radica en "las fuentes o tipos de información empleadas para realizar predicciones y la manera en que esta información es empleada con fines predictivos (p.15)"(7).

Antes de continuar, resulta importante aclarar que un método es simplemente una aproximación, forma o modo de abordar un problema determinado, y la evaluación de sus ventajas y desventajas será incompleta si no consideramos los fines para los cuales será empleado. De ello se desprende que no existen métodos adecuados o inadecuados per se. Por ejemplo, si el fin es describir o comprender la vivencia de una persona, el empleo de un método clínico descriptivo o fenomenológico será, sin duda, más adecuado que el uso de uno estadístico. Bajo la misma lógica, cuando los fines son predictivos, tal vez una aproximación cuantitativa resultaría más adecuada. Una vez establecida la aclaración, revisaremos algunos estudios que avalan la supremacía predictiva de los métodos estadísticos por sobre los clínicos.

Un estudio de White et al.(11) sobre confiabilidad inter-jueces en la evaluación neuropsicológica, reveló la existencia de niveles de correspondencia "moderadamente confiables" al momento de diagnosticar deterioro cognitivo general. El mismo estudio puso en evidencia que los niveles de correspondencia sólo alcanzan el estatus de "razonables" y "buenos" cuando se trata de diagnósticos más específicos. Otro estudio, conducido por Grove y Meehl(9), reporta que de un total de 136 estudios de investigación, en una amplia variedad de dominios predictivos, no más del 5% demostró que la predicción clínica fuera más precisa que la estadística. Un metaanálisis realizado por Grove y colegas(10)

1 En este contexto, la palabra "informal" hace referencia a la no utilización de métodos predictivos "formales" (e.g., estadísticos) para arribar a una conclusión diagnóstica.

sobre estudios de salud y comportamiento, reveló que las técnicas de predicción estadística fueron, en promedio, cerca de un 10% más precisas que las predicciones basadas en juicios clínicos. Dependiendo de los tipos de análisis, las predicciones clínicas fueron superadas por las estadísticas en un rango que va entre el 33% al 47% del total de los estudios examinados. Lo contrario ocurrió solamente entre un 6% y un 16% de los casos. Más aún, las predicciones clínicas funcionaron relativamente peor cuando los predictores incluyeron datos provenientes de entrevistas clínicas.

La superioridad de las técnicas de predicción estadísticas fueron consistentes, independientemente del tipo de datos que se utilizaron para realizar las predicciones, del tipo de jueces e incluso de los años de experiencia de estos últimos. Si consideramos que la experiencia clínica es alcanzada mayormente a través de la praxis, es posible suponer que distintas experiencias clínicas podrían conducir a la adopción de diferentes parámetros para determinar el cumplimiento de criterios[2], al uso de diferentes heurísticos y a la aparición de diversos tipos de sesgos. Por lo tanto, no sería ilógico sostener que la heterogeneidad inherente al método clínico puede contribuir de manera importante al aumento en el error de predicción al elaborar un diagnóstico.

Por lo tanto, el uso de métodos cuantitativos contribuye a reducir la cuota de subjetividad asociada a los métodos predictivos clínicos, mejorando así la exactitud de las predicciones diagnósticas. Desde el punto de vista de la "teoría bayesiana" de toma de decisiones(12), el mejor esquema predictivo será aquel que maximice las utilidades esperadas en aquella persona sobre la cual las predicciones son realizadas(8). Entonces, la elección de un modelo predictivo debería siempre estar orientada a maximizar las utilidades esperadas para el paciente.

En la próxima sección, adoptando una aproximación predictiva cuantitativa (i.e., estadística), revisaremos diferentes estadísticos de eficiencia diagnóstica que pueden resultar útiles a la labor clínica.

Estadísticos de eficiencia diagnóstica

Como señalamos al comienzo de la sección anterior, un resultado positivo en un test no implica directamente la presencia de la condición en cuestión. El resultado de un test es una cosa y su interpretación otra. Precisamente, es en este espacio interpretativo donde se da lugar a la aparición de imprecisiones que afectan la clasificación diagnóstica. Por ello, cobra relevancia la adopción de una aproximación cuantitativa que establezca la presencia de una condición en términos probabilísticos y que, además, proporcione criterios interpretativos uniformes. Al respecto, Millis es enfático al establecer la "absoluta necesidad de utilizar métodos cuantitativos explícitos en el proceso diagnóstico" (p.21)(6).

Desde una aproximación cuantitativa, contamos con algunos índices que aportan exactitud a la clasificación diagnóstica, conocidos genéricamente como estadísticos de eficiencia diagnóstica (EED)[3] (del inglés, Diagnostic Efficiency Statistics) (13). A continuación describiremos algunos EED, proporcionando simples algoritmos estadísticos para su cálculo, junto con las respectivas indicaciones para su correcta interpretación.

Sensibilidad y especificidad (Sensitivity and Specificity)

La sensibilidad es definida como la proporción de personas que poseen una

2 Aun cuando existen criterios unificados para la realización de diagnósticos en salud mental (e.g., DSM-IV-TR, ICD-10), no podemos asegurar que la interpretación de éstos sea completamente uniforme entre todos los profesionales de la salud mental.

3 También referidos en la literatura bajo los términos "índices de precisión clasificatoria" (Classification Accuracy Indices), o "índices de precisión diagnóstica" (Diagnostic Accuracy Indices).

condición y que son detectadas por un test(13). Como contraparte, la especificidad es definida como la proporción de personas que no poseen una condición y que no son detectadas por un test(13). Ambos índices están íntimamente relacionados, por ende su interpretación debe efectuarse siempre en función del otro.

