VALIDACIÓN DE PRUEBAS DIAGNÓSTICAS[1]

TARABLA, H. D.[2]

INTRODUCCIÓN

La confiabilidad (ausencia de error aleatorio) y la validez (ausencia de error sistemático -sesgo-) son elementos independientes de toda prueba. Un proceso diagnóstico puede repetir consistentemente resultados falsos positivos (ser confiable pero no válido) o puede dar resultados correctos con relativamente alta variabilidad entre submuestras (ser válido pero poco confiable). Generalmente se refiere a la falta de validez como una falla de la prueba en detectar individuos verdaderamente enfermos (poco sensible) y/o verdaderamente sanos (poco específica), aunque una baja confiabilidad también pueda afectar su sensibilidad y su especificidad. La confiabilidad y la precisión están asociadas con la ausencia de error aleatorio, mientras que validez se relaciona con la ausencia de sesgo. La exactitud, por otra parte, se relaciona con la ausencia de ambos tipos de errores.

Evaluar la exactitud de las pruebas diagnósticas implica confirmar tanto su confiabilidad como su validez, mientras que validarla se aplica más a la medición de su capacidad discriminatoria. La confiabilidad es un importante pre-requisito para que un proceso analítico sea exacto y para la validación de ese proceso. El diseño de estudios de validación de una prueba considera todos los temas vinculados con la cuanti-ficación del error sistemático o la falta de él. Esto es en definitiva cuantificar la habilidad de la prueba en detectar individuos verdadero positivos en la población enferma (sensibilidad) y verdadero negativos en la población sana (especificidad).

DISEÑO DE ESTUDIOS

Antes de diseñar un estudio para validar una prueba se debe tener muy en claro para qué va a ser utilizada (tamizado o confirmación, diagnóstico individual o grupal, etc.). El estudio puede ser experimental, observacional o mixto. Los ensayos experimentales son muy útiles para obtener resultados preliminares, pero generalmente tienden a sobrestimar el desempeño de la prueba. En todos los casos se deben seleccionar apropiadamente estrategias para reducir sesgos (ej. evaluaciones ciegas).

Los estudios para estimar la sensibilidad y la especificidad de las pruebas deben ser válidos tanto interna como externamente. En el primer caso, deben dar un estimador no sesgado del desempeño de la prueba bajo las condiciones en las que se realizó el trabajo (protocolo, laboratorio y muestras de

1.- Resumen de una disertación, XIII Reunión Científico-técnica de la Asociación Argentina de Veterinarios de Laboratorios de Diagnóstico, Merlo, San Luis, 16-17/11/00.

2.- Cátedra de Epidemiología, Facultad de Ciencias Veterinarias, Universidad Nacional del Litoral. Kreder 2805, (3080) Esperanza, provincia de Santa Fe - Estación Experimental Agropecuaria Rafaela, Instituto Nacional de Tecnología Agropecuaria, C.C. 22, (2300) Rafaela, provincia de Santa Fe.

Manuscrito recibido el 30 de marzo de 2001 y aceptado para su publicación el 15 de marzo de 2002.

referencia). En el segundo, la muestra bajo estudio debe ser representativa de la población de referencia. Los tres enfoques principales para estimar el poder discriminatorio de una prueba son: a) condicional al estado de salud, donde enfermos y sanos son conocidos a priori, b) transversal, donde se evalúan al mismo tiempo tanto las condiciones de sanos o de enfermos como las de positivos o de negativos a la prueba, y c) condicional a los resultados de otra prueba. La validación integral de una prueba debería incluir: a) una estimación preliminar de la sensibilidad y la especificidad, b) la determinación del límite mínimo de detección, c) la evaluación de posibles causas de diagnósticos falsos, y d) una evaluación a campo.

TAMAÑO DE LA MUESTRA

Para determinar el n de animales sanos y enfermos a incluir en el estudio se usa la fórmula para estimar el intervalo de confianza de una proporción. Sin embargo, ésta requiere de un conocimiento previo de los valores aproximados de sensibilidad y especificidad. En ausencia de estimaciones anteriores, en un primer ensayo se recomienda utilizar un mínimo de 100 individuos por grupo. Mayor sea la difusión de la prueba y/o su importancia económica o en salud pública, mayor debe ser el n para estrechar los límites de los intervalos de confianza de las estimaciones.

