Procedimiento de Microbiología Clínica

Recomendaciones de la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica

76. Estudios de evaluación del rendimiento analítico y clínico de productos sanitarios para diagnóstico *in vitro*

Editores

Emilia Cercenado Mansilla
Rafael Cantón Moreno

Coordinador

Sergio García-Fernández

Autores

Sergio García-Fernández
Andrea Vergara-Gómez
Ana María Sánchez-Díaz
Eliseo Albert Vicent

ISBN: 978-84-09-40709-5

EDITORES:

Emilia Cercenado Mansilla. Servicio de Microbiología. Hospital General Universitario Gregorio Marañón. Madrid.

Rafael Cantón Moreno, Servicio de Microbiología. Hospital Universitario Ramón y Cajal e Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS). Madrid.

SUGERENCIA DE CITACIÓN:

García-Fernández S, Vergara-Gómez A, Sánchez-Díaz AM, Albert Vicent E. 2022. 76. Estudios de evaluación del rendimiento analítico y clínico de productos sanitarios para diagnóstico *in vitro*. García-Fernández S (coordinador). Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2022.

Procedimientos en Microbiología Clínica

Recomendaciones de la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica

Editores:
Emilia Cercenado Mansilla
Rafael Cantón Moreno

76. Estudios de evaluación del rendimiento analítico y clínico de productos sanitarios para diagnóstico *in vitro*. 2022

Coordinador:
Sergio García-Fernández[1]

Autores:
Sergio García-Fernandez[1]
Andrea Vergara-Gómez[2]
Ana María Sánchez-Díaz[3]
Eliseo Albert Vicent[4]

Servicios de Microbiología. [1]Hospital Universitario Marqués de Valdecilla, Santander; [2]Hospital Clinic, Barcelona; [3]Hospital Universitario Ramón y Cajal, Madrid; [4]Hospital Clínico Universitario, Valencia.

ÍNDICE

DOCUMENTOS TÉCNICOS

PNT-ERAC-01. Evaluación del rendimiento analítico y clínico de una PCR a tiempo real para la detección de *Candida auris*
PNT-ERAC-02. Evaluación del rendimiento clínico de pruebas rápidas para el diagnóstico de SARS-CoV-2
PNT-ERAC-03. Evaluación de un sistema de estudio de sensibilidad antibiótica

ABREVIATURAS

ASM	*American Society for Microbiology*
AST	Pruebas de determinación de sensibilidad antibiótica *in vitro*
AUC	Área bajo la curva
CMI	Concentración mínima inhibitoria
Curva ROC	*Receiver Operating Characteristic curve*
CV	Coeficiente de variación
E	Especificidad
FN	Falso negativo
FP	Falso positivo
GCLP	Buenas prácticas de laboratorio clínico
LAMP	*Loop-mediated isothermal amplification*
LOD	Límite de detección
LOQ	Límite de cuantificación
LR	*Likelihood ratio*
MR	Método de referencia
ODS	*Odds ratio* diagnóstica
PS	Producto sanitario
PSDIV	Producto sanitario para diagnóstico *in vitro*
RE	Reglamento Europeo
RCT	*Randomized controlled trial*
S	Sensibilidad
S/Co	Señal respecto *cut off*
STARD	*Standards for Reporting Diagnostic Accuracy*
VN	Verdadero negativo
VP	Verdadero positivo
VPN	Valor predictivo positivo
VPP	Valor predictivo negativo

1. INTRODUCCIÓN

Buena parte de las decisiones médicas se basan en los resultados de los test diagnósticos que se realizan en los laboratorios clínicos y esas decisiones influyen en el diagnóstico y en el manejo de los pacientes. Es por ello que **los productos sanitarios para diagnóstico *in vitro*** (PSDIV) han de ser correctamente evaluados. Esas evaluaciones están sujetas a requerimientos, del mismo modo que lo están los fármacos u otros productos sanitarios destinados a mejorar la salud de los pacientes. Por estos motivos, se entiende la necesidad y la importancia de llevar a cabo estudios de evaluación de los PSDIV a fin de obtener resultados fiables y exactos que aseguren una calidad óptima para el cuidado de los pacientes, además de conservar y adecuar los recursos sanitarios.

La necesidad de una correcta evaluación de los PSDIV se ha demostrado ahora más que nunca en la actual pandemia causada por el SARS-CoV-2. Los test diagnósticos se deben evaluar atendiendo a su intención de uso, en diferentes poblaciones diana y teniendo en cuenta la prevalencia de la enfermedad. Hemos podido comprobar en tiempo real como los parámetros analíticos y clínicos de los PSDIV empleados en el diagnóstico de SARS-CoV-2 pueden ir modificándose en función de los tipos de muestra empleados, los días desde el inicio de síntomas o las modificaciones genéticas de las diferentes variantes del virus, entre otros. De esta manera se pone en evidencia que la evaluación de los PSDIV ha de ser un proceso continuo, y no únicamente circunscrito al momento del lanzamiento al mercado.

El 5 de abril de 2017, el Parlamento Europeo publicó un nuevo Reglamento sobre productos sanitarios de diagnóstico *in vitro*: (RE) 2017/746. El objetivo de este RE es establecer un marco normativo sólido, transparente, previsible y sostenible para los PSDIV que garantice un elevado nivel de seguridad y de protección de la salud, apoyando al mismo tiempo la innovación. Además, busca establecer altos estándares de calidad y seguridad para los PSDIV garantizando, entre otras cosas, que los datos generados en estudios de rendimiento sean fiables y sólidos y que la seguridad de los sujetos que participan en los estudios de rendimiento esté protegida (1).

El objetivo del presente procedimiento es proporcionar información útil y práctica que sirva de guía para llevar a cabo investigaciones que busquen evaluar el rendimiento de los PSDIV en diferentes etapas del desarrollo de los mismos. La intención, sin embargo, no es realizar una guía pormenorizada de cómo hacer una evaluación de un PSDIV para presentar datos a un organismo internacional y obtener marcado FDA (EE.UU.) o CE (Europa), ni obtener permisos de comercialización, sino facilitar su evaluación en un contexto clínico alineado con el nuevo RE 2017/746. En este procedimiento, el **documento científico** se estructura fundamentalmente en dos apartados. En el primero de ellos, **tipos de evaluaciones**, se describen cuatro fases de evaluación de un PSDIV, que abarcan la validez científica, el rendimiento analítico, el rendimiento clínico y los efectos en el paciente o en el sistema. Se establecen los fundamentos teóricos y los parámetros posibles de evaluar en cada fase. En el segundo apartado, **estudios de evaluación de PSDIV,** se toman como ejemplo diferentes tipos de PSDIV, para que de un modo concreto se repasen los posibles estudios de evaluación del rendimiento que se pueden realizar sobre un test diagnóstico. Los tres **documentos técnicos** de este procedimiento recogen una sistemática que sirve de ejemplo para realizar estudios de rendimiento analítico y clínico de un PSDIV, así como la evaluación de un sistema de determinación de sensibilidad antibiótica.

En procedimientos previos de la SEIMC ya se han abordado aspectos relacionados sobre la evaluación de PSDIV. En el Procedimiento número 48, se aborda la validación y verificación analítica de los métodos microbiológicos (2), y en el Procedimiento número 64, se trata la evaluación económica de las pruebas diagnósticas en Microbiología Clínica. Este tipo de evaluación está enmarcado en la fase cuatro de evaluación del presente procedimiento, sobre los efectos de los PSDIV en el paciente o en el sistema y, por este motivo, no profundizaremos en este aspecto en el presente documento (3).

2. CONCEPTOS BÁSICOS Y DOCUMENTOS DE INTERÉS

• **Producto sanitario (PS):** cualquier instrumento, dispositivo, equipo, programa informático, material u otro artículo, incluidos los programas informáticos, necesarios para su buen funcionamiento, destinado por el fabricante a ser utilizado en seres humanos con fines de:

- Diagnóstico, prevención, control, tratamiento o alivio de una enfermedad.
- Diagnóstico, control, tratamiento, alivio o compensación de una lesión o de una deficiencia.
- Investigación, sustitución o modificación de la anatomía o de un proceso fisiológico.
- Regulación de la concepción.

Y que no ejerza la acción principal que se desee obtener en el interior o en la superficie del cuerpo humano por medios farmacológicos, inmunológicos ni metabólicos, pero a cuya función puedan contribuir tales mecanismos.

• **Producto sanitario para diagnóstico *in vitro* (PSDIV):** cualquier producto sanitario que se utiliza para analizar muestras procedentes del cuerpo humano sin entrar en contacto con él con el fin de proporcionar información:

- Proceso fisiológico o patológico
- Deficiencias físicas o mentales congénitas
- Predisposición a una dolencia o enfermedad
- Determinar la seguridad y compatibilidad con posibles receptores
- Predecir la respuesta o reacción al tratamiento
- Supervisar las medidas terapéuticas.

Los recipientes para muestras se considerarán también PSDIV.

• **Evaluación del rendimiento de un PSDIV:** proceso mediante el cual se evalúan y analizan los datos de funcionamiento de un PSDIV.

• **Evaluación del rendimiento analítico:** estudios realizados para establecer o confirmar la capacidad de un PSDIV para detectar o medir un analito en particular.

• **Evaluación del rendimiento clínico:** estudios realizados para establecer o confirmar la capacidad de un PSDIV para generar resultados que se correlacionen con una condición física o estado fisiológico particular de acuerdo con la población objeto y el usuario previsto.

• **Reglamento (UE) 2017/746**: Nuevo Reglamento Europeo del 5 de abril de 2017 sobre los productos sanitarios para diagnóstico *in vitro* y por el que se derogan la Directiva 98/79/CE y la Decisión 2010/227/UE de la Comisión. Su función es establecer un marco normativo sólido, transparente, previsible y sostenible para los PSDIV que garantice un elevado nivel de seguridad y de protección de la salud, apoyando al mismo tiempo la innovación. Para ello establece normas elevadas de calidad y seguridad sobre los PSDIV, para garantizar, entre otras cosas, que los datos generados en estudios del funcionamiento clínico sean fiables y sólidos y que se proteja la seguridad de los sujetos que participen en estos estudios (1).

- https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32017R0746&from=es

• **Requisitos éticos y validación por el Comité de Ética**. El Comité tiene como misión garantizar el respeto a la dignidad, integridad e identidad del ser humano en lo que se refiere a la investigación con humanos, con muestras biológicas o con datos de origen humano, así como promover un comportamiento ético en la investigación. Sus funciones principales son:

1. Emitir informes sobre proyectos o trabajos de investigación que impliquen a seres humanos, sus muestras o sus datos personales atendiendo tanto a aspectos metodológicos, éticos como legales, así como velar durante su desarrollo por el cumplimiento de los diferentes aspectos éticos y las buenas prácticas de investigación.
2. Evaluar la cualificación del investigador principal y la del equipo investigador, así como la factibilidad del proyecto.
3. Velar por el cumplimiento de procedimientos que permitan asegurar la trazabilidad de las muestras de origen humano, sin perjuicio de lo dispuesto en la legislación de protección de datos de carácter personal.

Todo proyecto de investigación que implique a seres humanos, sus muestras o sus datos personales deberá ser evaluado por un comité ético. Los documentos básicos a presentar al comité para la evaluación de un proyecto de investigación, sin perjuicio de las peculiaridades del comité de cada centro de investigación biomédica, son:

1. Nombre del investigador principal, del promotor, de los investigadores colaboradores y título del proyecto,
2. *Curriculum vitae* del investigador principal.
3. Memoria científica.
4. Memoria económica
5. Hoja de consentimiento informado del paciente.

- **Documentos de interés relativos al Comité Ético**:
  - Declaración de Helsinki de la AMM – Principios éticos para las investigaciones médicas en seres humanos.
    - https://www.wma.net/es/policies-post/declaracion-de-helsinki-de-la-amm-principios-eticos-para-las-investigaciones-medicas-en-seres-humanos/
  - *International Ethical Guidelines for Health-related Research Involving Humans*.
    - ISBN: 978-92-9036088-9
  - Ley 14/2007, de 3 Julio, de investigación biomédica.
    - https://www.boe.es/eli/es/l/2007/07/03/14
  - Real Decreto 1716/2011, de 18 de noviembre, por el que se establecen los requisitos básicos de autorización y funcionamiento de los biobancos con fines de investigación biomédica y del tratamiento de las muestras biológicas de origen humano, y se regula el funcionamiento y organización del Registro Nacional de Biobancos para investigación biomédica.

- **Normas ISO:** estándares con reconocimiento internacional que tienen como objetivo establecer unos niveles de homogeneidad en relación con la gestión, prestación de servicios y desarrollo de productos en la industria. En la elaboración de este procedimiento se han empleado:
  - ISO 5725-2:2019 A*ccuracy (trueness and precision) of measurement methods and results*.
  - ISO 20776-2:2021 *Susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices.*
  - ISO 15189

- **Documentos CLSI:** Clinical and Laboratory Standards Institute.
  - EP05-A3: *Evaluation of Precision of Quantitative Measurement Procedures.*
  - EP06-A: *Evaluation of the Linearity of Quantitative Measurement Procedures.*
  - EP07-A2: *Interference Testing in Clinical Chemistry.*
  - EP09-A3: *Measurement Procedure Comparison and Bias Estimation Using Patient Samples.*
  - EP17-A2: *Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures*.
  - M100: *Performance Standards for Antimicrobial Susceptibility Testing.*

- ***Technical Guidance Series* (TGS):** documentación elaborada por la Organización Mundial de la Salud (OMS) con el objetivo de proporcionar información clara a los fabricantes de PSDIV que desean obtener una precalificación de su producto diagnóstico.
  - https://extranet.who.int/pqweb/vitro-diagnostics/technical-guidance-series.

- ***Technical Specifications Series* (TSS)**: documentación elaborada por la OMS. Los TSS establecen criterios de evaluación del rendimiento para un tipo de PSDIV en concreto. Cada documento proporciona información sobre requisitos mínimos que debe cumplir un fabricante para garantizar que el PSDIV es seguro y tiene un rendimiento óptimo.
  - https://extranet.who.int/pqweb/vitro-diagnostics/technical-specifications-series.

- ***The International Medical Device Regulators Forum*** **(IMDRF)**. Anteriormente denominado G*lobal Harmonization Task Force* (GHTF): es un grupo internacional de representantes de autoridades reguladorias de dispositivos médicos y asociaciones comerciales que proporcionan documentos de guía no vinculantes para las agencias reguladoras. Tiene como objetivo acelerar la armonización y convergencia internacional de la regulación de los productos sanitarios. El grupo cinco de trabajo de la GHTF elaboró documentación relacionada sobre la evaluación del rendimiento analítico y clínico de PSDIV.
  - https://www.imdrf.org/documents/ghtf-final-documents/ghtf-study-group-5-clinical-safetyperformance.