Cada test sugiere un punto de corte ideal para determinar un resultado positivo, el cual estará asociado al mejor equilibrio entre sensibilidad y especificidad. Cada vez que modifiquemos el punto de corte de un test, aumentaremos el nivel de sensibilidad en desmedro de la especificidad y viceversa(14). Aun cuando ambos índices son usualmente vistos como propiedades fijas de un test(15), lo serán sólo si éste es empleado en poblaciones con características similares. Por lo tanto, si el test ha de ser empleado en poblaciones con diferentes características, tales índices deberán ser recalculados(13).

Una creencia ampliamente difundida entre algunos clínicos es que una alta sensibilidad es una característica necesaria de todo test. No obstante, tal razonamiento no siempre es correcto. En algunos contextos es incluso recomendable adoptar una postura más conservadora y privilegiar una alta especificidad a expensas de una disminución en la sensibilidad(16). El equilibrio entre sensibilidad y especificidad dependerá, en parte, del contexto evaluativo y de los propósitos para los cuales el test será empleado.

En general, existen al menos dos consideraciones que deberíamos tomar en cuenta cada vez que utilizamos un test. Primero, cuando el punto de corte está asociado a una alta especificidad, un resultado negativo apoyará el descarte de la condición estudiada. Por ende, un test con alta especificidad nos prestará mayor utilidad para descartar un diagnóstico y no para confirmarlo. Segundo, a la inversa del caso anterior, cuando un punto de corte está asociado a una alta sensibilidad, un resultado positivo en el test nos permitirá confirmar la presencia de una condición, no así descartarla. Obviamente, la condición ideal es encontrar el mejor equilibrio entre sensibilidad y especificidad. Más adelante, enseñaremos una técnica útil para poder identificar el punto de corte que está asociado a un equilibrio óptimo entre ambos índices.

La tabla 1 muestra los EED para el hipotético *MM test*, con las respectivas fórmulas para su cálculo. El estándar de referencia estuvo constituido por 100 pacientes con diagnóstico confirmado de deterioro cognitivo[4]. El *MM test* tiene un total de 30 puntos, donde el máximo puntaje representa un desempeño cognitivo perfecto. El punto de corte ideal para un determinar un resultado positivo (i.e., presencia de deterioro cognitivo) fue establecido en 22 puntos o menos. Con la información proporcionada trataremos de tomar una decisión sobre doña Juana, de 66 años, que ha obtenido 23 puntos en el *MM test*.

4 En la sección final de esta revisión efectuaremos algunos comentarios y reflexiones respecto del establecimiento del estándar de referencia.

Tabla 1

Sensibilidad y especificidad del *MM test* para la detección de deterioro cognitivo.

		Deterioro cognitivo		
		Presente	Ausente	
Resultado diagnóstico del "MM test"	Positivo	**86**$^{(a)}$	4$^{(b)}$	90$^{(a+b)}$
	Negativo	7$^{(c)}$	**3**$^{(d)}$	10$^{(c+d)}$
		93$^{(a+c)}$	**7**$^{(b+d)}$	100$^{(a+b+c+d)}$

[VP/(VP +FN)]	=	[a / (a+c)]	=	Sensibilidad	=	92,5%
[VN/(VN +FP)]	=	[b / (b+d)]	=	Especificidad	=	42,9%

Nota: (a) = [VP] = Verdaderos Positivos; (b) = [FP] = Falsos Positivos ;
(c) = [FN] = Falsos Negativos; (d) = [VN] = Verdaderos Negativos.

Si operamos de manera mecánica, 23 puntos arrojan un resultado negativo. Ello permitiría descartar la presencia de deterioro cognitivo. Sin embargo, dado el bajo nivel de especificidad asociada al punto de corte (42,9%) lo aconsejable es la realización de exámenes adicionales u obtención de información adicional para descartar el diagnóstico con mayor seguridad. Por el contrario, si la señora Juana obtuviese 20 puntos en el mismo test, la interpretación cambia. Dado que el punto de corte está asociado a una alta sensibilidad (92,5%), ello aporta evidencia a favor de la confirmación de la condición. No obstante, como veremos más adelante, existen otros factores importantes a considerar antes de tomar alguna decisión al respecto. Uno de estos factores es que usualmente no existe un único puntaje de corte adecuado. Entonces, con el fin de ayudar al clínico en el proceso de toma de decisiones, es altamente recomendable que quienes publiquen o validen un test, provean de EED para diferentes puntos de corte[6]. Este asunto será abordado con más detalle en la siguiente subsección.

Curva ROC (Receiver Operating Characteristic)

Recientemente, señalamos que no siempre existe un único punto de corte adecuado. De hecho, en ocasiones, es incluso difícil de determinar cuál punto de corte puede ser el más apropiado. Entonces, la pregunta es: ¿cómo encontrar el punto de corte óptimo? Al mismo tiempo ocurre con frecuencia que tenemos dos tests distintos que, aparentemente, miden lo mismo pero reportan diferentes niveles de sensibilidad y especificidad. La pregunta, en este caso, es ¿cómo saber cuál test deberíamos escoger? A continuación propondremos el uso de una técnica útil para intentar responder ambas interrogantes.

La curva de las características operativas del receptor (del inglés, Receiver Operating Characteristic Curve), o simplemente curva ROC, es un gráfico que representa el funcionamiento operativo de un test. Dicho de un modo más simple, refleja la eficiencia diagnóstica general de un test. Una ventaja es que tal eficiencia

diagnóstica puede ser estimada, además, para diferentes puntos de corte.