SELECCIÓN DEL CRITERIO DE POSITIVIDAD

Los resultados de una prueba pueden ser expresados en forma cualitativa («positivos» o «negativos»), o por medio de algún tipo de gradiente numérico a lo largo de una escala ordinal o continua. Estos últimos pueden ser expresados en forma dicotómica por medio de la selección de un valor crítico o de corte en la escala por encima del cual los resultados obtenidos se consideran positivos y por debajo del mismo se clasifican como negativos. La selección del criterio de positividad («anormal») y negatividad («normal») para una prueba determinada se puede efectuar de utilizando seis métodos diferentes: a) distribución Gaussiana, b) percentiles, c) aceptación cultural, d) terapéutico, e) factores de riesgo y f) diagnóstico. Una buena prueba es aquella en la cual las distribuciones de frecuencias de los positivos y negativos a la prueba tienen el menor área posible de superposición.

Generalmente, para determinar el verdadero estado de salud de los individuos se utiliza la denominada prueba patrón, de referencia, regla o estándar de oro. Ésta es un método diagnóstico o una combinación de ellos que determina en forma absoluta y sin error si la condición (enfermedad, infección, etc.) está presente en un individuo. Sin embargo, en algunas enfermedades no existe tal prueba, mientras que en otras, este diagnóstico perfecto no es práctico, o bien es muy costoso, laborioso o invasor. En la mayoría de las enfermedades infecciosas es la identificación de los patógenos mediante su cultivo y aislamiento, pero en algunas ocasiones este es un proceso lento (Ej. tuberculosis), mientras que en otras no es posible aislar fácilmente al organismo responsable de la enfermedad (Ej. brucelosis, enfermedades virales). En enfermedades no infecciosas, el estándar de oro puede ser una biopsia, una necropsia o cualquier criterio que pueda diagnosticar inequívocamente la enfermedad.

Las estimaciones de sensibilidad y especificidad se complican aún más en aquellas enfermedades multicausales que producen respuestas medibles de distinta magnitud de

acuerdo al patógeno actuante. Por ejemplo, en el caso de la mastitis bovina, la capacidad de reacción de la glándula mamaria para producir o alterar diversos elementos que pueden ser medidos por las pruebas diagnós-ticas (Ej. California Mastitis Test) no es la misma ante cada especie bacteriana. Esto deriva en diferentes sensibilidades y especificidades para cada patógeno, que debe ser sumada a la variabilidad proveniente del animal (Ej. número y tiempo de lactancia), de la evolución de la enfermedad (Ej. si es una infección reciente o crónica), de la muestra de leche (Ej. primeros chorros, durante el ordeño o leche residual) y de la prueba patrón (Ej. si se pre-incuban o no las muestras antes del cultivo bacteriano). Por otra parte, las estimaciones de la especificidad de las pruebas diagnósticas para detectar infecciones intramamarias también están influenciadas por el hecho que los animales negativos provienen de rodeos que no son libres de infecciones intramamarias, por las variaciones fisiológicas en la composición de la leche relacionadas con el período de lactancia y por otros factores tales como el estrés ambiental.

ESTIMACIÓN DE LA SENSIBILIDAD Y ESPECIFICIDAD

Las estimaciones de la sensibilidad y de la especificidad se efectúan en un grupo de animales «enfermos» y «sanos» respectivamente, determinados de acuerdo a alguno de los criterios antes mencionados. Para el primer caso debe estar presente todo el espectro clínico y patológico de la enfermedad en la misma proporción que en la población de referencia. Para el segundo, los animales sanos deben ser representativos de la población sana donde se aplicará la prueba (Ej. si se trata de una prueba indirecta para medir la cantidad de células somáticas en leche no se incluirán vacas recién paridas secretando calostro) y no deben ser utilizados animales aparentemente sanos provenientes de rodeos infectados, dado que esto último llevará generalmente a subestimaciones de la especificidad de la prueba. El estado de salud de los animales a utilizar debe ser determinado con un alto grado de certeza por medio de pruebas biológicamente independientes de la que está siendo evaluada. A diferencia de los valores predictivos y de la eficiencia de la prueba, se asume que la sensibilidad y la especificidad no varían al cambiar la prevalencia real de la enfermedad. Sin embargo, y aunque no existe una dependencia formal entre sensibilidad y especificidad de la prueba y la prevalencia de la enfermedad, cualquier cambio que afecte la forma de las curvas de individuos enfermos y/o sanos o el rango de superposición entre ambas cambiará los patrones de sensibilidad y de especificidad de la prueba. Si la relación con la prueba diagnóstica es importante, estos cambios incluyen la distribución de razas, edad, sexo y exposición a afecciones antigénica-mente relacionadas. Por ello se debe recolectar información sobre las características demográficas de las poblaciones de animales que están sujetos a evaluación para revisar estos supuestos cada vez que sea posible e intentar controlarlos en el proceso de análisis estadístico.