- **Guía STARD**: *The Standards for Reporting of Diagnostic accuracy studies*. Iniciativa que pretende mejorar la calidad de los artículos e informes de los estudios de rendimiento de los PSDIV. Comprende un listado de elementos que un artículo/informe debe incluir con el objetivo de permitir a los revisores y lectores detectar posibles sesgos de un estudio y evaluar su generalización y aplicabilidad de los resultados. También propone incluir un diagrama de flujo que esquematice el diseño del estudio. De esta manera, visualmente se entiende el modo de reclutamiento de los pacientes, el orden de realización de los test diagnósticos o el número de pacientes sometidos al test en evaluación (4).

- ***Good clinical laboratory practice*** **(GCLP)**. Guías de Buenas Prácticas de Laboratorio. Estas guías están disponibles en la web de la OMS.
  - https://www.who.int/tdr/publications/documents/gclp-web.pdf.

- **Cursos de formación gratuitos sobre Buenas Prácticas de Laboratorio**.
  - https://globalhealthtrainingcentre.tghn.org/good-clinical-laboratory-practice-course/

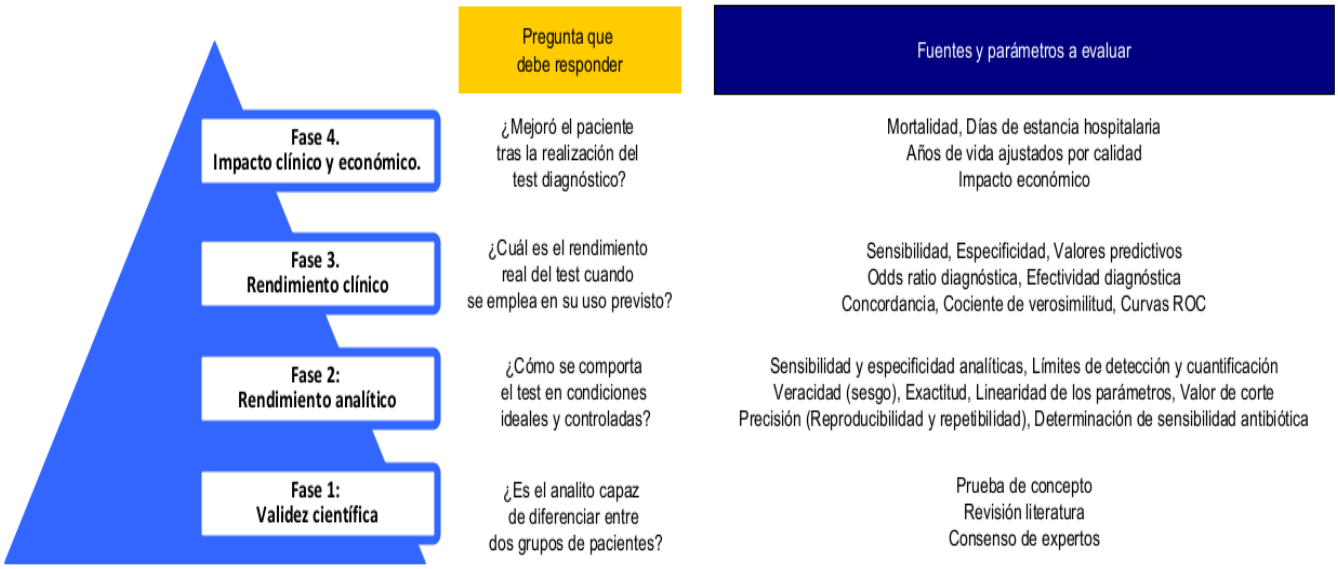
- **Enlaces de internet para cálculos de parámetros de evaluación de test diagnósticos:**
  - Cálculo de valores de sensibilidad, especificidad, valores predictivos, precisión, exactitud, etc.
    - https://onlineconfusionmatrix.com/
    - https://www.medcalc.org/calc/diagnostic_test.php
  - Cálculo del coeficiente kappa
    - https://idostatistics.com/cohen-kappa-free-calculator/#risultati
    - https://www.bioestadistica.uma.es/analisis/kappa/
    - https://www.graphpad.com/quickcalcs/kappa1/
  - Cálculo de *likelihood ratio* y probabilidad post-prueba
    - http://araw.mede.uic.edu/cgi-bin/testcalc.pl

3. TIPOS DE EVALUACIONES

Para poder comercializar un PSDIV debe realizarse previamente la evaluación de las características del funcionamiento y la seguridad en las condiciones normales del uso previsto del producto, tal y como se recoge en el artículo 56 del RE 2017/246. La evaluación del funcionamiento seguirá un procedimiento definido metodológicamente adecuado **(plan de evaluación del funcionamiento**) que permita demostrar (de conformidad con el artículo 56 y la parte A del anexo XIII del RE anterior) a) la **validez científica**, b) **el funcionamiento analítico** (rendimiento analítico o validez técnica) y c) **el funcionamiento o rendimiento clínico**. Los datos y conclusiones extraídos de esta evaluación constituirán la evidencia clínica de ese producto. Además, la evaluación del funcionamiento y su documentación se actualizarán durante todo el ciclo de vida del producto con datos obtenidos del plan de seguimiento del funcionamiento post-comercialización establecido por el fabricante (1).

Hasta la publicación del RE 2017/246 y de otras guías de armonización internacionales no existía un consenso de cuáles debían ser las fases de evaluación de un test. Se han propuesto distintos esquemas jerárquicos de evaluación de pruebas *in vitro*, de manera similar al modelo de 4-5 fases que se emplea para la evaluación clínica de nuevos fármacos (5). En la **figura 1** se resumen las fases de evaluación de los PSDIV y los parámetros que se investigan en cada uno de ellos.

Figura 1. Fases de evaluación del rendimiento analítico y clínico de un PSDIV. Elaborado con yEd 3.21.



Aunque en la mayoría de evaluaciones de un producto diagnóstico presentadas en fases, los estudios de validez científica (fase 1) preceden a los del resto de fases, no es infrecuente que los estudios de fase 2 se combinen con fase 1 o incluso los de fase 3. Del mismo modo, puede ocurrir que los tipos de estudio o diseños empleados en una de las fases se utilicen en otras (por ejemplo, en un estudio de validez científica se emplea un esquema de casos-controles, que también se puede emplear en el de rendimiento analítico o en el de rendimiento clínico) o que las muestras/pacientes incluidos en un estudio de fase 1, se empleen también en la fase 2 o fase 3 de desarrollo de ese PSDIV.

A continuación, se describen los aspectos principales de cada una de las cuatro fases de la evaluación y desarrollo de un PSDIV, indicando en cada etapa qué preguntas deben responderse y los diseños de estudio más adecuados para responderlas. En gran parte, la información para elaborar este procedimiento se ha obtenido de documentos del CLSI, normas ISO, documentos de la OMS, entre otros. En algunos aspectos, no es posible abarcar la totalidad de los posibles modos de evaluación de los test diagnósticos, pues existen documentos exclusivos para cada parámetro del rendimiento diagnóstico (por ejemplo, linealidad, cálculos de precisión, capacidad de detección, etc.), que lógicamente exceden el propósito y el alcance del presente procedimiento.

3.1. FASE 1: VALIDEZ CIENTÍFICA

Los estudios en esta fase tratan de demostrar la **validez científica de un analito**, es decir, la asociación de un analito (test diagnóstico) con un estado clínico o fisiológico. La pregunta que se trata de resolver en esta fase de evaluación es: "**¿es capaz el test diagnóstico de diferenciar entre dos grupos de pacientes?**". Esta asociación justifica el desarrollo y producción de un test diagnóstico, pues viene a demostrar que desempeña su función de manera óptima y segura.

El RE 2017/246 establece que la validez científica se demostrará basándose en una de las fuentes siguientes o en una combinación de estas:

- Información pertinente sobre la validez científica de productos que midan el mismo analito o marcador
- Literatura científica (revisada por pares)
- Opiniones o posiciones de consenso de expertos procedentes de asociaciones profesionales pertinentes
- Resultados de estudios de prueba de concepto

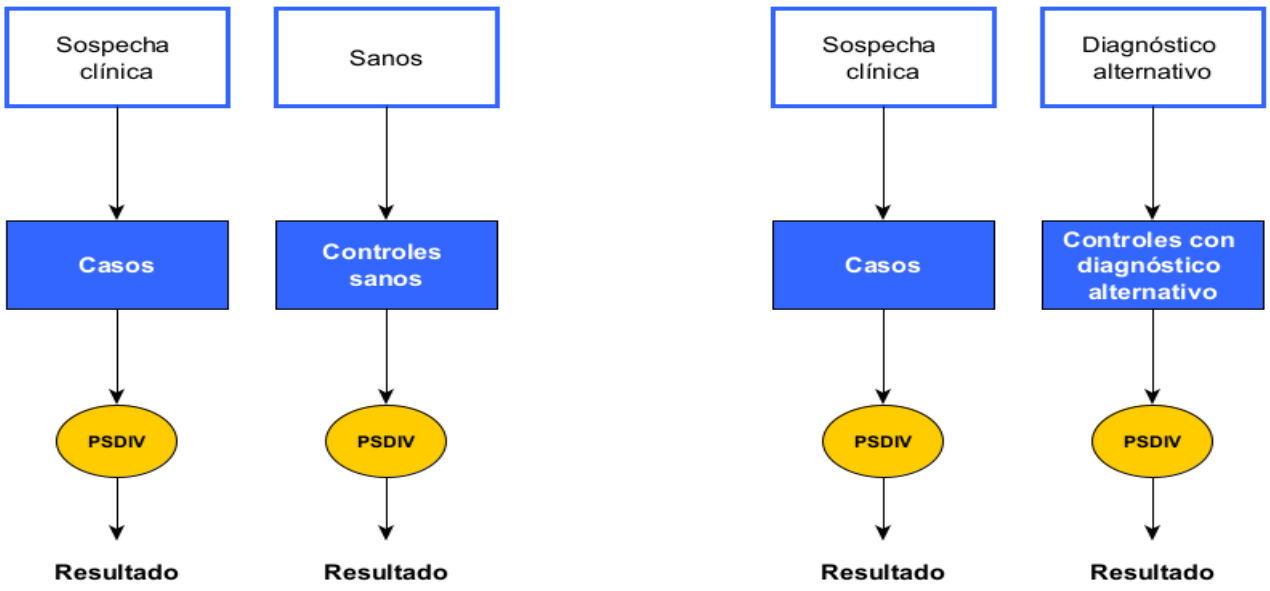
• Resultados de estudios del funcionamiento clínico

Por tanto, no es necesario comprobar la validez científica de un PSDIV si la asociación del analito con el estado clínico o fisiológico de interés ya está demostrada previamente. Podemos tomar como ejemplo la detección de galactomanano en muestras de lavado broncoalveolar como criterio para diferenciar a los pacientes con probable aspergilosis invasiva. En la literatura hay información pertinente sobre su validez, revisada por pares y su empleo se recomienda en guías clínicas de tratamiento y diagnóstico. De este modo, si un fabricante tiene una propuesta de un nuevo test diagnóstico para la detección de galactomanano, deberá justificar mediante una revisión sistemática los datos disponibles y su finalidad prevista. No será necesario demostrar nuevamente su utilidad.

Si, por el contrario, esta asociación no está demostrada, o se quiere emplear el analito/test diagnóstico para otro uso diferente, sí que será necesario demostrarla. El tipo de estudio que se puede aplicar en esta fase es el **estudio de casos-controles**, en el que se elige un grupo de personas con un estado clínico o fisiológico de interés (casos) y se compara con un grupo que no lo tenga (controles). Los controles pueden ser sujetos sanos, tomados de la población general, o pueden ser personas con un diagnóstico alternativo que produce síntomas y signos similares a los del estado clínico o fisiológico de interés. Este tipo de evaluación se puede realizar también partiendo de muestras almacenadas en los laboratorios (por ejemplo, muestras de sangre, suero u orinas) y accediendo a las historias clínicas de los pacientes para establecer los grupos de estudio. En el caso de que el test diagnóstico ofrezca resultados continuos (por ejemplo, procalcitonina, número de neutrófilos totales, etc.), la validez se comprueba si el valor medio de los resultados del test es diferente entre los casos con el estado clínico o fisiológico de interés, respecto a los que no lo tienen. En cambio, si el test diagnóstico es categórico (por ejemplo, antígeno de SARS-CoV-2), la validez se comprueba si el porcentaje de positivos o negativos es diferente entre ambos grupos (6).

En la **figura 2** se muestran dos posibles diseños de estudio de casos-controles para una evaluación de un test diagnóstico en su etapa inicial. En ambos se parte de dos grupos de estudio diferenciados. En la parte izquierda se emplea como grupo control sujetos sanos, y en la parte derecha se emplea un grupo de sujetos con un diagnóstico alternativo, que produce clínica similar al grupo definido como casos.

Figura 2. Ejemplo esquema de estudio de casos-controles para un estudio de validez científica de un PSDIV. Elaborado con yEd 3.21.



Se debe tener presente que, en esta fase de evaluación, se busca responder a la pregunta de si un analito o un test diagnóstico en cuestión es capaz de diferenciar entre dos grupos de pacientes. Por ello, es frecuente que la selección de los pacientes o muestras esté sujeta a importantes sesgos. Puede ocurrir

que sólo se hayan incluido pacientes muy graves como casos, o sujetos sanos como controles, lo que no representa a la realidad y produce resultados satisfactorios para el test diagnóstico. Aún con esta importante limitación, este tipo de estudios son fundamentales en el proceso de desarrollo de los PSDIV. Una falta de asociación entre el analito con el estado clínico o fisiológico de interés en esta fase de evaluación, con mucha probabilidad va a significar que no va a ser útil si se emplea en la práctica clínica.

Un ejemplo de estudio para demostrar la validez científica de un analito es el siguiente trabajo en el que se relacionan niveles de biomarcadores con la presencia de infección periprotésica. En el estudio se incluyeron 31 pacientes con infección periprotésica aguda, 51 con infección crónica y 139 con aflojamiento aséptico como grupo de control. El objetivo del estudio es demostrar que hay diferencias significativas entre los niveles de biomarcadores en los casos de infección respecto a los controles. Los autores encuentran que hay valores significativamente diferentes de los biomarcadores globulina, relación albúmina/globulina y fibrinógeno en los casos de infección. También comprueban la utilidad de los marcadores al obtener una elevada área bajo la curva mediante curvas ROC. De esta manera comprueban que los tres son biomarcadores prometedores en el diagnóstico de la infección protésica (7).