Para construir una curva ROC hemos de graficar la sensibilidad del test en función de su proporción de falsos positivos (1 - Especificidad). Usualmente, la sensibilidad se ubica en el eje de las abscisas y la proporción de falsos positivos en el eje de las ordenadas. Cada punto en el gráfico corresponde a un punto de corte específico del test. Al conectar los diferentes puntos obtendremos una curva que nos permitirá identificar cuál es el puntaje de corte asociado a una mejor relación sensibilidad vs. especificidad. Aquel punto de corte que se acerque más hacia la esquina superior izquierda del gráfico representará el nivel óptimo de funcionamiento.

Un indicador importante es el !"área bajo la curva" ROC (AUC; del inglés "Area under the curve"). Usualmente, mientras mayor sea el área bajo la curva (AUC), mayor será la eficiencia diagnóstica general del test. Aun así, es posible contar con dos test que posean AUC similares, pero diferentes curvas características. En este caso, hemos de evaluar si los puntajes de corte óptimos favorecen la sensibilidad o, bien, la especificidad. Por ello, es importante no perder de vista el contexto ni los propósitos evaluativos. El valor del AUC nos permitirá comparar las características operativas de diferentes tests y, por ende, nos ayudará a escoger cuál sería el más adecuado. A continuación, la figura 1 muestra las curvas ROC para el hipotético *MM test* y el recientemente validado *NN test.*

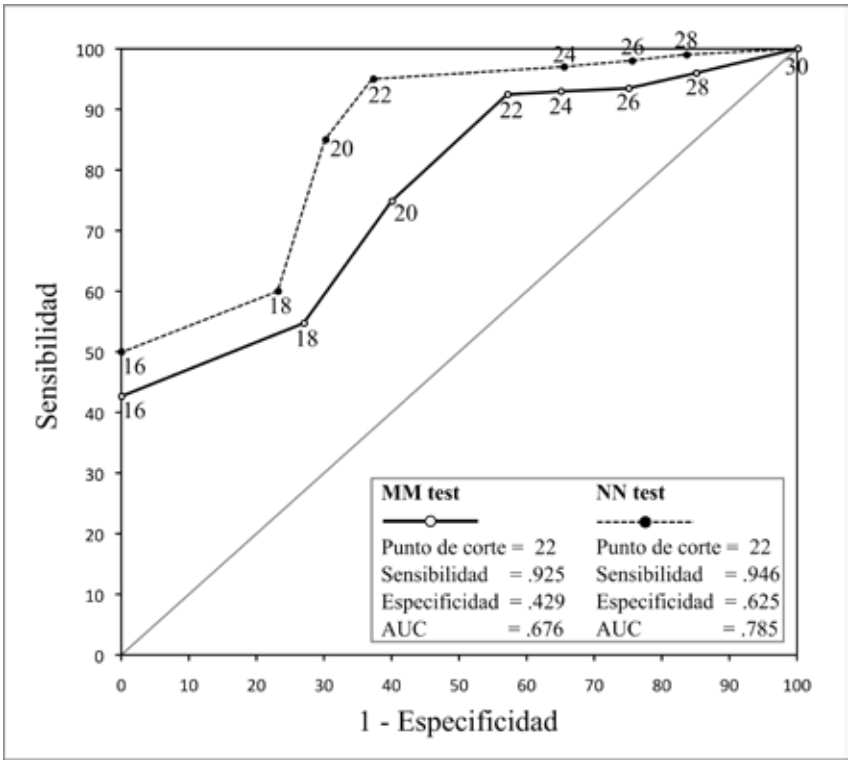

Figura 1. Curva ROC (Receiver Operating Characteristic curve) para el *MM test* y el *NN test*.

Primero trataremos de establecer los puntos de corte óptimos. Para el caso del *MM test* podemos observar que un punto de corte de 22 puntos es aquel que posee el mejor equilibrio entre sensibilidad vs. proporción de falsos positivos (i.e., 1 – Especificidad), en comparación con los otros puntos de corte para el mismo test. Si quisiéramos contar con mayor sensibilidad, podríamos aumentar el punto de corte a 24 (o incluso hasta 28 puntos). De este modo, puntajes inferiores a 24 (o 28, según sea el caso) darían con un resultado positivo. Si observamos atentamente la curva ROC, un punto de corte de 22 puntos está asociado a un sensibilidad de 92,5% y a una especificidad de 42,9% (i.e., 57,1% de falsos positivos), lo que ya es poco deseable. Entonces, en la medida que aumentemos el punto de corte, el incremento en los niveles sensibilidad será mínimo, en comparación con el importante aumento en la proporción de falsos positivos. Evidentemente, la ponderación de los costos vs. los beneficios resulta ser poco conveniente. En consecuencia, mantendremos un punto de corte de 22 puntos o menos para determinar un resultado positivo. Por otra parte, si quisiéramos aumentar la especificidad del *MM test* (i.e., reducir la proporción de falsos positivos) podríamos considerar bajar el punto de corte a 20 puntos. Ello lograría reducir la proporción de falsos positivos a un 39%. No obstante, el sacrificio en sensibilidad sería también ostensible, bajando de un 92,5% a un 75%. Ello constituye evidencia que nos sugiere, nuevamente, mantener el punto de corte previo.

