¿Por qué las pruebas de diagnóstico siempre mienten? O la dificultad de su validación

RAÚL C. MAINAR JAIME
Dpto. Patología Animal. Facultad de Veterinaria. Universidad de Zaragoza.

Llevo varios años impartiendo una clase en el Máster en Sanidad y Producción Porcina titulada "Interpretación de pruebas de diagnóstico" donde trato de explicar a los estudiantes la incapacidad de estas pruebas para decir toda la verdad sobre el estado de infección de un individuo. Y si no dicen toda la verdad, entonces ocultan algo. Evidentemente, no lo hacen de una manera deliberada, sino simplemente porque creo que, a día de hoy, no podemos señalar ninguna prueba de diagnóstico que sea perfecta, entendiendo por perfección que sea capaz de indicarnos con absoluta seguridad si un animal está infectado o no. ¿Implica esto que no nos deberíamos fiar de las pruebas de diagnóstico? En absoluto, solo que debemos ser capaces de interpretar adecuadamente lo que tratan de decirnos y esa interpretación a veces podría ir más allá de lo que el resultado de la prueba nos sugiere.
Las razones por las que eso ocurre son varias, la primera posiblemente yace en la propia dificultad para poner a punto una nueva prueba diagnóstica pero, una vez la prueba está desarrollada, las que considero más importantes tienen que ver con los problemas asociados con la validación de esa prueba de diagnóstico. Entender estos problemas nos debería ayudar a realizar una mejor interpretación de los resultados de las mismas.

CARACTERIZACIÓN DE UNA PRUEBA DE DIAGNÓSTICO: PRECISIÓN Y VALIDEZ

Cuando describimos una prueba de diagnóstico lo hacemos a partir de dos parámetros principales independientes: su precisión y su validez (*Greiner y Gardner, 2000*). La primera hace mención a la capacidad de una prueba para producir resultados consistentes cuando se repite en condiciones más o menos similares; esto es, la misma muestra es analizada en diferentes laboratorios, con diferentes técnicos, etcétera. En este caso hablamos de aspectos como variabilidad entre observadores (también denominada a veces como reproducibilidad). Podríamos incluso hablar de variabilidad intra-observador (la misma persona realizando la prueba varias veces sobre la misma muestra), pues para ciertas pruebas (por ejemplo, algunas técnicas como la prueba de aglutinación modificada o MAT —*modified aglutination test*— para detección de anticuerpos frente a *Toxoplasma gondii*), la experiencia del técnico es fundamental para determinar si un resultado se da como positivo o negativo.
En cualquier caso, no profundizaré más en lo que se refiere a la precisión de las pruebas de diagnóstico. Considero

que es un factor que depende fundamentalmente de la puesta a punto de la técnica por parte del laboratorio fabricante de la misma y me gustaría asumir que, en general, todas las pruebas pasan por adecuados estudios de reproducibilidad antes de salir al mercado.

Por el contrario, la validez se refiere al grado con el cual una prueba de diagnóstico mide lo que se supone debe medir, por ejemplo, si un animal está infectado o no, si presenta anticuerpos específicos, etcétera. Asumiendo una precisión aceptable, la validez sería por lo tanto el principal parámetro de interés para nosotros. Debo matizar que, aunque a lo largo del artículo nos vamos a referir de una manera general a hablar de diagnóstico de una infección, podríamos referirnos también a la detección de una característica en particular (por ejemplo, un tipo de inmunoglobulinas, proteínas de fase aguda, etcétera).

ESTIMACIÓN DE LA VALIDEZ DE UNA PRUEBA DE DIAGNÓSTICO: SENSIBILIDAD (SE) Y ESPECIFICIDAD (ES)

Cuando hablamos de validez de una prueba nos referimos a su capacidad para producir resultados correctos, y los parámetros para determinar esa validez son dos: su sensibilidad (SE) y su especificidad (ES). Cuando hablamos de SE nos referimos a la probabilidad de que la prueba identificará correctamente a los animales infectados (verdaderos positivos —VP—), mientras que la ES se define como la probabilidad de que la prueba identificará correctamente a los animales no infectados o verdaderos negativos (VN) (*Greiner y Gardner, 2000*). Así, estos parámetros miden la capacidad de discriminación que tiene la prueba para diferenciar entre individuos infectados y no infectados en una población. Se trata por lo tanto de características intrínsecas y constantes de la prueba de diagnóstico y por lo tanto no dependen ni de la prevalencia de infección ni de las características de los animales. Sin embargo, es común observar que las estimaciones de SE y ES para una misma prueba varían entre los estudios de validación publicados. Esta variación se puede atribuir principalmente a las diferencias entre las poblaciones de referencia y las estrategias (diseño) de muestreo que se han utilizado para el procedimiento de validación. Como veremos a continuación, en verdad, estimar la SE y la ES de una prueba no es fácil, y ahí radica el origen del título de este artículo.