En otro ejemplo, esta vez sobre infección del sistema nervioso central, se busca relacionar los niveles de proteína de unión a la vitamina D en muestras de líquido cefalorraquídeo, con la presencia de infección. En el estudio se incluyeron 90 casos con infección y 212 sin infección. Se comprueba que la concentración de proteína de unión a la vitamina D aumenta significativamente en el grupo de infección. Además, también se obtiene un buen valor predictivo mediante análisis de curvas ROC (8).

3.2. FASE 2: RENDIMIENTO ANALÍTICO

Una vez que se ha demostrado que un test es capaz de distinguir a los pacientes con una determinada condición de los que no la tienen, es necesario verificar que no se trata de resultados aislados. Los estudios de rendimiento analítico o funcionamiento analítico (como se denomina en el RE 2017/246) tienen como objetivo evaluar la capacidad de un PSDIV para **detectar/medir** un **determinado analito**, es decir, estiman las propiedades intrínsecas del PSDIV. Este tipo de estudios tienen que ser capaces de responder a la pregunta "**¿cómo se comporta este test diagnóstico en condiciones ideales y controladas?**".

Los PSDIV deben diseñarse y fabricarse de manera que las características de funcionamiento respalden el uso previsto, basándose en métodos científicos y técnicos adecuados. Concretamente, el diseño deberá abordar la sensibilidad y la especificidad analíticas, la veracidad (sesgo), la exactitud (resultante de la veracidad y la precisión), la precisión (repetibilidad y reproducibilidad), el control de las interferencias (endógenas y exógenas) relevantes conocidas, los límites de detección y cuantificación, el rango de medida, linealidad, valor de corte así como el tipo y manipulación adecuada de las muestras (1). La evaluación de las características de rendimiento analítico (y también clínico) debe realizarse de acuerdo a una serie de normas o recomendaciones internacionales. Ejemplo de ello son los documentos del CLSI EP05-A3 y EP09-A3 en los que se detallan las recomendaciones para la evaluación de la precisión y estimación del sesgo de los procedimientos de medición cuantitativa. Estas características de funcionamiento deberán mantenerse durante la vida útil del PSDIV tal y como indique el fabricante. Además, siempre que sea posible, los valores expresados numéricamente deberían estar en unidades comúnmente aceptadas y estandarizadas, y entendidas por los usuarios del dispositivo.

A continuación, se detallan los parámetros analíticos que debe incluir la evaluación analítica y se aborda en un apartado independiente por su particularidad, la evaluación de un sistema de determinación de sensibilidad antibiótica.

3.2.1. Tipos de muestras que se pueden incluir en una evaluación de rendimiento analítico

El tipo y número de muestras necesarias para este tipo de estudios va a depender de las características del

PSDIV, de su intención de uso y de la característica analítica que se quiera evaluar (por ejemplo, reproducibilidad, especificidad o rango lineal, etc.). De manera general existen tres tipos de muestras que pueden emplearse en la evaluación de un test diagnóstico:

- **Muestras clínicas** bien caracterizadas y anonimizadas.

- **Muestras artificiales** (*spiked samples*) en las que se añade el analito que se desea medir a una matriz idéntica a la de las muestras problema.

- **Paneles de muestras de referencia**.

Las muestras incluidas deben representar todos los posibles estados clínicos del paciente y todo el rango de medida del analito, siendo especialmente importante la inclusión de muestras que permitan demostrar el rendimiento del test en sus límites. Es recomendable que las muestras empleadas en la evaluación sean de la misma matriz que se pretende utilizar con la prueba (por ejemplo, suero, plasma o sangre total) y que representen a distintos grupos demográficos.

En la práctica, las muestras de baja reactividad (cercanas al valor de corte), de gran valor para comprobar los límites de rendimiento, pueden ser difíciles de obtener o escasas. En estos casos, es interesante la inclusión de **muestras artificiales** (por ejemplo, muestras negativas, en la matriz correspondiente, en las que se añade exógenamente una pequeña cantidad del analito de interés para obtener una muestra de baja reactividad, o diluciones de una muestra de alta concentración) en los estudios de rendimiento analítico, siempre que el enfoque tenga una amplia justificación científica.

Como ejemplo práctico, si quisiéramos evaluar el rendimiento analítico de un test para el diagnóstico rápido de malaria basado en la detección de antígenos de *Plasmodium* spp. en sangre, deberíamos poder demostrar la equivalencia entre distintos tipos de muestra (sangre, suero, plasma) y de anticoagulante empleado. Para ello deberíamos probar el test en al menos 25 muestras positivas para *Plasmodium* spp. y en 25 muestras negativas para cada tipo de muestra y anticoagulante. La equivalencia debería determinarse para cada especie de *Plasmodium*. Si no existiera equivalencia entre los tipos de muestra, debería caracterizarse el impacto que pudiera tener en el rendimiento de la prueba, por ejemplo, calculando la sensibilidad analítica en suero/plasma y en sangre total. Esto puede lograrse comparando series de diluciones de muestras positivas emparejadas de pacientes (sangre total frente a suero/plasma extraídas del mismo paciente a la vez) o determinándolo como parte de un estudio de rendimiento clínico (9).

3.2.2. Sensibilidad y especificidad analíticas

La sensibilidad analítica se refiere a la mínima cantidad o concentración de sustancia que es capaz de detectar el test diagnóstico. Se establece determinando la mínima concentración para la cual la probabilidad de detección es ≥ 95%. El término coincide con el **límite de detección** (LOD). Un LOD bajo, indica una mayor sensibilidad analítica. Para el cálculo de este valor se emplean estándares de referencia. Si no existen estándares de referencia para el analito de medida, el LOD se calcula a partir de diluciones seriadas de muestras que han resultado positivas por un método de referencia. En el caso de que el PSDIV se pueda emplear en diferentes tipos de muestras, el LOD se tiene que calcular para cada uno de los tipos (por ejemplo, detección de ARN del VHC en plasma con EDTA o en sangre completa). En el caso de los test diagnósticos cuantitativos (por ejemplo, carga viral de VHB) se establece también el **límite de cuantificación.** Son las concentraciones inferiores y superiores en las que la precisión y la veracidad están dentro de los criterios especificados. Para su cálculo se emplea material biológico de referencia. El documento del CLSI EP17-A2 proporciona información detallada para llevar a cabo una evaluación de la capacidad de detección de un procedimiento analítico.

Como ejemplo práctico, vamos a calcular el LOD de una técnica de PCR para un microorganismo cualquiera del que no se dispone de un estándar de referencia. Se debe partir de una muestra con una cantidad conocida del analito de medida. En este caso la sensibilidad analítica se expresa en número de copias por reacción (**Tabla 1)**.

• **Paso 1**: Primero se establece un LOD provisional. Se realizan diluciones seriadas 1:10 a partir de una cantidad de 1000 copias por reacción (muy probablemente por encima del LOD), hasta 1 copia por reacción (probablemente por debajo del LOD). Se realizan entre 3 - 5 réplicas.

  ○ En este punto ya sabemos que el LOD se encuentra al menos entre 10 (tasa de detección, 33%) y 100 copias por reacción (tasa de detección, 100%. Este será el LOD provisional).

• **Paso 2:** Se realizan diluciones seriadas 1:2 a partir de 100 (LOD provisional) copias por reacción hasta 1,563 copias por reacción. Se realizan al menos 20 réplicas de cada dilución para obtener una estimación más precisa del LOD. Para obtener una positividad ≥ 95%, se debe detectar el microorganismo en al menos 19/20 réplicas de cada dilución.

  ○ En este ejemplo, el LOD es 12,5 copias por reacción. Pues se obtiene una tasa de detección ≥ 95%.

Tabla 1. Ejemplo de cálculo del límite de detección (LOD)

Paso 1: LOD provisional	
Microorganismo, copias por reacción	**Tasa de detección**
1000	3 / 3
100	3 / 3 (LOD provisional)
10	1/ 3
1	0 / 3
Paso 2: LOD definitivo	
Microorganismo, copias por reacción	**Tasa de detección**
100	20 /20
50	20 /20
25	20 /20
12,5	20 /20
6,25	7 / 20
3,125	1 / 20
1,563	0 / 20

La **especificidad analítica** se refiere a la capacidad de un test diagnóstico para detectar solo el analito para el que fue diseñado en presencia de otras sustancias/agentes en la muestra. Para esto, se llevan a cabo **estudios de interferencia**. En ellos se busca determinar los posibles resultados falsos (negativos y positivos) que puede producir el test diagnóstico por la presencia de sustancias o condiciones que interfieran en la medición de las muestras. Estas sustancias interferentes pueden ser endógenas, presentes en una muestra en condiciones fisiológicas, o patológicas (por ejemplo, hemoglobina, lípidos, factor reumatoide, etc.); o exógenas, que se refieren a sustancias ajenas al organismo (por ejemplo, fármacos y sus metabolitos, etanol, cafeína, etc.). En los estudios de interferencia se incluyen al menos cinco muestras de cinco sujetos diferentes con la posible sustancia interferente. La prueba de interferencia se realiza en muestras positivas y negativas para el analito objetivo del test diagnóstico, sin y con adición de la sustancia interferente en cantidades fisiológicamente relevantes. El objetivo es determinar cualquier variación del resultado en las muestras con la posible sustancia interferente, respecto a las que no lo tienen. En la **tabla 2** se muestra un ejemplo de estudio de interferencia de sustancias endógenas.

Tabla 2. Ejemplo de estudio de interferencia de sustancias endógenas

	Resultado PSDIV en evaluación			
Muestras	Muestra sin interferente	Muestra con interferente 1 (xx g/mL)	Muestra con interferente 2 (x/mmol)	Muestra con interferente 3
nº 1	(valor)	(valor)	(valor)	(valor)
nº 2	(valor)	(valor)	(valor)	(valor)
nº 3	(valor)	(valor)	(valor)	(valor)
nº 4	(valor)	(valor)	(valor)	(valor)
nº 5	(valor)	(valor)	(valor)	(valor)
Notas al procedimiento: Los estudios de interferencia deben realizarse con muestras con y sin el analito de interés. En las muestras con el analito de interés, su concentración tiene que ser cercana al límite de detección. Cuando se evalúan sustancias endógenas, se emplean cantidades más altas de las que se encuentran en sujetos sanos o condiciones normales. Si no se obtienen muestras con niveles apropiados de una sustancia endógena interferente, se pueden añadir de manera exógena. Cuando se evalúan sustancias exógenas, se debe emplear una cantidad al menos tres veces superior al nivel máximo en plasma.				
Modificado de: *Technical Guidance Series for WHO Prequalification – Diagnostic Assessment: Principles of performance studies* (10)				

Se realizan también **estudios de reactividad cruzada**. Estos consisten en comprobar que no se produzca reactividad cruzada con otros agentes infecciosos presentes en la muestra, tanto víricos, como bacterianos, parasitarios o fúngicos. Generalmente se comprueba la posible reactividad cruzada con microorganismos que produzcan sintomatología similar o que formen parte de la microbiota presente en la muestra. En ocasiones también se comprueba la reactividad debido a las vacunaciones recientes. Se deben incluir entre tres y cinco muestras con cada uno de los posibles agentes interferentes, y realizar la medición por triplicado. En la **tabla 3** se muestra un ejemplo de evaluación de reactividad cruzada para una RT-PCR de SARS-CoV-2.

Tabla 3. Ejemplo de estudio de reactividad cruzada en una RT-PCR de SARS-CoV-2

Virus	Nº muestras testadas	PSDIV evaluado	Test de referencia
Coronavirus humano 229	5	(valor)	(valor)
Coronavirus humano 0C43	4	(valor)	(valor)
Coronavirus humano HKU1	4	(valor)	(valor)
Coronavirus humano NL63	3	(valor)	(valor)
SARS-CoV-1	5	(valor)	(valor)
MERS-CoV	4	(valor)	(valor)
Notas al procedimiento: Se deben emplear muestras con una concentración elevada de microorganismos. Si no hay muestras clínicas con el patógeno de interés, se pueden preparar de manera artificial. La cantidad para un virus debe ser como mínimo de 10^5 UFP/mL, y de 10^6 UFC/mL para bacterias.			

Es necesario analizar diferentes tipos de sustancias interferentes en función del tipo de test diagnóstico. Como ejemplo, en la **tabla 4** se recogen las sustancias endógenas, exógenas y los agentes de reactividad cruzada que se deben incluir en un estudio de interferencia de una prueba rápida para el diagnóstico de VIH o en una prueba de cuantificación de carga viral de VHB. Si se demuestra que hay interferentes o reacciones cruzadas para las que se observa una variación de resultado, se debe hacer constancia en la evaluación del

PSDIV como una limitación del rendimiento y, además, el fabricante lo tiene que incluir en las instrucciones de uso. Por último, los resultados de cada interferente estudiado se comunicarán por separado y no como un agregado del número total de muestras analizadas en el estudio de interferencia. Se puede consultar el documento del CLSI EP07-A2 para información detallada sobre los estudios de interferencia.