Si evaluamos la misma situación para el nuevo *NN test*, el punto de corte de 22 puntos está claramente asociado a un mejor balance entre sensibilidad y especificidad. El punto de corte ideal puede ser fácilmente identificado en el gráfico. Más aún, si quisiéramos mejorar la especificidad del *NN test*, la disminución del puntaje de corte a 20 puntos mantendría aún niveles más que aceptables de sensibilidad. Por lo tanto, no sería una mala opción en caso de requerir mejor especificidad.

Si comparamos visualmente ambas curvas ROC, podríamos estimar de inmediato que el *NN test* tiene un funcionamiento operativo general mejor al del *MM test*. No obstante, una para obtener una estimación precisa del área bajo la curva, hemos de considerar el índice AUC[5]. Entonces, ante la pregunta ¿cuál test deberíamos escoger?, observaremos los valores AUC respectivos para tomar una decisión adecuada. El área bajo la curva del *MM* es AUC = .676. Por su parte, el *NN test* tiene un índice AUC = .785. En este caso, el *NN test* posee un índice AUC que indica una mejor capacidad discriminativa. Por lo tanto, entre ambos, sería recomendable utilizar el *NN test*.

Sin embargo, lo anterior sólo nos indica que el *NN test* opera mejor que el *MM test*, pero nada dice respecto de si alguno de ellos posee una buena capacidad discriminativa. Para aclarar este punto, diremos que los valores del índice AUC fluctúan entre 0 y 1. Valores cercanos a 1 indicarán, por lo tanto, una excelente capacidad discriminativa. Valores equivalentes a .5 caerán exactamente sobre la diagonal trazada en el gráfico, e indica que el test es inútil en términos discriminativos. Hosmer y Lemeshow(18) sugieren los siguientes parámetros para interpretar el AUC:

(a)	AUC = .5	= Nula capacidad discriminativa
(b)	AUC = .70 - .79	= Capacidad discriminativa aceptable
(c)	AUC = .80 - .89	= Capacidad discriminativa excelente
(d)	AUC ≥ .9	= Capacidad discriminativa sobresaliente

5 Un buen tutorial respecto de cómo calcular de manera simple el valor AUC, puede ser encontrado en Obuchowski NA. Receiver Operating Characteristic Curves and Their Use in Radiology. Radiology. 2003; 229:3-8.

Si nos basamos en estos parámetros, podríamos aseverar que las características operativas del *MM test* son menos que aceptables y que el *NN test* se ubica en el límite superior de la categoría "capacidad discriminativa aceptable". Por lo tanto, podríamos ahora optar con mayor tranquilidad por el *NN test*. No obstante, de tener acceso a otro test con mejores EED, de seguro sería el recomendado.

Hasta ahora nos hemos referido a las bondades y deficiencias de dos tests hipotéticos, pero poco hemos dicho respecto de quién obtiene un resultado positivo en ellos: el paciente. La pregunta ahora es, ¿podríamos obtener alguna información sobre el paciente interpretando el valor AUC? La respuesta es: sí, pero de manera parcial. Por ejemplo, el *NN test* tiene un valor AUC = .785. Una posible interpretación del área bajo la curva nos permite afirmar que "si seleccionamos aleatoriamente a un paciente con deterioro cognitivo y a uno sin tal diagnóstico, el primero tendrá un 78,5% de probabilidad de obtener un resultado positivo en el test en comparación con el segundo". Esta interpretación nos entrega información más específica, y nos ayuda en el proceso de toma de decisión. Pese a ello, esto no nos permite afirmar nada respecto del resultado positivo en un test de un paciente en particular. En la siguiente subsección avanzaremos un poco más en tal dirección, con el fin de obtener información más específica y relacionada con el paciente.

Razón de verosimilitud (Likelihood Ratio)

Aun cuando tengamos un test con excelentes características operativas, existe siempre la posibilidad de obtener falsos positivos (i.e., pacientes con un resultado positivo en el test que, efectivamente, no presentan la condición) y falsos negativos (i.e., pacientes con un resultado negativo en el test que, efectivamente, presentan la condición). Paralelamente, están los casos correctamente identificados (i.e., verdaderos positivos y verdaderos negativos). Todas estas posibilidades están intrínsecamente ligadas a la sensibilidad y especificidad del test.

Existe la posibilidad de resumir ambos estadísticos de eficiencia diagnóstica en un solo índice, conocido como razón de verosimilitud[6] o LR (del inglés Likelihood Ratio). La razón de verosimilitud se refiere siempre a la "probabilidad de tener la condición", y, su designación positiva (+) o negativa (–), se refiere simplemente al resultado del test. Al igual que la sensibilidad y la especificidad, el LR puede ser considerado como una propiedad fija del test.

Por una parte tenemos el LR(+), que nos indica la probabilidad de que un resultado positivo en el test provenga de una persona que efectivamente posea la condición(13). El LR (+) equivale a la razón entre los verdaderos positivos y los falsos positivos(14). Por otra parte, tenemos el LR(–) que nos dice cuál es la probabilidad de que un resultado negativo en el test provenga de una persona que no posea la condición. El LR(–) es la razón entre los falsos negativos y los verdaderos negativos(14).