COMPARACIONES ENTRE PRUEBAS DIAGNÓSTICAS

En la práctica es muy dificultoso cumplir con todas las premisas para determinar sensibilidad y especificidad. Por ello, se acostumbra a medir los resultados de la nueva prueba comparándolo con los de otra prueba que no es biológicamente independiente de la que está siendo evaluada. Sin embargo, se debe recordar que estas estimaciones, aunque

útiles, no reemplazan a la validación de la prueba y pueden estar sesgadas dado que es improbable que las muestras de población sana y enferma incluyan el espectro completo de las poblaciones de animales sanos y enfermos respectivamente. Por otra parte, si las pruebas que se quieren comparar están biológica-mente relacionadas, las mismas tenderán a dar los mismos resultados al ser aplicadas en un mismo individuo. Por todo ello, en estas circunstancias es conveniente llamar a estas estimaciones como «sensibilidad y especificidad relativas» a la prueba en la cual se basó el diagnóstico. Estas comparaciones son más valiosas si la sensibilidad y la especificidad de esta última se aproximan al 100 %.

Se debe tener en cuenta que la comparación de dos o más pruebas sólo indica el grado de concordancia entre los resultados de las pruebas, pero de ninguna manera si una es más sensible o específica que la otra. Además, estas comparaciones deben hacerse con un diseño ciego, de manera que la persona que efectúa una prueba desconozca el resultado de la otra para evitar serios sesgos.

Cuando se comparan dos pruebas diagnósticas aplicadas en poblaciones diferentes o en distintos subgrupos de la misma población, los índices de concordancia y la probabilidad de error pueden variar de acuerdo a la prevalencia real de la enfermedad en las poblaciones o subpoblaciones sujetas a comparación.

La manera más simple y frecuente para medir el grado de acuerdo entre los resultados de dos pruebas es la proporción de concordancia general o Concordancia Observada. Esta es igual a la suma de la probabilidad de los resultados positivos y negativos a ambas pruebas. Sin embargo, se necesita más que una medida del grado de acuerdo entre dos pruebas, dado que, salvo en circunstancias extremas, es dable esperar algún grado de concordancia debido solamente a la intervención del azar.

Una medida que corrige esto es la razón denominada Kappa (K), que incluye el cálculo de la Concordancia Esperada, que incorpora los valores marginales de cada prueba. Un valor de K = 1 representa una concordancia perfecta, más allá de la intervención de azar, K = 0 no hay concordancia.

En la comparación entre pruebas diagnósticas, se necesita como mínimo un Kappa de 0,50 para indicar un nivel de acuerdo moderado. El grado de asociación estadística es usualmente medido con la prueba de Mc Nemar.

DETERMINACIÓN DEL PUNTO DE CORTE

La exactitud de una prueba diagnóstica depende del área de superposición entre las distribuciones de los resultados positivos y negativos. Para su cálculo en las pruebas que expresan sus resultados bajo una escala continua puede utilizarse una técnica de análisis de una curva conocida generalmente por su sigla en inglés ROC («Receiver Operating Characteristic»). Esta curva se construye graficando la sensibilidad contra (1 - especificidad), bajo varias posibilidades de criterios de positividad y describe la habilidad de una prueba para discriminar entre animales enfermos y sanos.

Una prueba perfecta tiene un área del 100 %, mientras que una prueba que no mejora lo que se obtendría por azar tiene una área del 50 % limitada por una línea recta diagonal desde el ángulo inferior izquierdo al superior derecho. La curva ROC es una herramienta simple para la aplicación del método diagnóstico o del valor predic-tivo

en la selección del criterio de «anormalidad» y puede ser calculada por medio de programas de computación. Si la importancia de los diagnósticos falsos positivos es igual a la de los falsos negativos, el punto de corte óptimo está dado por el valor del punto que más se acerca al ángulo superior izquierdo. Sin embargo, en los casos que el costo de efectuar un diagnóstico falso positivo es diferente al de uno falso negativo, este punto de corte puede variar aumentando la especificidad o la sensibilidad respectivamente. Seleccionando un punto de corte bajo se obtendrá una buena sensibilidad, incremen-tando el valor predictivo negativo y disminuyendo los diagnósticos falso negativos. Si, por lo contrario, se elige un punto de corte alto se mejorará la especificidad de la prueba, aumentando el valor predictivo positivo y disminuyendo la proporción de diagnósticos falsos positivos. La curva ROC también puede ser utilizada para comparar dos pruebas diagnósticas. La mejor prueba será aquella cuya área bajo la curva sea mayor. Para obtener un resumen cuantitativo de múltiples estudios de validación se utiliza una herramienta conocida como Meta-análisis, mientras que actualmente se encuentran en desarrollo métodos estadísticos alternativos para evaluar el poder discriminatorio de las pruebas en ausencia de prueba patrón en Medicina Veterinaria.