Tabla 4. Estudios de interferencia y reactividad cruzada

	Sustancias endógenas	Sustancias exógenas	Reactividad cruzada
Prueba rápida VIH	Receptores de múltiples transfusiones. Embarazadas. Hemoglobina, lípidos, bilirrubina. Inmunoglobulinas elevadas.	Antimicrobianos: antimaláricos, antirretrovirales o antituberculosos. Fármacos comunes: paracetamol o aspirina. Etanol, cafeína	**Virus:** Hepatitis (A, B, C), citomegalovirus, Epstein–Barr, varicela-zoster, fiebre amarilla, influenza A y B, encefalitis por garrapatas, HTLV I y II. **Bacterias/parásitos**: *Plasmodium* spp., *Leishmania* spp., *Trypanosoma brucei*, *Mycobacterium tuberculosis*. Receptor de vacuna de la gripe. Seropositividad VIH inducida por vacunas.
Carga viral DNA VHB	Hemoglobina, lípidos, bilirrubina, albúmina. Inmunoglobulinas elevadas. Enzimas hepáticas elevadas: ALT, AST, GGT. Cirrosis alcohólica. Enfermedades autoinmunes: factor reumatoide, lupus eritematoso sistémico.	Antimicrobianos: antivirales frente a VHB, antirretrovirales, antimaláricos, antituberculosos Fármacos comúnmente usados en la región de uso del test diagnóstico. Biológicos que aumentan ácidos nucleicos circulantes.	**Virus:** VIH 1 y 2; Hepatitis (A, C, D), BK, citomegalovirus, Epstein–Barr, varicela-zoster, herpes simplex 1 y 2, HTLV I y II, parvovirus B19. **Bacterias/parásitos/hongos:** *Plasmodium* spp., *Leishmania* spp., *Trypanosoma cruzi, Trypanosoma brucei, Staphylococcus aureus, Staphylococcus epidermidis, Propionibacterium acnes, Neisseria gonorrhoeae, Candida albicans.*
Modificado de: *WHO Technical Specification Series* (11, 12)			

3.2.3. Veracidad (sesgo)

La veracidad es la **concordancia** entre la media de un gran número (infinito) de **resultados**, obtenidos mediante un método de prueba o test en evaluación, con el **valor de referencia** verdadero o aceptado como correcto. La medida de la veracidad de una medición se ve afectada por el error sistemático y se aplica, tanto en ensayos cuali como cuantitativos, cuando se dispone de un estándar o método de referencia (MR) (**Figura 3**). Puesto que no es posible realizar un número infinito de mediciones, de manera práctica, la evaluación de la veracidad se expresa cuantitativamente en términos de **sesgo/desviación**.

El sesgo puede deberse al propio método, al/los laboratorios donde se realice el test o a defectos en el diseño del estudio que conduzcan a conclusiones que no reflejan adecuadamente la verdad sobre el rendimiento de un PSDIV. Los sesgos más frecuentes que pueden producirse en la evaluación de un test son:

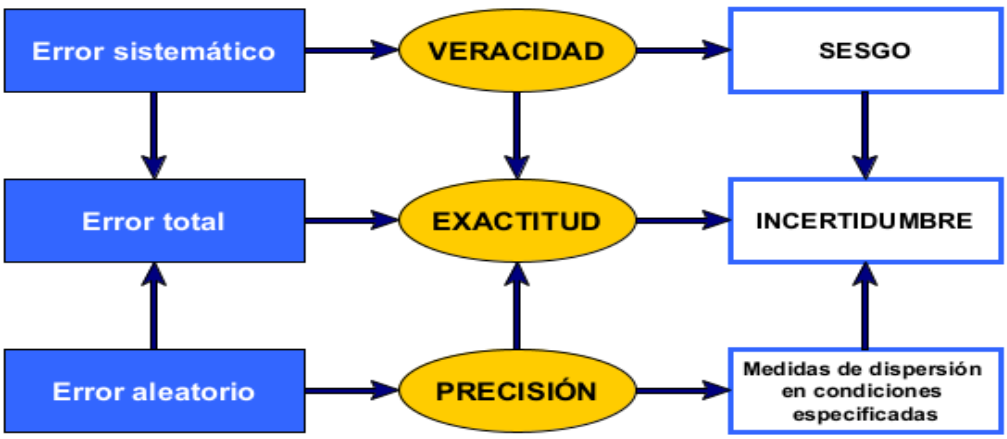
a) **Sesgo de composición del espectro**: se produce cuando se emplean muestras que no son representativas de la población diana o del uso previsto del test. Este tipo de sesgo puede reducirse empleando muestras con concentraciones de analito que representen todos los estadios de una enfermedad/condición y con una adecuada diversidad demográfica así como incluyendo muestras que contengan sustancias que puedan interferir potencialmente.
b) **Sesgo de muestreo**: ocurre cuando se analizan muy pocas muestras.
c) **Sesgo de selección**: se produce cuando se evita introducir muestras que puedan ser problemáticas (por ejemplo, muestras con título bajo del analito)
d) **Sesgo de análisis**: tiene lugar cuando el método de caracterización es insuficiente/deficiente al no permitir determinar el verdadero estado de un analito en una muestra incluida en un panel de prueba. Este sesgo puede minimizarse empleando estándares de referencia adecuados o un algoritmo de pruebas

validado para todas las muestras de un estudio. El fabricante tiene la obligación de garantizar que el estándar de referencia elegido esté validado y que la validación se lleve a cabo en un laboratorio competente; por ejemplo, uno que trabaje de acuerdo con la norma ISO 15189 o equivalente.
e) **Sesgo de revisión**, de observador o de información: ocurre cuando el profesional encargado de realizar el test/prueba conoce el resultado de una prueba anterior o el estado clínico del paciente del que proviene. Como ejemplo podemos tomar una inmunocromatografía para detectar la presencia de un antígeno de la cápside de un virus. Si el operador/técnico sabe que la muestra es positiva o proviene de un paciente infectado, puede tender a interpretarla como positiva, aún si el resultado es un positivo débil. Este tipo de sesgo puede minimizarse asignando códigos a las muestras y aleatorizando el orden en que se testan. Es importante garantizar que los técnicos que realicen las pruebas registren los resultados con el mayor detalle posible. Por ejemplo, para una prueba rápida cualitativa (que proporciona resultados de: "positivo", "negativo" o "indeterminado"), deben registrarse al menos resultados semicualitativos (por ejemplo, la intensidad de las bandas calificadas como -, +, ++, +++, ++++). Los resultados registrados de este modo son importantes para la validez del estudio, ya que permiten comprender mejor los cambios en el rendimiento del PSDIV (por ejemplo, la degradación de la señal a lo largo del tiempo).

De manera práctica, el sesgo se calcula comparando la media de los resultados **(X**) del método candidato/ en evaluación con un valor de referencia adecuado (**Xref**) mediante uno de los siguientes enfoques: a) análisis de muestras de referencia, b) experimentales y de recuperación utilizando muestras adicionadas y c) comparación mediante resultados obtenidos por otro método. Además, los estudios para evaluar el sesgo tienen que cubrir el alcance del método y pueden precisar el análisis de distintos tipos de muestra y/o diferentes niveles de analito. El sesgo puede expresarse en términos absolutos (*b=X-Xref*), en términos relativos en porcentaje (*b(%)=X-Xref/Xref x100*) o como recuperación relativa (recuperación aparente) *R`(%)=X/Xref x100*. Si no existe un material de referencia adecuado, pueden emplearse **estudios de recuperación**, que son aquellos en los que se adiciona una cantidad conocida de analito a una determinada matriz para poder dar una indicación de sesgo probable. En ocasiones los fabricantes de un determinado PSDIV pueden tener interés en demostrar que el test/método en evaluación proporciona resultados equivalentes a los de un método existente. En este caso el objetivo es establecer que no hay un sesgo significativo en relación con los resultados generados por el método existente (aunque este pudiera estar por sí mismo sesgado). Para ello pueden emplearse muestras de referencia o muestras de ensayo típicas ya que el método alternativo proporcionará el valor de referencia, lo que permite poner a prueba el método en muestras 'reales' que sean representativas de las que el laboratorio encontrará de manera rutinaria.

Un ejemplo de cálculo del sesgo se presenta en el apartado 4.3 de este procedimiento, en los estudios de evaluación de sistemas de determinación de sensibilidad antibiótica.

Figura 3. Relación entre los errores que se producen durante una medición y los parámetros analíticos veracidad, precisión y exactitud. Elaborado con yEd 3.21.



3.2.4. Precisión (reproducibilidad y repetibilidad)

La precisión de un test diagnóstico hace referencia a la **concordancia** entre los resultados obtenidos al aplicar el mismo test repetidas veces y en las mismas condiciones establecidas. La variabilidad en los resultados de una medición puede deberse al operador, al equipo y la calibración empleada, al entorno en que se encuentra el PSDIV (temperatura, humedad, contaminación, etc.), o al tiempo transcurrido entre determinaciones, entre otros. Por este motivo la precisión abarca dos conceptos:

a) **Repetibilidad:** hace referencia a la variación que se observa en una medición sobre una misma muestra, cuando el mismo operador emplea el test diagnóstico repetidamente, en las mismas condiciones y mismo lugar.
b) **Reproducibilidad**: se refiere a la variación que se observa cuando diferentes operadores emplean el mismo test diagnóstico en el mismo o diferente lugar.

Generalmente la variabilidad entre las mediciones realizadas por diferentes operadores y/o con equipos diferentes suele ser mayor que la variabilidad entre las mediciones realizadas en un corto intervalo de tiempo por un mismo operador con el mismo equipo.

La precisión depende de los errores aleatorios (inherentes a todo proceso de medición) del test diagnóstico y se evalúa valorando la dispersión entre varias mediciones realizadas sobre las mismas muestras (**Figura 3**). De manera práctica se evalúa mediante el **coeficiente de variación** (CV), que es el cociente entre la desviación de las mediciones y el valor medio de éstas expresado en porcentaje. La precisión depende generalmente de la concentración de analito por lo que debe determinarse en una serie de concentraciones del intervalo de interés. La evaluación de la precisión se lleva a cabo realizando mediciones repetidas en muestras representativas (en términos de matriz y concentración del analito, homogeneidad y estabilidad) pero no necesitan el empleo de muestras de referencia. El número mínimo especificado de repeticiones suele estar comprendido entre 6-15 para cada muestra empleada. Las réplicas también tienen que ser independientes, incluso los pasos de preparación de las muestras deben repetirse. Los ensayos de reproducibilidad deben incluir las estimaciones de variabilidad entre días, series, lugares, lotes, operadores e instrumentos (precisión intermedia). Para los test con resultados cualitativos o dicotómicos, la precisión se puede expresar como la **concordancia interobservador o entre observadores** y puede estimarse empleando el **coeficiente kappa de Cohen (apartado 3.3.2.).**

El número y tipo de muestras, laboratorios y condiciones para realizar la evaluación de la precisión de un método de medida se especifican en la norma ISO 5725-2. Las tablas 5 y 6 son dos propuestas de presentación de los resultados de un ensayo de precisión expresada como repetibilidad (**tabla 5)** y reproducibilidad **(tabla 6**).

Tabla 5. Ejemplo de tabla de recogida de datos de un ensayo de precisión (repetibilidad)

Panel QC	Nº de réplicas del test	S/Co		Coeficiente de variación
		Media	SD	
Control Negativo	(valor)	(valor)	(valor)	(valor)
QC-1 (Positivo título bajo)	(valor)	(valor)	(valor)	(valor)
QC-2 (Positivo título intermedio)	(valor)	(valor)	(valor)	(valor)
QC-3 (Positivo título alto)	(valor)	(valor)	(valor)	(valor)
Modificado de: *Technical Guidance Series for WHO Prequalification – Diagnostic Assessment: Principles of performance studies* (10)				

Tabla 6. Ejemplo de tabla de recogida de datos de un ensayo de precisión (reproducibilidad) empleando el control de calidad positivo bajo (QC-1) de un panel de referencia.

Resultados para el QC-1 (positivo bajo)	Nº de réplicas del test	S/Co		Coeficiente de variación intra-condición (%CV)
		Media	SD	
Reproducibilidad global	(valor)	(valor)	(valor)	(valor)
Inter-día	(valor)	(valor)	(valor)	(valor)
Inter-operador	(valor)	(valor)	(valor)	(valor)
Inter-lote	(valor)	(valor)	(valor)	(valor)
Inter-equipos	(valor)	(valor)	(valor)	(valor)
Modificado de: *Technical Guidance Series for WHO Prequalification – Diagnostic Assessment: Principles of performance studies* (10)				

Un ejemplo de estimación de la precisión se presenta en el estudio multicéntrico de evaluación del rendimiento analítico del ensayo Alinity m VHC (Abbot) para la detección y cuantificación de ARN de VHC. En este estudio se emplearon 406 muestras de suero y 1401 de plasma de pacientes con hepatitis crónica por VHC procedentes de 9 centros sanitarios distintos. La precisión se estableció en cuatro sitios de estudio utilizando un panel no comercial, compuesto por diferentes concentraciones (comprendidas 1,00-2,00 log10 UI/ mL) de ARN del VHC (para cada genotipo de ARN del VHC 1a, 1b y 3a) y generado por dilución de las muestras clínicas positivas en plasma humano normal. La reproducibilidad se caracterizó por un coeficiente de variación del 4,3% para el control positivo bajo y del 1,7% para el control positivo alto en los tres lotes de controles del VHC utilizados durante el estudio (**Tabla 7**) (13).

Tabla 7. Ejemplo de estimación de la precisión en el estudio multicéntrico de evaluación del rendimiento analítico del ensayo Alinity-m VHC.

	Genotipo/ control	ARN VHC diana (Log_{10} UI/ml)	Nº de réplicas	Media del ARN VHC medido (Log_{10} UI/ml)	SD (Log_{10} UI/ml)	CV (%)
Precisión	GT 1a-1	2,00	58	2,05	0,17	8,1
	GT 1a-2	1,40	54	1,38	0,16	11,9
	GT 1b-1	2,00	56	2,23	0,15	6,6
	GT 1b-2	1,40	59	1,54	0,21	13,5
	GT 3a-1	2,00	56	2,24	0,17	7,5
	GT 3a-2	1,40	56	1,65	0,16	9,8
Reproducibilidad	CPB	2,71-2,81	215	2,70	0,12	4,3
	CPA	6,02-6,23	215	6,06	0,11	1,7
SD: desviación estándar; CV: coeficiente de variación; GT: genotipo, CPB: control positivo bajo; CPA: control positivo alto *CPB y CPA, el rango refleja las distintas concentraciones dianas de los lotes de control empleados durante el estudio. Modificado de: *Chevaliez y cols (2020)* (13).						

3.2.5. Exactitud (resultante de la veracidad y la precisión)

La exactitud es una medida de la proximidad de un único resultado con un **valor de referencia** o aceptado como correcto. Además, este parámetro está relacionado con la incertidumbre de una medición ya que la desviación respecto al valor de referencia se debe a la suma del error aleatorio y sistemático. Por lo tanto,

esta medida tiene componentes tanto de la veracidad como de la precisión del test empleado (**figura 3**).

Los experimentos de exactitud deben realizarse en distintos laboratorios coordinados por un panel de expertos establecido específicamente para este propósito. Para medir esta característica del test es necesario disponer de a) un estándar de referencia, b) muestras artificiales (matriz a la que se le adiciona una cantidad conocida del analito a medir), o c) un método reconocido como exacto y empleado como método de referencia frente al que comprobar si existen diferencias estadísticamente significativas.