Como interpretación general, un LR mayor a 1 sugiere que el resultado del test aumenta la probabilidad de que la condición efectivamente exista. Por definición, los LR(+) deben ser mayores que 1 y los LR(–) deben ser fracciones positivas entre 0 > LR(–) <1. Si tanto LR(+) como LR(–) son iguales a 1, ello es indicador de nula capacidad discriminativa y por ende no aportan información diagnóstica. Haciendo un paralelo con la curva ROC, un LR = 1 caería en cualquier punto sobre la línea diagonal del gráfico ROC. En resumen, mientras más alejados del valor 1 se encuentren LR(+) y LR(–), mayor será el poder resolutivo del test(14). Esto quiere decir literalmente que el clínico no necesitaría realizar ma-

6 También conocido como cociente de probabilidades.

yores inferencias posteriores. A esto es a lo que se refiere Meehl con la predicción basada en métodos cuantitativos, objetivos y formales. Siguiendo con el ejemplo del *MM test*, el LR(+) asociado es obtenido como sigue:

$$LR(+) = \text{Sensibilidad} / (1 - \text{Especificidad}) \quad (1)$$

$$LR(+) = .925 / (1 - .429) = 1.62 \quad (2)$$

La interpretación correcta de este resultado indica que alguien con deterioro cognitivo tendrá 1.62 veces más probabilidades de obtener un resultado positivo en el *MM test* que alguien que no manifieste deterioro cognitivo. Evidentemente, mientras mayor sea el valor del LR+, más certeza tendremos al momento de confirmar la presencia de un diagnóstico. Ahora, ¿qué valor LR posee el *NN test* para resultados positivos?

$$LR(+) = \text{Sensibilidad} / (1 - \text{Especificidad}) \quad (3)$$

$$LR(+) = .946 / (1 - .625) = 2.52 \quad (4)$$

De acuerdo con la nueva estimación, un resultado positivo tendrá una interpretación diferente. En este caso podemos afirmar que una persona con deterioro cognitivo tendrá 2.5 veces más probabilidades de obtener un resultado positivo en el *NN test* que alguien que no la manifieste.

Por su parte, el LR(–) es obtenido a partir de la siguiente fórmula:

$$LR(-) = (1 - \text{Sensibilidad}) / \text{Especificidad} \quad (5)$$

Mientras menor sea su valor (i.e., más cercano a cero) y, dado un resultado negativo en un test, podremos excluir una condición con mayor certeza. Para el caso del *MM test,* el LR(–) = 0.175 y para el *NN test* el LR(–) = 0.086. Por lo tanto, ante un resultado negativo, el *NN test* nos daría mayor certeza que el *MM test* al descartar la presencia de deterioro cognitivo. Este resultado era esperable, en tanto el *NN test* mostró una mejor especificidad que el *MM test*.

Hasta aquí, hemos incorporado información que nos permite obtener estimaciones probabilísticas que dan cuenta del nivel de certeza diagnóstica. No obstante, podemos refinar aún más nuestra certeza diagnóstica si hacemos uso de información relativa a la prevalencia de la condición en la población. Esta información juega un rol importante en la precisión de nuestras predicciones y será abordado en la siguiente subsección.

Valor predictivo (Predictive Value)

Hasta ahora, hemos revisado EED cuyas propiedades no se ven afectadas por el efecto de la prevalencia de una condición. No obstante, revisaremos a continuación un índice conocido como valor predictivo, que incorpora dicha información y nos permite realizar estimaciones contextualizadas, precisas y que pueden ser extrapoladas a un paciente en particular.

La prevalencia es entendida básicamente como la proporción de personas que padecen una condición de interés en una población determinada. Dependiendo del tipo subpoblación, podemos obtener las tasas de prevalencia para diferentes sexos, grupos etarios, zonas geográficas, u otras variables de interés. En términos generales, podemos decir que cuando la prevalencia de una condición es muy baja (e.g., 1%) se hace menos probable que un individuo cualquiera la presente. Como contraparte, si la prevalencia de

una condición es extremadamente alta en un grupo determinado (e.g., 97%) se hace altamente probable que un individuo la presente.

Al igual que para la razón de verosimilitud (LR), existe un valor predictivo positivo (VPP) y valor predictivo negativo (VPN), dependiendo del resultado obtenido en un test. El VPP corresponde a "la probabilidad condicional de poseer un diagnóstico (i.e., deterioro cognitivo) dado un resultado positivo en un test"[13]. A la inversa, el VPN corresponde a "la probabilidad condicional de no poseer un diagnóstico dado un resultado negativo en un test"(13). El VPP y VPN pueden ser obtenidos a partir de las siguientes fórmulas:

$$VPP = \frac{\text{Prevalencia} \times \text{Sensibilidad}}{(\text{Prevalencia} \times \text{Sensibilidad}) + [(1 - \text{Prevalencia}) \times (1 - \text{Especificidad})]} \quad (6)$$

$$VPN = \frac{(1 - \text{Prevalencia}) \times \text{Especificidad}}{[(1 - \text{Prevalencia}) \times \text{Especificidad})] + [\text{Prevalencia}) \times (1 - \text{Sensibilidad})]} \quad (7)$$

A continuación, mediante un ejemplo obtendremos e interpretaremos el VPP. Luego, a modo de ejercicio, el lector podrá calcular e interpretar el VPN sobre la base de los mismos datos.