La estimación de ambos parámetros se realiza generalmente mediante la comparación de los resultados de la prueba en cuestión frente a lo que denominamos un *Gold standard* o, para entenderlo más fácilmente, una prueba de referencia. Esta prueba de referencia sería un método de diagnóstico o combinación de métodos que determinan sin error si el animal está infectado o no. Así, podríamos crear una población de individuos infectados y no infectados de acuerdo con esta prueba de referencia (población *Gold standard*), y mediante la comparación de los resultados de nuestra prueba frente a esta población podríamos obtener una estimación directa de la SE y de la ES de la prueba (*Tabla 1*). Posteriormente, para cada valor de SE y ES se le podría calcular su intervalo de confianza del 95% (IC 95%) y tendríamos así una estimación adecuada de esos parámetros.

En el caso de pruebas de diagnóstico directo es la detección del agente o alguna de sus características singulares (mediante cultivo bacteriológico o parasitológico, PCR, etcétera), lo que determinará si la prueba es positiva o negativa, por lo que un resultado positivo no dará lugar a dudas, y por ello este tipo de pruebas se consideran 100% específicas (todos los resultados positivos son animales verdaderamente infectados, no hay FP). Otra cosa es la interpretación de un resultado negativo; es decir, de la SE de la prueba, pues si cabe la posibilidad de FN.

Por el contrario, en las pruebas de diagnóstico indirecto, que lo que cuantifican es una determinada respuesta orgánica del animal a la infección, generalmente se define primero un determinado punto de corte o valor de respuesta umbral a partir del cual se considera un resultado positivo. En este caso la SE y ES se estimarán para ese punto de corte y podremos observar tanto FP como FN. Esta sencilla aproximación suele estar en la mayoría de los casos, o quizá en todos, sesgada, principalmente por la inexistencia de pruebas *Gold standard* adecuadas que nos permitan obtener poblaciones *Gold standard* idóneas, como vamos a ver a continuación.

TABLA 1

Estimación de la sensibilidad y especificidad de una prueba diagnóstico mediante comparación con una población de referencia o *Gold standard*.

	Población *Gold standard*		
	Infectados	**Sanos**	
Prueba +	VP (verdadero positivo)	FP (falso positivo)	Total +
Prueba -	FN (falso negativo)	VN (verdadero negativo)	Total -
	Total infectados	Total sanos	TOTAL

Sensibilidad (SE)=Pr(Prueba+|Infectado) ➔ SE = VP/Total infectados.
Especificidad (ES)=Pr(Prueba -|Sano) ➔ ES = VN/Total sanos.

ESTIMACIÓN DE LA SENSIBILIDAD (SE)

Como hemos dicho, para estimar la SE necesitamos una muestra de individuos correctamente identificados como infectados (o población *Gold standard*) frente a los que utilizar nuestra prueba a valorar. Pero no sólo eso, también se requiere que esa muestra de individuos infectados que se vaya a utilizar como *Gold standard* sea representativa de toda la población de individuos infectados por el agente patógeno de interés. Y esto es así porque, como bien sabemos, los animales en el campo se enfrentan a diferentes niveles de exposición al agente patógeno que consideremos. Y estos mismos animales presentan susceptibilidades diferentes a la infección en función de su edad, estado sanitario, estado fisiológico, nivel de estrés, etcétera. Todos estos factores hacen que los resultados de la infección sean diferentes. Puede haber animales que sufran una infección localizada mientras otros sufren una infección más generalizada, lo que puede afectar a la SE de una prueba directa. Las respuestas del hospedador a la infección también pueden ser de diferente magnitud; es decir, con niveles de respuesta inmune (p. ej. anticuerpos en sangre) diferentes, lo que sin duda afectará a la SE de una prueba indirecta. Así, la capacidad de identificar correctamente a los animales infectados puede variar tanto para las pruebas de diagnóstico directo como para las de diagnóstico indirecto en función de estos factores.