Para este tipo de evaluaciones se recomienda el empleo de, al menos, cinco muestras distintas que serán enviadas desde un laboratorio central (que coordine el envío a todos los laboratorios implicados en la evaluación/ validación) que garantice que las muestras que se envían a todos los laboratorios son idénticas y que permanezcan invariables durante el transporte y los diferentes intervalos de tiempo que pueden transcurrir antes de la utilización del PSDIV durante la evaluación. Desde un punto de vista estadístico, para estimar la exactitud de un método de medición se asume que cada resultado del test para una determinada muestra es la suma de tres componentes:

$$Y = m+B+e$$

- **m** es la media general (expectativa)
- **B** es el componente de sesgo del laboratorio en condiciones de repetibilidad
- **e** es el error aleatorio que se produce en cada medición en condiciones de repetibilidad

Para una información pormenorizada de los ensayos para evaluar la exactitud de una técnica diagnóstica recomendamos la consulta del documento ISO-5725.

3.2.6. Rango de medida

El **rango de medida** o **intervalo de trabajo** representa el **rango de concentraciones** de analito que puede ser detectado directamente por el test diagnóstico de manera **fiable**, sin necesidad de dilución o concentración previa. El extremo inferior del rango de medida viene definido por el límite de cuantificación (LOQ) y el superior por el valor/valores a partir de los cuales se observan anomalías significativas en la sensibilidad analítica. Dentro del intervalo de trabajo puede existir un intervalo de respuesta lineal, en el que el valor del test tiene una relación lineal con la concentración de analito (apartado 3.2.7.).

Para calcular el intervalo de trabajo de una prueba cuantitativa son necesarias muestras con concentraciones conocidas (que cubran todo el rango de interés) y blancos de muestra. Se recomienda incluir patrones de calibración cuyas concentraciones sean ± 10-20% del rango esperado y cuyas concentraciones se distribuyan uniformemente a lo largo de este rango. Los valores del analito medido por el test diagnóstico (por ejemplo, niveles de IgG frente a CMV) se representan gráficamente (eje Y) frente a los valores de concentraciones conocidas de las muestras (eje X) (curva de calibración). La evaluación del intervalo de trabajo se realizará visualmente, mediante estadística de regresión lineal y el gráfico de residuales del modelo elegido.

El rango de medida debe calcularse para cada tipo de muestra ya que las interferencias producidas por componentes de la muestra pueden dar lugar a la obtención de respuestas no lineales. Además, el test diagnóstico puede tener distinta capacidad de detección del analito problema en función de la matriz de muestra. Un ejemplo de la importancia del tipo de muestra lo tenemos en el trabajo de Hildenbrand y cols. en el que evalúan la utilidad clínica de la prueba AmpliPrep/COBAS® de Roche TaqMan® CMV para la cuantificación de carga viral de CMV en muestras no plasmáticas (BAL, LCR y orina). Para determinar el rango de medición en cada una de las tres matrices no plasmáticas se introdujeron artificialmente concentraciones conocidas (2,48–5,48 log10) de CMV en un panel de muestras clínicas de BAL, LCR u orina negativa para CMV. La carga viral de CMV se determinó por triplicado para cada muestra, en dos o tres determinaciones independientes. Al comparar la carga viral esperada con la observada (COBAS®) encon-

traron diferencias inferiores a 0,5 log para todos los niveles de carga viral y biomatrices, y obtuvieron un coeficiente de correlación cercano a uno, indicando un ajuste muy bueno en ese intervalo de trabajo (14).

3.2.7. Linealidad de los parámetros

La linealidad o intervalo lineal de una prueba es el rango de valores pertenecientes al intervalo de trabajo, en los que el resultado/valor de la prueba tiene una relación lineal (directamente proporcional) con la concentración de analito. En el documento EP06 del CLSI se especifican las recomendaciones para diseñar, analizar e interpretar los estudios de linealidad de los procedimientos de medición cuantitativa. Es de utilidad tanto para fabricantes como para agencias reguladoras que supervisan a los fabricantes de PSDIV o a laboratorios de uso final.

De manera simplificada, para evaluar la linealidad de una prueba se emplean una serie de muestras de concentración conocida o diluciones de muestras muy concentradas en el analito problema. Las mediciones o los valores obtenidos por el test en evaluación se comparan con los valores asignados y se representan gráficamente. Se calcula la ecuación de la recta para los valores obtenidos y el coeficiente de correlación. La gráfica resultante deberá ser lineal, con un coeficiente de correlación superior a 0,99. El coeficiente de correlación permite estimar la relación entre dos mediciones y si los valores experimentales siguen una función lineal.

Para la evaluación pueden usarse disoluciones patrón con cuatro o cinco concentraciones distintas de analito. Siempre que sea posible se utilizará **material de referencia certificado**, **calibradores** con valores asignados y en su defecto podrán usarse **muestras** anonimizadas de **pacientes** cuya concentración se conozca, siempre que representen concentraciones altas y bajas. Se recomienda hacer tres-cuatro réplicas de medición para cada concentración y emplear como valor final la media aritmética de esas mediciones.

En el trabajo de Nam y cols. (2021), llevan a cabo la evaluación del rendimiento analítico del sistema Alinity i para la detección de distintos analitos entre los que se encuentran IgG frente a CMV, *Toxoplasma* spp. y Rubeola o anticuerpos anti HBs. Entre las características de rendimiento evalúan el rango de trabajo y rango lineal de cada prueba encontrando rangos lineales con muy buenos coeficientes de correlación (>0,99), como se observa en la **figura 4**. (15).

Figura 4.- Intervalos de trabajo y rangos de linealidad para las pruebas de detección de anticuerpos anti-HBs, IgG de CMV, IgG frente *Toxoplasma* spp. e IgG frente a Rubéola del equipo Alinity i (Abott) (Modificado de: Nam y cols. 2021) (15).

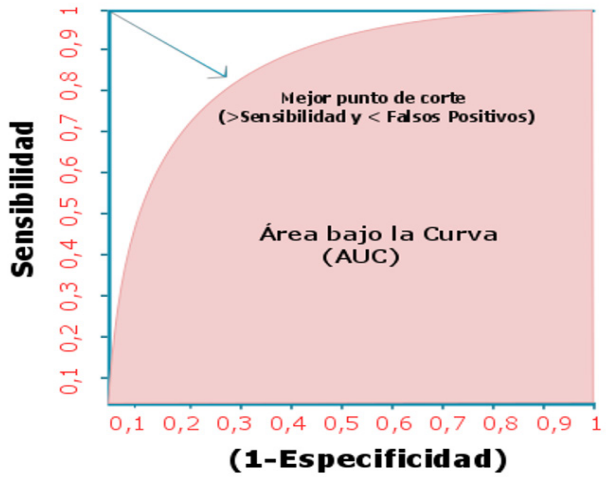
Analyte	Test range	Observed linear range	Slope	Intercept	r^2	Recovery (%)
Anti-HBs (mIU/mL)	2.0-1000.0	2.26-870.74	1.012	−5.636	0.998	95.6-103.3
CMV IgG (AU/mL)	1.1-250	1.4-236.9	1.010	0.312	0.999	97.8-103.5
Toxoplasma IgG (IU/mL)	0.2-200	0-171.8	0.983	0.781	0.995	93.8-105.9
Rubellar IgG (IU/mL)	0.5-500	0.4-500	1.027	5.402	0.992	100-109.9

3.2.8. Valor de corte (*CUT-OFF*)

Para aquellos PSDIV cuyo resultado sea una variable continua (por ejemplo, señal respecto *cut off* (S/Co), carga viral de un virus o concentración de un analito en sangre) suele ser necesario, durante las fases de evaluación 2 y 3, establecer un valor de corte o umbral que permita transformar ese resultado en una variable categórica (positivo/negativo, reactivo/no reactivo, enfermo/sano), en relación con el estándar de referencia. Por tanto, el valor **de corte** o ***cut-off*** es el umbral por encima del cual el resultado del test diagnóstico se considera positivo, y por debajo del cual, se considera negativo. La elección del valor de corte es un punto clave en el desarrollo de un test diagnóstico ya que la sensibilidad y la especificidad del mismo variarán con él.

Para analizar las compensaciones entre la sensibilidad y la especificidad en todos los valores de corte posibles suele emplearse una herramienta matemática denominada "**Curva ROC**" (*Receiver Operating Characteristic*). La curva ROC se construye al representar gráficamente la sensibilidad de una prueba diagnóstica (eje Y) frente a los resultados falsos positivos (*1- especificidad; eje X*) que adoptan valores comprendidos entre 0 y 1 (0-100 %). De esta manera cada punto representa la probabilidad de diagnosticar correctamente a sanos y enfermos. Para seleccionar el valor de corte que proporciona la sensibilidad y la especificidad más alta conjuntamente puede emplearse el **índice J de Youden** (*J=S+ E -1*). Este estadístico adquiere valores comprendidos entre -1 y 1 y está definido para todos los puntos de una curva ROC. El valor de corte óptimo será aquel con un mayor índice J, que se corresponde gráficamente con el punto de la curva ROC más cercano al ángulo superior izquierdo (valor 1,0 correspondiente a S y E del 100%) (**Figura 5**). Un PSDIV que tenga una baja efectividad diagnóstica, tendrá un índice de Youden próximo a cero, mientras que un PSDIV perfecto tendría un índice de Youden de 1. La principal desventaja de emplear este índice es que no es sensible a cambios en sensibilidad (S) y especificidad (E). De esta forma un PSDIV A, con una S de 0,9 y E de 0,4 y otro PSDIV B, con una S de 0,6 y E de 0,7 tendrían el mismo índice de Youden (0,3). Por lo tanto, no debe emplearse como un índice aislado a la hora de evaluar un PSDIV.

Figura 5. Curva ROC y área bajo la curva.



A la hora de interpretar una curva ROC es importante también el concepto *área bajo la curva* (*area under the curve*, AUC). El AUC representa la probabilidad de que la prueba clasifique correctamente a un par de individuos, sano y enfermo, seleccionados al azar. Los valores del AUC oscilan entre 0,5 (representa el azar) y 1, considerándose la prueba más discriminativa cuanto más se acerque a un AUC de 1. Generalmente se considera que las pruebas cuyas curvas ROC tengan AUC >0,9 son muy discriminativas,

entre 0,7-0,9 moderadamente discriminativas y 0,5-0,7 con bajo poder discriminativo (16). Por tanto, la capacidad discriminatoria de una prueba disminuye al disminuir el AUC de la curva ROC. Cuando la curva coincide con la diagonal (AUC=0,5) la prueba no tiene capacidad discriminatoria y valores por debajo de 0,5 aparecen cuando existe una clasificación incorrecta de sanos y enfermos. El AUC, por ser un estimador muestral de un parámetro poblacional se debe informar siempre junto con su IC 95 %.

Las curvas ROC, además de ser útiles en la elección de un valor de corte, se emplean en la evaluación de la capacidad discriminativa de una prueba diagnóstica (capacidad para discriminar sanos de enfermos) y en la comparación de la capacidad discriminativa de 2 o más test diagnósticos cuyos resultados se expresan como escalas continuas.

Como ejemplo de empleo de curvas ROC para elegir un valor de corte, en el estudio de Ortiz de la Tabla y cols. (2018) evalúan la precisión diagnóstica del inmunoensayo quimioluminiscente (CLIA) automatizado (VirClia®) para la detección de IgG e IgM frente a *Mycoplasma pneumoniae* en pacientes con neumonía adquirida en la comunidad (NAC). Para ello miden los niveles de anticuerpos IgG e IgM en sueros de fase aguda y convaleciente de 137 pacientes con NAC mediante ELISA cuantitativo y esta técnica CLIA. Construyeron curvas ROC con los cocientes de IgG e IgM en fase convaleciente y aguda (C/A) para cada técnica observando que los resultados eran comparables. Además, emplean estas curvas ROC para calcular el valor de corte óptimo del cociente C/A para cada tipo de Ig. En concreto, para el CLIA, el mejor cociente C/A para IgG fue de 2,6 (S: 94,9%; E: 99,9%), y para IgM de 1,4 (S: 65,8%; E: 100%) (17).

3.2.9. Estudios de evaluación de métodos de determinación de sensibilidad antibiótica

Las pruebas de determinación de sensibilidad antibiótica *in vitro* (AST) se realizan a partir de microorganismos aislados de un foco infeccioso, sobre todo cuando pertenecen a especies que puedan ser resistentes a los antibióticos empleados más frecuentemente. Este tipo de pruebas son además muy útiles en los estudios de vigilancia epidemiológica y en la comercialización de nuevos antibióticos a la hora de comparar su actividad intrínseca con la de otro antibiótico existente. Generalmente se utilizan técnicas basadas en procesos de dilución que tienen como objetivo determinar las concentraciones mínimas inhibitorias o CMI (concentración más baja de un agente antimicrobiano (mg/L) que previene la aparición de colonias visibles de un microorganismo bajo unas condiciones determinadas *in vitro*). Estas técnicas de dilución pueden realizarse en caldo o en agar, siendo las primeras las más empleadas, especialmente la que se realiza en micropocillos de placas de microtiter (microdilución).

3.2.9.1. Recomendaciones de la ASM. Tradicionalmente, la evaluación de un nuevo sistema para la determinación de sensibilidad antibiótica se ha realizado siguiendo las recomendaciones de documentos de consenso como el Cumitech 31 A de la ASM (*American Society for Microbiology*). De acuerdo con este procedimiento, la evaluación debe permitir la detección de **errores categóricos**, es decir aquellos debidos a discrepancias en la categorización de aislados como sensibles, intermedios o resistentes entre los métodos en evaluación o entre el método en evaluación y el método de referencia. Para la evaluación de un nuevo sistema de AST recomendaban utilizar como técnica de referencia la **microdilución** y, aunque la técnica de difusión con disco podría emplearse, no se recomendaba al no permitir el cálculo del acuerdo esencial. La evaluación de un nuevo sistema de AST puede realizarse empleando un método de referencia o sin él.

A) Evaluación empleando un método de referencia. Cuando se evalúa un nuevo sistema de AST usando como comparador un método de referencia, se recomienda ensayar al menos **cien** aislados por panel y que, al menos la mitad de ellos, sean resistentes a algún antibiótico. Para cuantificar el acuerdo y las discrepancias entre ambos sistemas se calculan los siguientes parámetros/índices:

- **Acuerdo esencia**l (EA): número de acuerdos (CMI) ± 1 dilución doble entre el nuevo test y el de referencia.

$$AE = \frac{nº\ de\ acuerdos\ \pm 1\ dilución}{nº\ \ total\ de\ microorganismos} x100$$

- **Acuerdo categórico** (CA): acuerdo en los resultados interpretados (S/I/R) entre el nuevo sistema de AST y el método de referencia empleando criterios de interpretación CLSI.

$$CA = \frac{nº\ de\ acuerdos\ (S, I\ , R)}{nº\ \ total\ de\ microorganismos}\ x\ 1$$

El AE y el AC pueden calcularse para todos los microorganismos y antibióticos de manera combinada o de manera individual para cada antibiótico.