Como es ampliamente conocido, la prevalencia del deterioro cognitivo aumenta según aumenta el rango etario. De acuerdo con la Encuesta Nacional de Salud[19] (ENS Chile 2009-2010) la prevalencia de deterioro cognitivo entre el rango de 60 a 69 años asciende a un 7,2% de la población (I.C. 4,1% – 12,1%). Para el rango comprendido entre 70 a 79 años, la prevalencia aumenta a 12,8% (I.C. 7% – 22,5%). Para adultos mayores a 80 años, la prevalencia alcanza el 20,9% (I.C. 4,1% – 12,1%). Consideremos entonces que doña Juana (66 años) y doña María (78 años) han obtenido un resultado positivo en el *NN test*. Como indicamos anteriormente, su sensibilidad es 94,6% y su especificidad 64,1%. Luego, calcularemos el VPP para ambas. Doña Juana pertenece al rango etario de 60 a 69 años, por ende la prevalencia de deterioro cognitivo asociada asciende a 7,2%. Con esta información, más los valores de sensibilidad y especificidad podemos calcular el VPP como sigue:

$$VPP = \frac{.072 \times .946}{(.072 \times .946) + [(1 - .072) \times (1 - .625)]} = .164 \quad (8)$$

Este resultado indica que doña Juana tiene un 16,4% de probabilidad de padecer deterioro cognitivo, dado su resultado positivo en el *NN test*. Recordemos que hemos estimado una probabilidad condicional. Claramente, la evidencia a favor de tal diagnóstico es relativamente débil. Por lo tanto, correspondería efectuar evaluaciones adicionales y continuar estudiando el caso. Por otra parte, doña María pertenece al rango etáreo de 70 a 79 años, cuya prevalencia asciende al 12,8%. Entonces, el VPP es equivalente a:

$$VPP = \frac{.128 \times .946}{(.128 \times .946) + [(1 - .128) \times (1 - .625)]} = .270 \quad (9)$$

Tal resultado indica que, para doña María, la probabilidad de padecer deterioro cognitivo dado su resultado positivo en el *NN test* es de un 27%, es decir, poco menos de un tercio. Aun cuando, en este caso, el VPP es mayor, ello no constituye evidencia concluyente.

Estos resultados reflejan que en la medida que la prevalencia disminuye, el poder predictivo de cualquier test disminuirá. Entonces podemos sostener que, mientras menor sea la prevalencia de una condición, un test será mejor utilizado para descartar la presencia de una condición más que para confirmarla(13). De modo inverso, en la medida que la prevalencia sube, aumentará el poder predictivo de cualquier test. Entonces, cuando la prevalencia es alta, el test será mejor utilizado para confirmar la presencia de la condición y no para descartarla[13]. Evidentemente, el mejor escenario predictivo se constituye cuando la prevalencia de la condición es exactamente 50%. En este caso, la prevalencia no ejercerá ninguna influencia sobre las características operativas del test y, por tanto, sería como realizar estimaciones basadas en los EED anteriormente revisados. No obstante, cada vez que el clínico no considera el valor de la tasa de prevalencia al realizar sus predicciones, está asumiendo "inadvertida e implícitamente" que dicha tasa equivale al 50%. Ello implica que, cada vez que la prevalencia sea inferior a dicho valor, estaremos sobreestimando la probabilidad de ocurrencia de un diagnóstico, dado un resultado positivo en cualquier test. De este modo, el riesgo de sobrediagnóstico aumenta significativamente, sobre todo en aquellos casos en que la prevalencia es particularmente baja. Por ende, aconsejamos al lector poner especial atención sobre este punto.

A continuación, estableceremos algunas consideraciones y reflexiones generales respecto de la adopción de una aproximación cuantitativa, como complemento a los procedimientos diagnósticos.

Discusión

De acuerdo con la Asociación Americana de Psicología (APA), el establecimiento de lineamientos resulta esencial para el desarrollo profesional y la protección del público, más allá de la existencia de diversidad teórica y práctica dentro de las distintas áreas de la psicología(20). Entre ellas encontramos, por cierto, a la neuropsicología. Asimismo, se indica que en la era de la práctica psicológica basada en la evidencia (Evidence-Based Practice in Psychology, EBPP), tales lineamientos deben estar basados en un cuidadoso y sistemático balance entre la experiencia clínica y la evidencia empírica[3]. Dentro de ese contexto, es que esta revisión ha intentado realzar y enfatizar la importancia de contar con herramientas que permitan mejorar la eficiencia y precisión de los diagnósticos clínicos en neuropsicología, junto con promover su uso. No obstante, no constituye un objetivo encandilar al lector con las maravillas y bondades de una aproximación predictiva cuantitativa. Como toda aproximación, tiene sus ventajas y desventajas. Por ello, expondremos a continuación algunas consideraciones respecto de la adopción de una aproximación cuantitativa en la predicción diagnóstica clínica.

Primero, aun cuando la adopción de una aproximación formal contribuye a dotar de mayor verosimilitud a nuestras hipótesis diagnósticas, en ningún caso permite establecer certezas absolutas. A lo que sí podemos aspirar es a obtener niveles de certeza que sean lo suficientemente confiables como para constituir evidencia a favor de una o más hipótesis diagnósticas. Segundo, hemos de establecer que la eficiencia diagnóstica de cualquier test se verá afectada por la dificultad para establecer estándares de referencia claros y adecuados. Entenderemos por estándar de referencia como, el mejor método (o combinación de métodos) disponible para determinar la presencia o ausencia de una condición de interés(21). El establecimiento