Por ello, si queremos estimar de forma adecuada la SE de una prueba deberíamos considerar de alguna manera todos esos factores que pueden afectar a la respuesta del animal ante la exposición al agente. La única forma de alcanzar ese objetivo es obtener una muestra de animales infectados representativa de todos los posibles niveles de exposición al agente etiológico y todas las posibles respuestas de los animales a esos niveles de infección. Y nuestra prueba de referencia o *Gold standard* debería ser capaz de identificar todos ellos como infectados. ¿Qué prueba *Gold standard* es capaz de identificar correctamente toda la variabilidad de individuos infectados de esa hipotética población de individuos infectados por el agente de interés? Como podemos imaginar, eso es prácticamente imposible. Y sin ese *Gold standard* perfecto, difícilmente podremos crear una población *Gold standard* adecuada con la que comparar nuestra prueba, por lo que la estimación de la SE de nuestra prueba siempre estará sesgada. Por ello, a veces, cuando se comparan estudios, puede parecer que la SE de una determinada prueba varía, pero realmente lo que varía es la población utilizada como *Gold standard*. Es fácil entender que, si para la población *Gold standard* infectada utilizamos principalmente animales sintomáticos, la probabilidad de obtener un resultado positivo será mayor (mayor SE) que si en la población infectada incluimos también animales asintomáticos o recientemente infectados.

Además, teniendo en cuenta la esperada variabilidad en las respuestas a la infección en una población de animales infectados, y la imposibilidad de obtener una población *Gold standard* adecuada, otro aspecto importante a considerar es el IC95% en el que se moverá nuestra estimación de la SE. Para obtener una estimación precisa de la SE (es decir, con un pequeño IC 95%) será también necesario un gran número de animales, limitando las posibilidades de llevar a cabo un estudio adecuado.

Ante la dificultad que supone esta aproximación a la estimación del valor de SE de una prueba, la alternativa ha sido generalmente la de crear nuestra propia "población de referencia" de animales infectados frente a la cual utilizar la prueba que queremos evaluar. El gran problema asociado con este enfoque es la necesidad de garantizar la infección de todos los animales de nuestra población de referencia, lo que exige normalmente administrarles dosis altas del patógeno en cuestión y, desde luego, no necesariamente relacionadas con las que se pueden encontrar en un entorno natural. Además, se utilizan animales de unas determinadas características (de edad, peso, estado sanitario, etcétera), lo que sin duda homogeniza los resultados obtenidos, disminuyendo su variabilidad. En general, con este enfoque, que es el más asequible, estaremos sobreestimando la verdadera SE de la prueba, y si el número de animales utilizado para la evaluación es pequeño, que suele ser el caso por el coste que conlleva utilizar un número grande de animales, la precisión de esa estimación será más bien baja. Por ello, habrá que desconfiar si leemos que una prueba tiene una SE del 100%. De hecho, en ese caso, muy probablemente se refiera a una SE relativa a los resultados de otra prueba que se ha usado como referencia, pero que no será una prueba *Gold standard* adecuada, pues como he comentado, seguramente se desconocerá la SE y ES de esta última por las razones ya comentadas.

ESTIMACIÓN DE LA ESPECIFICIDAD (ES)

En este caso lo que se necesita es una población *Gold standard* negativa y, al igual que antes, representativa de la población no infectada por el agente etiológico que estemos considerando. La dificultad de encontrar esa población en este caso no es menor, pues existen muchas infecciones que cursan de forma asintomática, lo que impide conocer el estado sanitario real de un determinado individuo, salvo que este se sacrifique y se realice un estudio microbiológico meticuloso para la determinación de la presencia del patógeno en alguno de los órganos diana. Algo, evidentemente, muy costoso si se pretende utilizar un número grande de animales para obtener una estimación precisa de la ES.

Además, en el caso de pruebas de diagnóstico indirecto como las basadas en la detección de anticuerpos específicos frente a la infección, podemos encontrar un problema adicional: la existencia de animales no infectados con el agente que estemos valorando, pero sí con alguno relacionado que pueda provocar respuestas humorales similares, dando lugar a FP (*Tabla 1*). Un ejemplo clásico son las reacciones serológicas falsas positivas en pruebas de diagnóstico de brucelosis que están asociadas a infecciones por *Yersinia enterocolitica* serotipo O:9, pues esta bacteria posee una cadena del lipopolisacárido O (O-LPS) casi idéntica a la de *Brucella*, que actúa como un potente antígeno (*Jungersen et al., 2006*). Así, estos animales no infectados con el patógeno de interés pero que pueden presentar reacciones cruzadas en la prueba serológica, también deberían ser incluidos en esa población negativa *Gold standard* representativa.