- **Errores muy graves** (VME): discrepancias originadas cuando el método de prueba categoriza como sensible a un aislado que es resistente por el método de referencia. Clínicamente son los errores más graves y solo pueden detectarse incluyendo microorganismos resistentes a cada antibiótico.

$$VME = \frac{nº\ de\ discrepancias\ VME}{\begin{array}{c} nº\ \ total\ de\ microorganismos\ R\ por\ método \\ actual\ verificado\ por\ método\ de\ referencia \end{array}} x100$$

Los VME deben determinarse empleando al menos 35 aislados resistentes.

- **Errores graves** (ME): discrepancias originadas cuando el método de prueba categoriza como resistente a un aislado que es sensible por el método de referencia. Este tipo de errores se detectan analizando únicamente los microorganismos sensibles a cada antibiótico.

$$ME = \frac{nº\ de\ discrepancias\ ME}{\begin{array}{c} nº\ \ total\ de\ microorganismos\ S\ por\ método \\ actual\ verificado\ por\ método\ de\ referencia \end{array}} x100$$

- **Errores menores** (MinE): discrepancias originadas cuando el método de prueba categoriza como intermedio a un aislado que según el método de referencia es sensible o resistente y viceversa. Puede calcularse para todos los microorganismos y antibióticos de manera combinada o de manera individual para cada antibiótico.

$$ME = \frac{nº\ de\ discrepancias\ MinE}{nº\ total\ de\ microorganismos} x100$$

Para el cálculo de ME y MinE deben emplearse al menos 100 aislados.

En la **tabla 8** se esquematizan los principales parámetros que deben calcularse en la evaluación de un sistema de AST y sus valores recomendados según la ASM.

Tabla 8. Parámetros que deben evaluarse en un sistema de AST.

Parámetro	Método en evaluación	Método de referencia	Valor recomendado
Acuerdo esencial (EA)	Acuerdo en CMI ± 1 dilución /n total aislados		**≥ 90%** (combinados)
Acuerdo categórico (CA)	Acuerdo en categoría (S/I/R) /n total de aislados		
Errores muy graves (VME)	**S**	**R**	**≤ 3%**
Errores graves (ME)	**R**	**S**	**≤ 3%**
Errores menores (MinE)	**I**	**S/R**	**≤7%** (combinados ME+MinE)
	S/R	**I**	

Si los límites recomendables para estos parámetros se sobrepasan para cualquier antibiótico, el test de prueba debe considerarse no verificado y retirarse para ese uso o llevar a cabo medidas correctivas junto con el fabricante para tratar de solventar las discrepancias. Tras las acciones correctivas el sistema de prueba revisado debe utilizarse en paralelo con el método de referencia en al menos 20 aislados.

B) **Evaluación sin método de referencia.** En este tipo de evaluación, más habitual en laboratorios con recursos limitados, se comparan los resultados del test de prueba con el método en uso en el laboratorio. Los criterios que debe cumplir este nuevo test son idénticos a los definidos cuando se emplea un método de referencia, salvo que el método de prueba puede no ser incorrecto cuando existen discrepancias entre ambos sistemas. En este caso se pueden calcular el **acuerdo esencial** (acuerdo CMI ± 1 dilución) y **categórico** (acuerdo en resultados interpretados, S/I/R entre ambos sistemas), la **tasa de errores mayores** (un sistema categoriza como S y el otro como R), **tasa de errores menores** (un sistema da un resultado I y el otro S o R) pero no el porcentaje de errores muy graves al no poderse asumir que los resultados del test en uso fueran los correctos.

De manera general se recomienda que los ME sean ≤ 5 % (≤ 10 % de manera combinada ME y MinE) y que el acuerdo esencial y categórico sean igual o superiores al 90%.

La utilización de los criterios de la ASM se ha empleado en numerosos trabajos de investigación utilizando criterios de interpretación tanto del CLSI como del EUCAST. Sin embargo, con el cambio de concepto de la categoría I (sensibilidad con exposición aumentada) el cálculo de errores menores no tendría sentido.

3.2.9.2. Recomendaciones ISO 20776-2. Este documento establece los criterios de evaluación de los sistemas para estudio de la sensibilidad antibiótica utilizando como comparador el método estándar de microdilución en caldo (ISO 20776-1:2019). Para ello deben emplearse cultivos puros de bacterias aerobias, que crezcan fácilmente *overnight* y que puedan crecer en pocillos de placas de microdilución con medio Mueller-Hinton (volúmenes ≤ 200). Una de las principales diferencias de esta norma con la previa (ISO 20776-2:2007) y con las recomendaciones de la ASM es que no contempla el cálculo de parámetros que precisan la interpretación del valor de CMI (categorización de un aislado como S, I, R). Es decir, desaparecen el acuerdo categórico y los errores graves, muy graves y menores. Con este cambio se pretende evitar que el rendimiento tenga que ser reevaluado cada vez que cambien los puntos de corte y de esta forma se tengan en cuenta únicamente parámetros asociados al funcionamiento del ensayo y no a la interpretación de resultados.

En esta norma se recogen también las especificaciones para la evaluación de sistemas de determinación de sensibilidad antibiótica cualitativos, que no se detallan en este procedimiento.

Metodología

Método de prueba: tiene que permitir la determinación de la CMI y que ofrezca al menos un rango de 4 diluciones dobles consecutivas y que permita la determinación del EA.

Método de referencia: para la evaluación debe emplearse un método de referencia (descrito en ISO 20776-1) que se utilizará de manera simultánea al sistema de prueba en todos los lugares de evaluación o en un único sitio para todos los aislados incluidos en el estudio. Si el sistema de referencia y de prueba se utilizan en el mismo sitio, deben utilizar el mismo día con el mismo inóculo.

Selección de aislados y preparación de inóculo: deben incluirse al menos 300 aislados clínicos (al menos 100/sitio: 25 aislados contemporáneos y 75 de *stock*) que representen el mayor número de géneros y especies posible. Es recomendable que los aislados no estén relacionados y que representen distintos grados de sensibilidad a los distintos antibióticos. Estos aislados pueden ser de ***stock*** (aquellos recuperados de una muestra clínica que se ha guardado o que pertenece a una colección) o **aislados contemporáneos** (aquellos de muestras clínicas recuperados en los seis meses previos y que apenas se han subcultivado). Cuando sea posible se ensayarán al menos 25 aislados cuya CMI se encuentre en escala. La estandarización del inóculo para el test de prueba se llevará a cabo siguiendo las instrucciones de uso del fabricante.

Control de calidad: deben incluirse aislados de control de calidad y testarse cada día que se utilice el test de prueba. Las CMIs de estos aislados deben estar incluidas en los rangos descritos en CLSI M100 o en el documento de control de calidad del EUCAST.

Reproducibilidad: se comprobará ensayando un mínimo de diez aislados por triplicado durante al menos tres días en cada laboratorio que participe en la evaluación. El número de aislados en escala debe indicarse en el informe final.

Solución de discrepancias: se considera un resultado discrepante cuando la CMI de un microorganismo por el método de prueba se encuentre fuera del EA (± 2 diluciones superior o inferior a la del método de referencia). Si la discrepancia se debe a un error técnico (contaminación, condiciones de incubación inadecuadas, error del propio sistema...) puede solventarse repitiendo tanto el método de prueba como el de referencia y reemplazando el nuevo resultado por el problemático. No obstante, cuando las discrepancias no se deban a errores técnicos, tienen que resolverse re-testando por triplicado o duplicado (si se incluye la medida discrepante) en una sola vez el/los aislados problemáticos. Debe calcularse la moda y resolver si existe un acuerdo esencial con las nuevas determinaciones. La interpretación final de los resultados obtenidos por el método de prueba y el de referencia se basará en la comparación de la moda o la mediana de tres réplicas de ambos métodos (**tabla 9**).

Tabla 9. Interpretación de resultados adicionales para la solución de discrepancias. (Modificado de la norma ISO 20776-2)

CMI (mg/L)							
Aislado	Método de prueba			Método de referencia			EA de medias/ medianas
	Inicial	Repetición	Moda/ Mediana	Inicial	Repetición	Moda/Mediana	
A	1	1,2,2	2	4	2,4,4	4	Si
B	1	1,2,4	2	4	2,4,4	4	Si
C	8	2,2,4	2	1	1,2,4	2	Si
D	1	1,1,2	1	4	2,4,4	4	No

Criterios de aceptación y análisis de datos

Para la evaluación deben calcularse el acuerdo esencial (%, de la misma manera que recomienda la ASM) y el sesgo. El EA debe ser superior o igual al 90% y presentarse además de forma separada para microorganismos grampositivos, gramnegativos fermentadores y gramnegativos no fermentadores. El sesgo tiene que ser ≤ +/- 30 % y en caso de que se exceda ese porcentaje y no se resuelvan las discrepancias mediante el re-ensayo por triplicado, deberá añadirse un comentario en el etiquetado del test. Se recomienda llevar a cabo un análisis minucioso de las discrepancias para determinar si afectan solo a un tipo de microorganismos o si tienen que incluirse como una limitación del test de prueba **(ver PNT-ERAC-03).** Las CMIs de las cepas de control de calidad deben estar en rango en al menos un 95% de los resultados de todos los laboratorios que participan en la evaluación durante el periodo de estudio.

Los resultados de reproducibilidad del test de prueba deben estar dentro del rango de ± 1 dilución de la moda/mediana y no ser superiores a tres diluciones dobles para un antibiótico determinado en ≥95% de los resultados.

Una vez finalizada la evaluación es necesario redactar un informe en el que se incluyan las características de rendimiento del sistema en prueba para cada grupo/género de microorganismos y cada antimicrobiano. El rendimiento de la prueba debe calcularse para cada antimicrobiano y deben nombrarse los laboratorios que han participado en la evaluación.

3.3. FASE 3: RENDIMIENTO CLÍNICO

Los estudios de rendimiento clínico de un test diagnóstico tienen como objetivo evaluar si dicho test es capaz de lograr la **finalidad prevista** que afirma el fabricante cuando se emplea **en la población de uso** a la que va destinado. Se trataría de responder a la siguiente pregunta: **"¿cuál es el rendimiento real del test cuando se usa según lo previsto?"**. En este caso, las muestras quizás no sean tan perfectas como las empleadas para los estudios de validez analítica.

3.3.1. *Inputs*

Es importante pensar y organizar correctamente estos estudios antes de comenzar. Los estudios en esta fase deben definir claramente los aspectos que se describen a continuación.

• **Test diagnóstico evaluado**. Puede ser una prueba, la misma prueba que se repite en diferentes momentos, o incluso una combinación de diferentes pruebas, como una prueba con menor especificidad seguida de una prueba con mayor especificidad en los inicialmente positivos. Idealmente, lo que se evaluará será el algoritmo completo. Se deberá aportar la siguiente información cuando proceda:

- **Nombre completo del test y del fabricante**. Puede ser una evaluación de una única prueba o una comparación de varios test. Es importante también indicar la versión de la prueba y de las instrucciones de uso del fabricante. Si participan varios centros en el estudio, todos deberán usar la misma versión. Si hubiera cambios a lo largo del estudio, sería necesaria una validación para demostrar que el rendimiento clínico no se ve afectado.
- **Tipo de muestras** (sangre, suero, plasma, frotis nasofaríngeo, biopsia cutánea, etc.), **protocolo de recogida** (limpieza y desinfección previa de la zona, horario concreto, tipo de recipiente, etc.) y **condiciones de manipulación** (en cabinas de bioseguridad, en la cama del paciente, etc.), **transporte** (a temperatura ambiente, refrigerada, etc.) y **almacenamiento** (temperatura ambiente, refrigeración, congelación, con o sin estabilizantes, duración de la estabilidad, etc.). Se deben describir también los criterios de aceptación y rechazo de las muestras. Hay que tener en cuenta que los resultados obtenidos en la evaluación con un determinado tipo de muestra, no son extrapolables a otros tipos de muestra, sino que debe ser también evaluado en otros tipos de muestra.

- **Diana** del ensayo: ADN, ARN, antígeno, IgG, IgM, etc. Se debe indicar claramente la diana del test o hacer referencia a las instrucciones del proveedor. Para las pruebas de biología molecular, se deberá indicar la región diana del genoma donde se unirán los *primers*; para las pruebas de antígeno, el antígeno específico que va a detectarse; para las pruebas serológicas, las proteínas a las que se dirige el anticuerpo, el tipo de inmunoglobulinas detectadas (IgA, IgG, IgM, totales), así como si son anticuerpos neutralizantes o no neutralizantes.
- **Fundamento** de método: PCR, RT-PCR, fluorescencia, inmunocromatografía, etc.
- **Plataforma** necesaria para su uso y requisitos previos (extracción de ácidos nucleicos, inactivación por calor, centrifugación, etc.).
- Resultados de la **evaluación analítica** del método. Los indicadores clave de rendimiento analítico de las pruebas utilizadas deben ser conocidos antes de comenzar la evaluación clínica (ver punto 3.2.).
- **Interpretación** de los resultados. Debe definirse la interpretación para los resultados positivos, negativos e indeterminados. Preferiblemente, los puntos de corte se seleccionan a priori, por ejemplo, basándose en las indicaciones del fabricante, o en estudios previos. Debe interpretarse sin conocimiento del estándar de referencia, a menos que el resultado de la prueba no requiera de ningún juicio (subjetividad) por el intérprete. Si no existen datos previos para determinar los puntos de corte, habrá que establecerlos con los datos del estudio, pero se requiere una validación externa del punto de corte óptimo en una población seleccionada y representativa para evitar sesgos.
- Se debe describir también qué hacer con los **resultados inválidos o indeterminados**, así como reportar estos casos.

• **Intención de uso**. Antes de empezar la evaluación de una prueba, los investigadores deben definir para qué va a usarse y qué lugar ocupa el test evaluado en el algoritmo que ya existe en torno al problema para el que va a usarse:

- Diagnóstico.
- Cribaje.
- Monitorización.
- Pronóstico.
- Respuesta al tratamiento.
- Estudiar el estado inmune.
- Estratificación del estado de la enfermedad.
- Selección de pacientes para una determinada terapia o tratamiento.
- Etc.