de un estándar de referencia constituye en ocasiones un reto. Incluso, para el caso de algunas condiciones, tales estándares son simplemente inexistentes(22). No obstante, para nuestra tranquilidad, la carencia de estándares de referencia es una situación más bien común en medicina y epidemiología, y no está circunscrita solamente a la neuropsicología(6). De acuerdo a Joseph, Gyorkos y Coupal(22), podría argumentarse que esta situación es virtualmente invariable en tanto que son muy pocos los tests que pueden ser considerados 100% precisos. Este hecho, escasamente reconocido y reportado, ha despertado el interés de algunos investigadores, quienes han propuesto algunas directrices para la adecuada determinación de los estándares de referencia. Por ejemplo, la iniciativa STAR D(6, 21, 23) (del inglés Standards for Reporting of Diagnostic Accuracy) tiene por objetivo determinar criterios definidos y claramente establecidos para reportar la eficiencia diagnóstica de un test. Este tipo de iniciativas permite mejorar las condiciones en que un instrumento es desarrollado y validado, y por ende, permite mejorar su eficacia diagnóstica. Entre tales criterios, se encuentra precisamente el establecimiento claro de un estándar de referencia y los argumentos que sustentan su elección. Tercero, la variabilidad a la que están afectas las tasas de prevalencia puede alterar la precisión de las estimaciones de eficiencia diagnóstica[7]. Esta variabilidad tiene diversas fuentes, tales como cambios demográficos, la existencia de población no consultante o la disminución (o aumento) en las tasas de incidencia o, incluso, el sub y sobrediagnóstico. Este último factor no debe ser minimizado o desestimado, dado que es más frecuente de lo que se piensa. Particularmente, el sobrediagnóstico se da con mayor frecuencia en servicios de salud primaria y en aquellos trastornos que tienden a ser más prevalentes (e.g., depresión)(24-26). Una manera de controlar la variabilidad de las tasas de prevalencia es la utilización de EED. Sin embargo, al no ser ésta la única fuente de variabilidad, nuestra sugerencia apunta a estar debidamente informados y permanentemente actualizados respecto de eventuales variaciones en las tasas prevalencia. Ello es particularmente importante al estimar algunos EED que se ven particularmente afectados por la prevalencia de una condición, por ejemplo, el valor predictivo.

Finalmente, y a modo de reflexión, diremos que quien elabora un diagnóstico clínico tiene la responsabilidad ineludible de estar al tanto de aquellos métodos, tanto diagnósticos como terapéuticos, que puedan traducirse en algún beneficio potencial o real para el paciente. Asimismo, quien privilegie el uso de algunas técnicas diagnósticas o terapéuticas en conocimiento de la disponibilidad de otras más eficaces, contraviene lo establecido por los principios éticos del psicólogo y su respectivo código de conducta (APA, Principio A: Benevolencia y no maleficencia)(27). Desde esta óptica, la elección de los métodos de evaluación diagnóstica a utilizar deja definitivamente de ser un tema de elección o predilección personal, y debería estar dentro de las recomendaciones de los estudios de evidencia clínica. Sin embargo, es bien sabido que la utilización de aproximaciones clínicas basadas en la evidencia se han concentrado mayormente en temáticas relativas al tratamiento, mientras que su aplicación al ámbito diagnóstico ha tenido un desarrollo bastante menor(28). Considerando que dentro de los objetivos de las pruebas diagnósticas están la detección o exclusión de trastornos, la contribución al manejo terapéutico, y la evaluación del pronóstico, entre otros(28), es innegable que el diagnóstico constituye el eje de todo proceso clínico, razón por la que no debe estar exento del uso de las mejores técnicas o métodos disponibles para su adecuado establecimiento.

Las consecuencias de un diagnóstico errado podría generar estrategias terapéu-

7 Acá hemos descrito el valor predictivo, no obstante existen otros EED que también se ven afectados por el valor de las tasas de prevalencia de la condición.

ticas inapropiadas, poniendo en riesgo el pronóstico y la calidad de vida de un paciente. Dentro de los procesos de razonamiento y toma de decisiones clínicos descritos, están el tradicional (no analítico) y el analítico. El primero se basa en la experiencia del clínico, mediante el reconocimiento inconsciente o automatizado de algunos patrones, que aunque cumple un rol importante, podría llevar a errores y sesgos diagnósticos. Como contraparte, aunque no excluyente, está la estrategia diagnóstica analítica que se basa fundamentalmente en el uso de probabilidades condicionales[8]. La complementación de ambos tipos de estrategias diagnósticas es lo más aconsejado para la obtención de un diagnóstico acertado(29). Por consiguiente, sin desmerecer el valor atribuible a la experiencia y a la intuición, sería conveniente complementar nuestras hipótesis diagnósticas con información obtenida por medio de procedimientos objetivos y formales. Una aproximación diagnóstica meramente intuitiva podría generar consecuencias pragmáticas tan complejas y lamentables para los pacientes como las que Rosenhan(30) demostró en su clásico experimento de mediados de los setenta. La actividad clínica en el área de la salud mental no suele estar exenta de eventuales procesos de estigmatización. Por ende, la determinación y posterior etiquetamiento de algunos diagnósticos (acertados o no) podrían derivar en conductas de estigma y discriminación, con el consecutivo efecto negativo en la calidad de vida general, tanto de un paciente como de su familia(31,32,33). La situación podría, eventualmente, tornarse aún más grave cuando el costo personal y social de una estigmatización se produce por un diagnóstico equivocado. Todo diagnóstico errado conlleva el riesgo de efectuar intervenciones iatrogénicas que, ciertamente, generan un impacto negativo en la recuperación integral de los pacientes, favoreciendo el deterioro social, y dificultando la adquisición de mayor autonomía y reincorporación a aquellas actividades o roles sociales apropiados.

Por todo lo anterior, nuestra invitación es a no desestimar el uso de métodos que podrían contribuir a la generación de cambios positivos y duraderos en la vida de muchos de nuestros pacientes.