BIBLIOGRAFÍA

BUCK, A. A. & G. G. GART. 1966. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. Am. J. Epidemiol. 83: 586-592.

CARRINGTON REID, M.; M. S. LACHS & A. R. FEINSTEIN. 1995. Use of methodological standards in diagnostic test research. J. Am. Med. Ass. **274**: 645-651.

COHEN, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 : 37-46.

DOHOO, I. R. 1981. Effects of misclassification on statistical inferences in Epidemiology. Letter to the Editor. Am. J. Epidemiol. 118: 485-486.

FARAONE, S. V. & M. T. TSUANG. 1994. Measuring diagnostic accuracy in the absence of a gold standard. Am. J. Psychiatry 151: 650-657.

FLEISS, J. L. 1981. Statistical Methods for Rates and Proportions, 2nd Ed., John Wiley & Sons Inc., New York, 321 pp.

GARDNER, I. A. & M. GREINER. 1999. Advance Methods for Test Validation and Interpretation. Freie Univ. Berlin, 74 pp.

GART, J. J. & A. A. BUCK. 1966. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am. J. Epidemiol. 83 : 593-602.

GREINER, M. 1995. Two-graph receiver operating characteristic (TG-ROC) - a Microsoft Excel template for the selection of cutoff vaalues in diagnostic tests. J. Inmunol. Methods 185: 145-146.

GREINER, M.; D. SOHR & P. GOBEL. 1995. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. J. Inmunol. Methods 185: 123-132.

HANLEY, J. A. & B. J. MCNEIL. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 143 : 29-36.

HUI, S. L. & S. D. WALTER. 1980. Estimating the error rates of diagnostic tests. Biometrics 36: 167-171.

IRWIG, L.; P. P. GLASZIOU ; G. BERRY; C. CHOCK; P. MOCK & J. M. SIMPSON. 1994. Efficient study designs to assess the accuracy of screening tests. Am. J. Epide-

miol. 140: 759-769.

IRWIG, L.; A. N. A. TOSTESON; C. GATSONIS; J. LAU; G. COLDITZ; T. C. CHAIMERS & F. MOSTELLER. 1994. Guidelines for meta-analysis evaluating diagnostic tests. Ann. Intern. Med. 120: 667-676.

JACOBSON, R. H. 1998. Validation of serological assays for diagnosis of infectious diseases. Rev. Sci. Tech. OIE 17: 469-486.

KNAPP, R. G. & M. CLINTON MILLER. 1992. Clinical Epidemiology and Biostatistics. Natl. Med. Series, Williams & Wilkins, Harwal Publ. Co., Pennsylvania, 435 pp.

LANDIS, J. R. & G. G. KOCH. 1977. The measurement of observer agreement for categorical data. Biometrics. 33: 159-174.

MARTIN, S. W. 1976. The evaluation of tests. Can. J. Comp. Med. 41: 19-25.

METZ, C. E. 1978. Basic principles of ROC analysis. Semin. Nucl. Med. 8: 283-298.

OIE. 1997. Principles of validation of diagnostic assays for infectious diseases. In: Manual of Standards for Diagnostic Tests and Vaccines. Ed: Office International des Epizooties (OIE), Paris, pp. 8-15.

SCHAFER, H. 1989. Constructing a cut-off point for a cuantitative diagnostic test. Stat. Med. 8: 1381-1391.

THORNER, R. M. AND REMEIN, Q. R. 1961. Principles and procedures in the evaluation of screening of disease. US Dept. Hlth. Educ. & Welfare, Public Hlth. Monograph N1 67, 24 pp.

SMITH, R. D. 1991. Evaluation of diagnostic tests. In: Veterinary Clinical Epidemiology. Ed: Smith, R. D. Stoneham, Butterworth-Heinemann, pp. 29-43.

SNEDECOR, G. W. & W. G. COCHRAN. 1980. Statistical Methods. Iowa St. Univ. Press, 7th Ed., 507 pp.

TARABLA, H. D. 2000. Epidemiología Diagnóstica. Ed.: Ctro. Pub., Secr. Ext., Univ. Nac. Litoral, Santa Fe, 120 pp.

VAN DER SCHOW, Y. T.; A. L. M. VERBEEK & S. H. J. RUIJS. 1995. Guidelines for the assessment of new diagnostic tests. Invest. Radiol. 30: 334-340.

YAMAMOTO, R. 1975. Characteristics an evaluation of screening tests. Mycoplasmosis Workshop Am. Assoc. Avian Pathol. 112th Annu. Mtg. Am. Vet. Med. Asocc., Anaheim, California. Univ. California Davis, EPM Dpt., 5 pp.