• **Población objetivo**. Los estudios deben realizarse en individuos seleccionados dentro de la población en la que se utilizará la prueba, según el uso previsto. Para realizar la evaluación se debe definir el método de inclusión de los participantes en el estudio, incluyendo criterios de inclusión y exclusión, con el objetivo de reclutar participantes representativos de la población objetivo. Idealmente, si el uso previsto de la prueba es en un entorno sanitario, se deberían incluir individuos consecutivos de la población objetivo sin previo conocimiento de si los individuos tienen la condición a estudio o no. Para estudios en los que el uso previsto de prueba es tomar decisiones de salud pública, se puede usar una muestra aleatoria representativa de la población objetivo. Los estudios que incluyen individuos con una enfermedad conocida y controles sanos introducen un sesgo de selección y efectos relacionados con el espectro clínico de la enfermedad. Es importante también indicar el periodo de tiempo durante el que se reclutaron los pacientes.

• **Requerimientos**. Cada tipo de prueba tiene diferentes requisitos en términos de equipamiento y otros recursos necesarios, experiencia del operador, tipos de muestras, almacenamiento de muestras y tiempo de respuesta.

• **Prevalencia del problema en la comunidad**.

• **Lugar de realización del test**. El tipo de centro sanitario, como laboratorios de referencia, centros de salud, autotest, etc., así como la ubicación geográfica, con diferentes patologías endémicas, características

climáticas, variabilidad genética de los individuos o desarrollo socioeconómico del lugar.

• **Test de referencia**. Se trata del mejor método disponible para establecer la presencia o ausencia de la condición o característica de interés. Puede ser una sola prueba o método, o una combinación de métodos y técnicas, incluido el seguimiento clínico. Debe evitarse el uso de test diagnósticos *in house,* siempre y cuando no estén perfectamente validados. El estándar de referencia debe separar claramente individuos que tienen la condición a estudio de aquellos que no; por ejemplo, aquellos que tienen o han tenido la infección de aquellos que no tienen o no han tenido la infección, o aquellos que son infecciosos de aquellos que no son infecciosos. Independientemente del uso previsto, en estudios de rendimiento clínico, la interpretación de la prueba o pruebas a estudio y de la prueba estándar de referencia debe realizarse sin conocer los resultados de la otra prueba (o pruebas). La exactitud de la prueba a estudio se calcula asumiendo que el estándar de referencia es 100% exacto. Si esto no es cierto, las estimaciones de precisión del nuevo método podrán verse afectadas. Por ello, hay que tener en cuenta las limitaciones del test de referencia a la hora de interpretar los resultados. Todas las muestras deben probarse utilizando tanto la prueba en investigación como la prueba de referencia. En general, la prueba a estudio y la de referencia deben realizarse lo más cerca posible en el tiempo y, si es posible y aplica, sobre la misma muestra. En cualquier caso, debe recogerse la fecha de la toma de muestra y del análisis. Fidalgo y cols. (2021), evaluaron un sistema semi-automático basado en PCR a tiempo real para el diagnóstico de infecciones gastrointestinales. En este estudio no se analizaron las mismas muestras, sino dos periodos de tiempo diferentes: entre 2013-2018 las muestras fueron procesadas con los métodos estándar y a partir de 2019, desde su implementación en el laboratorio, con el sistema semi-automático (18).

En el caso de que no haya un estándar de referencia, la clasificación de los sujetos del estudio se realizará en algunos casos en base a información sobre el riesgo de exposición, otras pruebas (biomarcadores, pruebas de imagen, etc.), respuesta al tratamiento y resultados clínicos en el seguimiento. En ocasiones, para test muy novedosos, puede ser que no se disponga de un estándar de referencia o que éste ofrezca una menor sensibilidad, en este caso resulta válido usar referencias combinadas. Otra posibilidad, en ausencia de un estándar de referencia, es aportar el resultado del rendimiento del test como una proporción o porcentaje de concordancia positiva o negativa (PPA o NPA) con respecto a su comparador (19).

Debe desarrollarse un protocolo que recoja todo lo anterior, además de los objetivos y el diseño del estudio. Además, es importante indicar dónde se va a realizar el estudio (laboratorio de hospital, laboratorio del fabricante, cama del enfermo, etc.), así como quién lo va a llevar a cabo (profesional de laboratorio, personal ajeno al laboratorio, etc.).

• **Tamaño muestral.** El tamaño de la muestra debe ser el número de individuos incluido en el estudio, no el número de muestras testadas. Si se incluye más de una prueba en algunos individuos incluidos en el estudio, la prueba repetida no debe incluirse en las estimaciones de sensibilidad y especificidad. Sin embargo, sí que pueden incluirse varias muestras del mismo individuo para el cálculo de la sensibilidad y especificidad cuando se realice el estudio a lo largo del tiempo. Cuando se repita una prueba, se debe informar el motivo de la repetición y notificarse. Si se está analizando más de una prueba en todos los individuos, el tamaño de la muestra sigue siendo el número de individuos incluidos en el estudio. En el caso de baja prevalencia de la condición a estudio, hay diseños de estudios alternativos que pueden aplicarse (20).

• **Tipos de muestras**. Deben incluirse muestras clínicas obtenidas de la población objeto de uso del test diagnóstico, tanto que presenten la condición a estudio como de controles. Serán muestras preferentemente consecutivas. Puede tratarse de muestras que se recogen específicamente para realizar el estudio o muestras que están ya almacenadas. En este caso, deberán indicarse las condiciones y el tiempo que han estado almacenadas. Será necesario indicar también, si una vez terminado el estudio, las muestras se guardarán en un biobanco por ejemplo o se destruirán.

• **Aspectos éticos.** Los derechos, seguridad y bienestar de los participantes del estudio deben protegerse. El estudio deberá generar información nueva y los beneficios para la salud deberán superar los riesgos. La confidencialidad de los datos de los pacientes debe ser respetada y se debe solicitar el consentimiento infor-

mado (incluyendo objetivos de estudio, derechos y riesgos de los participantes, e información de contacto) antes de la recogida de muestras y datos personales, así como la aprobación del comité de ética pertinente.

3.3.2 Análisis y presentación de los resultados. *Outputs.*

Los informes de evaluación clínica deben seguir las directrices STARD (**tabla 10**) e incluir un diagrama de flujo como el propuesto en este procedimiento, adaptado a cada estudio, para informar sobre el número de personas incluidas en el estudio, el número de individuos excluidos, el número de personas incluidas cuyas muestras no fueron analizadas, y el número de individuos que tenían muestras analizadas pero que no se incluyeron en el estudio por otros motivos (no se realizó la prueba de referencia, tuvo resultados inválidos o indeterminados, etc.) (**Figura 6**) (4).

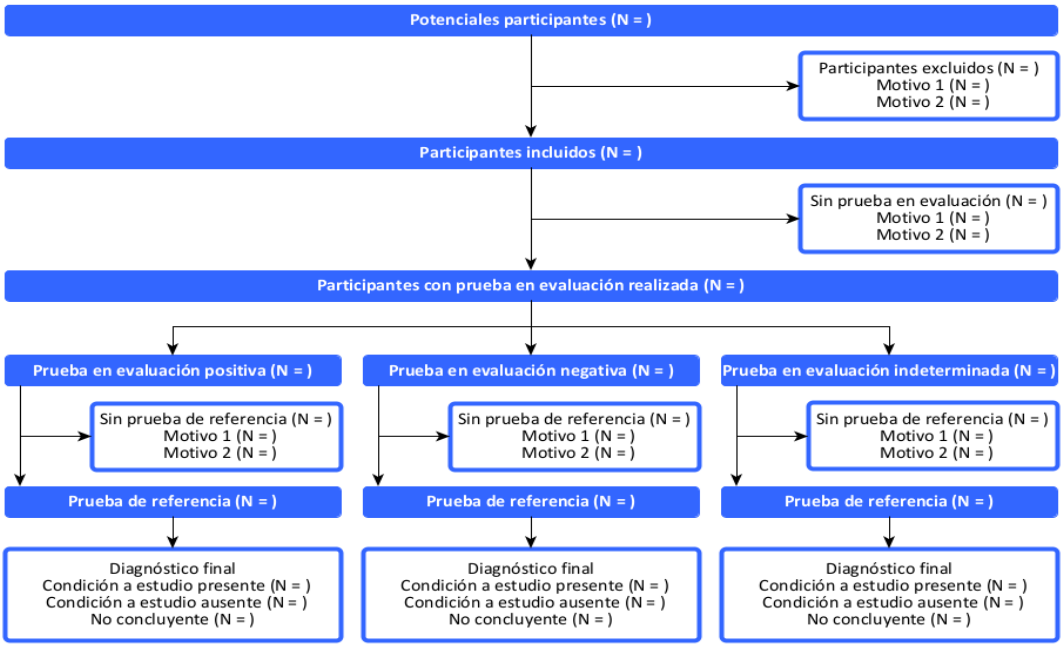
Tabla 10. Lista de elementos esenciales a incluir al comunicar los resultados de estudios de precisión diagnóstica.

Sección y tema	Elemento	Potenciales problemas
Título o resumen		
	Identificación como estudio de precisión diagnóstica utilizando al menos una medida de precisión (sensibilidad, especificidad, valores predictivos o AUC).	Resultados de precisión diagnóstica informados, pero no incluidos como objetivo del estudio.
Resumen		
	Resumen estructurado del diseño del estudio, métodos, resultados, y conclusiones.	El diseño del estudio no está claro o no se incluyen todos los apartados.
Introducción		
	Antecedentes científicos y clínicos, incluido el uso previsto, diana y qué lugar ocupa la prueba a estudio.	Falta de claridad sobre el uso previsto.
	Objetivos del estudio e hipótesis.	No indicar si el objetivo del estudio es establecer la validez científica, el rendimiento clínico o la precisión diagnóstica. No indicar si el rendimiento clínico es un objetivo de estudio.
Métodos		
Diseño del estudio	Indicar si la recopilación de datos fue antes de la realización de la prueba a estudio y el estándar de referencia (estudio prospectivo) o después (estudio retrospectivo).	No indicar cuándo se recopilaron los datos. Reclutar pacientes en estudios basados en los resultados de las pruebas.
Participantes	Criterios de inclusión.	No informar o registrar los síntomas u otras características utilizadas para incluir a los pacientes en el estudio. No informar el tiempo transcurrido entre la realización de las pruebas y aspectos clínicos clave.
	Dónde y cuándo se seleccionan los participantes.	No dejar claro qué departamentos del hospital estaban involucrados. Uso de muestras enviadas para pruebas de laboratorio de rutina, pero sin declarar su origen y fecha de envío.
	Los participantes formaron un grupo consecutivo, aleatorio o de conveniencia.	No definir un criterio claro o ir cambiando a lo largo del estudio.
Pruebas	Prueba en estudio, con suficiente detalle para permitir la replicación.	No indicar el tipo de muestra y/o las condiciones de recogida y conservación. No indicar quién obtuvo la muestra o quién realizó e interpretó la prueba. No indicar la versión.
	Prueba de referencia, con suficiente detalle para permitir la replicación.	El estándar de referencia a menudo se informa con detalles insuficientes para permitir la replicación, o a menudo usando métodos internos no publicados con resultados analíticos y clínicos poco claros.
	Justificación de la elección de la prueba de referencia (si existen alternativas).	A veces es difícil seleccionar la referencia, sobre todo si no existe una y hay que tener en cuenta diferentes criterios para diagnosticar la condición a estudio.
	Definición y justificación de los valores límite para determinar la positividad de las pruebas o las categorías de resultados, distinguiendo lo preespecificado de lo exploratorio.	No indicar si se han usado los puntos de corte preespecificados u otros. El umbral de positividad y cómo se determinó a menudo no está reportado.
	Indicar si la información clínica y los resultados del estándar de referencia estaban disponibles para quien realizó o interpretó la prueba a estudio.	Información disponible para los evaluadores de la prueba a estudio no reportada. No es posible determinar qué prueba se realizó primero (y por tanto si fue un estudio ciego).
Análisis	Métodos para estimar o comparar medidas de precisión diagnóstica.	Rara vez se explica el cálculo de la sensibilidad y la especificidad, así como la forma en que se se definen las categorías de aquellos con y sin la condición objetivo.
	Cómo se manejaron los resultados indeterminados de la prueba a estudio y de la prueba de referencia.	A menudo no se informa. Diagramas de flujo que demuestran que los resultados indeterminados no se incluyeron.
	Cómo se manejaron los resultados ausentes de la prueba a estudio y de la prueba de referencia.	Raramente reportado; los estudios a menudo sólo informan pruebas positivas y negativas, con resultados de pruebas fallidas o ausentes excluidas o no documentadas.
	Cualquier análisis de variabilidad en la precisión diagnóstica, distinguiendo lo preespecificado de lo exploratorio.	Normalmente no se incluye.
	Tamaño muestral y cómo se calculó.	Descripción limitada del cálculo de tamaño muestral.
Resultados		
Participantes	Flujo de participantes, incluyendo un diagrama.	Pocos estudios incluyen un diagrama con el flujo de participantes, tiempos, así como resultados indeterminados y faltantes.
	Características demográficas y clínicas de los participantes del estudio.	A veces no se reporta esta información o es muy limitada.

Tabla 10 continuación. Lista de elementos esenciales a incluir al comunicar los resultados de estudios de precisión diagnóstica.

	Distribución de la gravedad de la enfermedad en los participantes con la condición a estudio.	Rara vez se proporcionan definiciones y distribuciones de gravedad. La prevalencia de la condición a menudo no se informa.
	Distribución de diagnósticos alternativos en los participantes con la condición a estudio.	Los diagnósticos alternativos a veces pueden ser parte del estándar de referencia. Un diagnóstico alternativo no tiene por qué excluir la condición de estudio.
	Intervalo de tiempo e intervenciones clínicas entre la realización de la prueba de referencia y la prueba a estudio.	Por lo general, esto no es un problema, ya que la prueba en estudio y el estándar de referencia se realizan al mismo tiempo, por ejemplo, utilizando las muestras o muestras pareadas.
Pruebas	Tabla con los resultados cruzados de la prueba a estudio y los resultados del estándar de referencia.	Esta tabla a veces no se incluye, sino que directamente se muestran los valores de S, E, VPP, VPN, etc.
	Estimaciones de la exactitud diagnóstica y su precisión (intervalos de confianza del 95%).	A veces no se informan los intervalos de confianza.
	Cualquier evento adverso por realizar la prueba a estudio o el estándar de referencia. Dificultades en su realización.	A veces no se informan los problemas en la realización de las pruebas o derivados de ello.
Discusión		
	Limitaciones del estudio, incluidas las fuentes potenciales de sesgo, la incertidumbre estadística y la generalizabilidad.	Asumir que se obtendrán los mismos resultados observados en un laboratorio de referencia o en un entorno clínico con pacientes con una alta prevalencia de la condición a estudio en otros entornos clínicos.
	Implicaciones para la práctica, incluido el uso previsto y el papel clínico de la prueba a estudio.	Rara vez se explica el papel de la prueba a estudio, aunque a veces se puede deducir del diseño del estudio. Exagerar las implicaciones de los resultados en términos de importancia para la práctica o suponiendo generalizabilidad a otros entornos.
Otra información		
	Número y nombre de registro del estudio.	Raramente reportado. Estudios de rendimiento clínico a menudo no registrados previamente.
	Dónde se puede acceder al protocolo completo del estudio.	Raramente reportado.
	Fuentes de financiación y otro tipo de apoyo; papel de los financiadores.	
Adaptado de la guía STARD 2015 (21)		

Figura 6. Propuesta de diagrama de flujo de los participantes para un estudio de evaluación del rendimiento clínico de un PSDIV. (Modificado de la guía STARD 2015) (21).