Referencias

1. Rufo-Campos M. La neuropsicología: historia, conceptos básicos y aplicaciones. Rev. Neurol. 2006;43:S57-S58.
2. Akhutina TV. LS Vigotsky, AR Luria. La formación de la neuropsicología. Rev. Esp. Neuropsicol. 2002;4:108-129.
3. Heilbronner RL, Taylor HG, Wills K., Boone K., Bigler E, Fortuny LAI et al. American Academy of Clinical Neuropsychology (AACN) practice guidelines for neuropsychological assessment and consultation. Clin Neuropsychol 2007;21:209-231.
4. Hannay HJ, Bieleauskas LA, Crosson BA, Hammeleke TA, Hamsher K., Kofler SP. Proceedings of the Houston conference on specialty education and training in clinical neuropsychology, september 3-7, 1997, University of Houston Hilton and Conference Center. Arch Clin Neuropsychol 1998;13:157-158.
5. Meehl PE. Philosophy of science: Help or hindrance? Psychological Reports. 1993;72:707-33.
6. Millis SR. What clinicians really need to know about symptom exaggeration, insufficient effort, and malingering: Statistical and measurement matters. In: Morgan J., Sweet J. (Eds.). Neuropsychology of malingering casebook. Hove, East Sussex, Reino Unido: Psychology Press, 2008.
7. Meehl PE. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Northvale, NJ, USA: Jason Aronson, 1996.
8. Grove WM. Clinical versus statistical prediction: the contribution of Paul E. Meehl. J Clin Psychology 2005;61:1233-43.

8 El lector interesado puede revisar el teorema de Bayes, propuesto por el matemático Rvdo. Thomas Bayes (1701-1761).

9. Grove WM, Meehl PE. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The Clinical-Statistical Controversy. Psychol Public Pol Law 1996;2:293-323.
10. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. Psychol Asses 2000;12:19-30.
11. White RF, James KE, Vasterling JJ, Marans K., Delaney R., Krengel M. et al. Interrater reliability of neuropsychological diagnoses: a Department of Veterans Affairs cooperative study. J Int Neuropsychol Soc 2002;8:555-65.
12. O'Hagan A. Bayesian statistics: principles and benefits. In: Van Boekel M., Stein A., Van Bruggen A. (Eds.). Bayesian statistics and quality modelling in the agro food production chain. Wageningen, The Netherlands: Kluwer Academic Publishers, 2003.
13. Streiner D. Diagnosing tests: using and misusing diagnostic and screening tests. J Pers Assess 2003;81:209-219.
14. Spitalnic S. Test Properties 2: Likelihood Ratios, Bayes' Formula, and Receiver Operating Characteristic Curves. Hospital Physician. 2004;40:53-8.
15. Streiner D., Norman G. PDQ epidemiology, 2nd. ed. Toronto, Ontario, Canada: Decker, 1996.
16. Iverson GL. Identifying exaggeration and malingering. Pain Pract 2007;7:94-102.
17. Obuchowski NA. Receiver Operating Characteristic Curves and Their Use in Radiology. Radiology. 2003;229:3-8.
18. Hosmer DW, Lemeshow S. Applied logistic regression: Hoboken, NJ:Wiley-Interscience, 2000.
19. MINSAL. Deterioro cognitivo del adulto mayor. In: Encuesta Nacional de Salud ENS Chile 2009-2010, Vol. II. ed. Santiago de Chile: Gobierno de Chile, Ministerio de Salud, Departamento de Epidemiología; Pontificia Universidad Católica de Chile; Universidad Alberto Hurtado, Observatorio Social; 2010. 442-461. Revisadoen http://www.encuestasalud.cl/ens/wp-content/uploads/2011/2009/InformeENS_2009-2010_CAP2015.pdf el 8 de mayo de 2012.
20. Sturm CA, Hancock KA, Cerbone AR, de La Cancela V., Connell MA et al. Determination and documentation of the need for practice guidelines. Am psychol 2005;60:976-8.
21. 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.
22. Joseph L., Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 1995;141:263-72.
23. Bossuyt PM, Reitsma JB, E Bruns D., Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Fam Pract 2004 ;21(1):4-10.
24. Boland RJ, Diaz S., Lamdan RM, Ramchandani D., McCartney JR. Overdiagnosis of depression in the general hospital. Gen Hosp Psychiatry. 1996;18:28-35.
25. Aragonés E., Piñol JL, Labad A. The overdiagnosis of depression in non-depressed patients in primary care. Fam Pract 2006;23:363-8.
26. Parker G. Is depression overdiagnosed? Yes. BMJ 2007;335:328.
27. American Psychological Association. Ethical principles of psychologists and code of conduct. Washington D.C.; American Psychological Association, 2002.
28. Knottnerus JA, Van Weel C., Muris JWM. Evaluation of diagnostic procedures. BMJ 2002;324:477-80.
29. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ 2005;39:98-106.
30. Rosenhan DL. On being sane in insane places. Science. 1973;179:250-8.
31. Garand L., Lingler JH, Conner KO, Dew MA. Diagnostic labels, stigma, and participation in research related to dementia and mild cognitive impairment. Res Gerontol Nurs 2009;2:112-21.
32. Chapman DP, Williams SM, Strine TW, Anda RF, Moore MJ. Dementia and Its Implications for Public Health. Preventing Chronic Disease 2006;3:1-13.

33. Corrigan PW, Watson AC. Understanding the impact of stigma on people with mental illness. World Psychiatry 2002;1:16-20.

Correspondencia a:
Alonso Ortega G.
alonso.ortega@uv.cl.
Walter Lips C.
walter.lips@uv.cl.