Vemos como, también aquí, la posibilidad de encontrar esa población *Gold standard* negativa representativa de toda la población de animales no infectados se complica bastante. La opción mayoritaria planteada en los estudios de validación es la utilización de una población libre de infección, generalmente proveniente de centros de experimentación con alta bioseguridad y, en muchos casos, a partir de animales libres de patógenos específicos (SPF —*Specific Pathogen Free*—, en sus siglas en inglés), o incluso de animales recién nacidos antes de la toma de calostro (agammaglobulinémicos). Estos animales, al no haber estado expuestos a los microorganismos propios de su especie y no presentar anticuerpos frente a ellos, difícilmente podrán dar FN, por lo que también ahora se estará sobre estimando la ES de la prueba.

En este caso, sin embargo, disponemos de otra alternativa para la estimación más adecuada de la ES, que es seleccionar como población *Gold standard* negativa animales de poblaciones históricamente libres de la infección de interés. Ello permitirá la obtención de valores de ES mucho más cercanos a la realidad. Aunque hay que tener en cuenta que la validación de la prueba de diagnóstico debería hacerse, preferiblemente, en las poblaciones sobre las que se va a utilizar, por lo que utilizar poblaciones muy diferentes para la validación también puede introducir sesgos importantes en la estimación final de estos parámetros.

OTROS ENFOQUES PARA LA ESTIMACIÓN DE LA SE Y ES DE UNA PRUEBA

Ante las dificultades existentes para hacer adecuadas estimaciones de SE y ES por el método clásico de comparar con una población *Gold standard*, en los últimos años se han desarrollado alternativas basadas en modelos probabilísticos. Se trata de los denominados métodos latentes que permiten el cálculo de estos parámetros en ausencia de poblaciones *Gold standard*. No es objetivo de este artículo explicar estos nuevos enfoques, pero conceptualmente son relativamente fáciles de entender, pues están basados en el hecho de que los animales solo puede presentar dos estados (infectado o no infectado), que las pruebas solo pueden dar dos resultados (positivo o negativo), y que la SE y ES de las pruebas es constante (tal como hemos indicado anteriormente). Teniendo en cuenta esas consideraciones, se pueden utilizar dos o más pruebas diagnósticas en dos o más poblaciones de animales con prevalencias de infección desconocidas pero diferentes, y obtener una estimación de la SE, ES y prevalencia de las poblaciones tras la comparación de los resultados entre ellas (*Johnson et al., 2019*). Una ventaja añadida de ›

este enfoque sería la posibilidad de utilizar un gran número de animales, lo que permite diseñar el estudio con poblaciones más representativas de los diferentes niveles de infección esperados.

CONCLUSIÓN

Como hemos visto hasta ahora, la estimación de la SE y ES de una prueba diagnóstica no es fácil. De hecho, las dificultades para llevar a cabo una adecuada validación hacen que, generalmente, se utilicen enfoques tendentes a sobreestimar los valores de SE y ES. Los veterinarios tenemos que ser conscientes de ello y al menos exigir a las casas fabricantes de los kits de diagnóstico comerciales los procedimientos de validación utilizados en sus pruebas. Ello nos permitirá interpretar de una forma más adecuada los resultados de los análisis que realicemos con ellos. Sin embargo, hay que recordar que, para la correcta interpretación final de los resultados, además de los valores de SE y ES, hará falta algo más de información. Será interesante, por ejemplo, conocer el punto de corte utilizado (en las pruebas serológicas) y también el valor obtenido en el análisis para saber lo cerca o lejos que estamos de ese punto de corte. También será necesario conocer las características de los animales analizados (por ejemplo, si presentan o no factores de riesgo para esa infección, la prevalencia de la población, etcétera). Con todo ello podremos obtener una mejor imagen del estado de salud de nuestros animales y podremos estimar incluso los denominados Valores Predictivos (la proporción de verdaderos positivos/negativos del total de resultados positivos/negativos de una prueba), que nos pueden dar una mejor estimación de la probabilidad de que nuestro animal esté realmente infectado.

REFERENCIAS

- Greiner M, Gardner IA. Epidemiologic issues in the validation of veterinary diagnostic tests. *Preventive Veterinary Medicine* 2000; mayo 30; 45(1-2): 3-22.
- Johnson WO, Jones G, Gardner IA. Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Preventive Veterinary Medicine* 2019 junio 1; 167: 113-27.
- Jungersen G, Sørensen V, Giese SB, Stack JA, Riber U. Differentiation between serological responses to *Brucella suis* and *Yersinia enterocolitica* serotype O:9 after natural or experimental infection in pigs. *Epidemiology & Infection* 2006; abril; 134(2): 347-57.