La representación de los resultados del estudio en forma de tabla en la que se crucen los resultados de la prueba a estudio y los de la prueba de referencia es muy útil y facilita la lectura. Cualquier dato que falte o resultado indeterminado para cualquiera de las pruebas se debe informar y no excluir de los resultados. Todos los cálculos deben ir acompañados de los correspondientes intervalos de confianza, teniendo en cuenta el tamaño muestral y empleando los métodos estadísticos adecuados. En caso de resultados discordantes, debe indicarse cuál es el procedimiento para la resolución de estos casos.

El rendimiento clínico puede expresarse de varias formas, tal como se describe a continuación.

Resultados individuales

Se debe informar el número de participantes con resultados verdadero positivo, falso positivo, verdadero negativo y falso negativo, y no sólo las estimaciones de precisión derivadas, que se describirán a continuación. Estos resultados se pueden presentar en una tabla de 2x2 que muestre el número de participantes individuales cuyas pruebas de referencia y evaluada coincidieron y no coincidieron (**Tablas 11 y 12**).

Tabla 11. Tabla de contingencia en un estudio de rendimiento clínico en el que se compara un PSDIV con el estándar de referencia.

		Prueba evaluada	
		Positiva	**Negativa**
Estándar de referencia	**Positiva**	Verdadero positivo (VP)	Falso negativo (FN)
	Negativa	Falso positivo (FP)	Verdadero negativo (VN)

Tabla 12. Tabla de contingencia en un estudio de rendimiento clínico en el que se comparan dos PSDIV entre sí.

		PSDIV A	
		Positiva	**Negativa**
PSDIV B	**Positiva**	Concordantes	Discordantes
	Negativa	Discordantes	Concordantes

Sensibilidad (S) y especificidad (E) diagnósticas

La sensibilidad es el porcentaje de individuos que presentan la condición de interés y que obtienen un resultado de la prueba positivo. Representa la capacidad de la prueba para identificar correctamente a un individuo con dicha condición. Por su parte, la especificidad es el porcentaje de individuos que no presentan la condición de interés y que obtienen un resultado de la prueba negativo. En este caso representa la capacidad de la prueba para identificar correctamente a un individuo que no presenta dicha condición. Son características inherentes a la prueba. Para el cálculo de la S y E, se emplean las siguientes fórmulas:

S = VP/(VP+FN)	E = VN/(FP+VN)

***Odds ratio* diagnóstica (ORD)**

La *odds ratio* diagnóstica (ORD) es la razón de la *odds* de que el enfermo dé positivo (proporción entre verdaderos positivos y falsos negativos: VP/FN) con respecto a la *odds* de dar positivo estando sano (co-

ciente entre falsos positivos (FP) y verdaderos negativos (VN): FP/VN). Depende de la S y E, y, por tanto, no depende de la prevalencia de la enfermedad. Para el cálculo de la ORD, se emplea la siguiente fórmula:

$$ORD = (VP/FN)/(FP+VN)$$

El valor nulo es el uno, indicando que la prueba no tiene capacidad discriminatoria entre sanos y enfermos. Un valor mayor de uno indica capacidad discriminatoria, que será mayor cuanto mayor sea el valor. Por último, valores entre cero y uno indican que la prueba no solo no discrimina bien entre enfermos y sanos, sino que los clasifica de forma errónea y nos da más valores negativos entre los enfermos que entre los sanos.

Valor predictivo positivo (VPP) y negativo (VPN)

El VPP indica el porcentaje de individuos con una prueba positiva que presentan la condición de interés (fiabilidad de un resultado positivo), mientras que el VPN indica el porcentaje de individuos con prueba negativa que no presentan dicha condición (fiabilidad de un resultado negativo). Para el cálculo de los VPP y VPN, se emplean las siguientes fórmulas:

$$VPP = VP/(VP+FP) \qquad VPN = VN/(FN+VN)$$

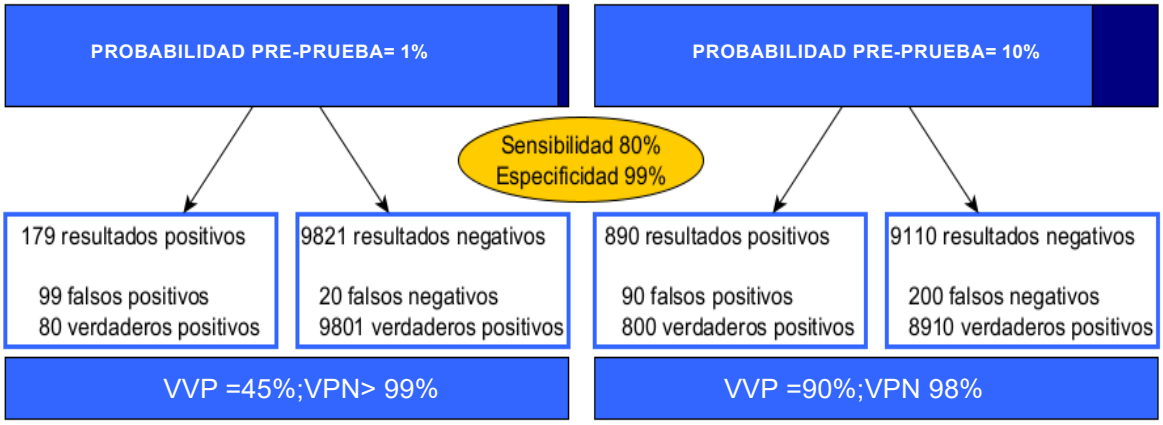
Como ejemplo, en la siguiente imagen (**Figura 7**) se muestran los resultados individuales, así como la sensibilidad, especificidad, VPP y VPP, con sus correspondientes intervalos de confianza, de cuatro pruebas para el diagnóstico rápido del VIH en campo (22).

Figura 7. Resumen de sensibilidades, especificidades, valores predictivos positivos y valores predictivos negativos de las pruebas rápidas de VIH en comparación con las del algoritmo de referencia (EIA más Multispot). TP: *true positive*; FP, *false positive*; TN, *true negative*; FN, *false negative*. (*Modificado de Juarez y cols.*2016) (22).

		Result (*n*)[b]						Predictive value (%)[c]	
Test	Total[a]	TP	FP	TN	FN	Sensitivity, (% [95% CI])	Specificity (% [95% CI])	Positive	Negative
Determine	1,503	465	0	1,037	1	99.8 (98.81–99.99)	100.0 (99.64–100)	100.0	99.9
Uni-Gold	1,508	467	0	1,039	1	99.8 (98.47–99.95)	100.0 (99.65–100)	100.0	99.8
SD Bioline	1,509	468	0	1,040	1	99.8 (98.82–99.99)	100.0 (99.65–100)	100.0	99.9
INSTI	1,505	467	2	1,035	1	99.8 (98.82–99.99)	99.8 (98.30–99.98)	99.6	99.9

Los VPP y VPN varían según la prevalencia de la condición en la población a estudio (probabilidad preprueba). La prevalencia de una condición se considera como la probabilidad del individuo de presentar esa condición en su entorno y contexto clínico particular. Se recomienda realizar un cálculo de estos valores teniendo en cuenta diferentes prevalencias de la condición a estudio en la población (**Figura 8**). En la siguiente web (https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/eua-authorized-serology-test-performance) la FAD recoge los valores predictivos positivo y negativo de diferentes pruebas serológicas de COVID-19 según la prevalencia en población objetivo, ya que esta ha ido cambiando y cambia continuamente. Volviendo al ejemplo anterior, si los test rápidos se fueran a usar en un entorno con una prevalencia más elevada o, por el contrario, más baja de VIH, sería necesario realizar una nueva evaluación.

Figura 8. Valor predictivo positivo y negativo de un PSDIV según la probabilidad pre-prueba. Elaborado con yEd 3.21. Modificado de Doust y cols. (2021) (23).



Efectividad diagnóstica (índice de exactitud, ***diagnostic accuracy***)

Otra medida global del rendimiento clínico es la efectividad diagnóstica, que se expresa como la proporción de sujetos clasificados correctamente (VP+VN) entre el total de sujetos (VP+VN+FP+FN). Está condicionada por la prevalencia. Para una misma S y E, la efectividad diagnóstica será mayor cuanto menor sea la prevalencia.

Índice de Youden

El índice de Youden se puede emplear para evaluar el poder discriminativo de un PSDIV y también para comparar varios PSDIV. Para el cálculo del índice de Youden, se emplea la siguiente fórmula:

$$\text{Índice de Youden} = (S + E) - 1$$

Coeficiente kappa de Cohen

El coeficiente kappa es una medida del acuerdo entre dos PSDIV que clasifican cada uno de los elementos en categorías mutuamente excluyentes, empleando una matriz 2x2. Incorpora en su fórmula una corrección que excluye la concordancia debida exclusivamente al azar. Para la evaluación de concordancia de tres o más categorías (por ejemplo, datos asociados a positivo, negativo e indeterminado) se emplea el coeficiente kappa de Fleiss, que emplea matrices nxm. El coeficiente kappa se interpreta de la siguiente manera (ver página 9 de este documento; https://www.graphpad.com/quickcalcs/kappa1/):

índice kappa	Concordancia
<0,20	Pobre
0,21-0,40	Débil
0,41-0,60	Moderada
0,61-0,80	Buena
0,81-1	Muy buena

Cociente o razón de verosimilitud (*likelihood ratio*, LR)

El cociente de verosimilitud es la razón entre la posibilidad de observar un resultado en los pacientes con la condición a estudio entra la posibilidad de ese resultado en pacientes sin la condición. Dicho de otra manera, es la probabilidad de tener la enfermedad en oposición a no tenerla, teniendo un resultado del test positivo y la probabilidad de no tener la enfermedad en oposición a tenerla, teniendo un resultado del test negativo. Existen cocientes de probabilidad para test con resultado positivo (cociente de probabilidad positivo, CPP; o *likelihood ratio* positivo, LR+) y negativo (cociente de probabilidad negativo, CPN; o l*ikelihood ratio* negativo, LR-).

		CPP	CPN
Cálculo		**S/(1-E)**	**(1- S)/E**
Interpretación	Excelente	>10	<0,1
	Buena	5 - 10	0,1 - 0,2
	Regular	5 - 2	0,2 - 0,5
	Inútil	<2	0,5 - 1

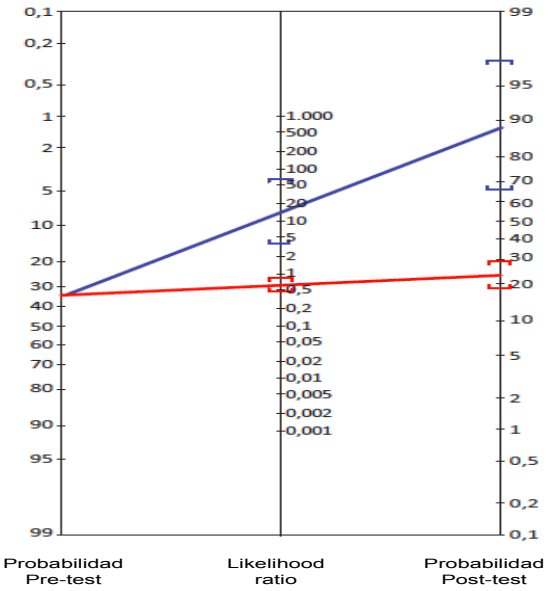
Es una herramienta de gran utilidad para la toma de decisiones clínicas frente a la solicitud de algún test diagnóstico, ya que son valores inherentes a este e independientes de la prevalencia de la enfermedad, al igual que ocurre con la S y la E. Se puede usar para analizar pruebas con resultados dicotómicos, en los que solo es posible determinar presencia o ausencia de enfermedad (negativo o positivo), o bien con resultados categóricos, por ejemplo, mediante exámenes que tienen categorías de gravedad (leve, moderada o severa).

Cuando se solicita una prueba de laboratorio, la probabilidad de confirmar el diagnóstico puede ser previa, antes de solicitar la prueba, o posterior, después de recibir el resultado. Una prueba para resultar útil deberá tener excelentes valores de S y E. Sin embargo, el grado de confianza que genera la prueba nunca es total porque siempre hay un grado de incertidumbre. La utilidad de una prueba viene determinada por la disminución en el grado de incertidumbre presente antes y después de que esta sea realizada.

El nomograma de Fagan permite calcular la probabilidad post-prueba una vez que se conoce la probabilidad pre-prueba (prevalencia) y la razón de verosimilitud. Tiene tres columnas: la primera es la probabilidad de tener la enfermedad antes de aplicar la prueba (prevalencia), la segunda es la razón de verosimilitud y la tercera la probabilidad post-prueba. Con una regla se traza una línea entre la probabilidad pre-prueba y la razón de verosimilitud. La prolongación de esta línea corta en la tercera columna la probabilidad de tener la enfermedad en función del resultado de la prueba (**figura 9**) (24, 25).

En el trabajo de Abarca y cols. (2013) se determina la utilidad diagnóstica de la IgM anti-*Bartonella henselae* para el diagnóstico de la enfermedad por arañazo de gato, utilizando como estándar de referencia (recomendado por el CDC) la determinación de IgG mediante inmunofluorescencia. En este trabajo no sólo calcularon la S, E, VPP y VPN, sino que también calcularon los LR y la probabilidad post-test mediante el nomograma de Fagan. En 37 pacientes con IgG anti-*B. henselae* positiva, la IgM fue positiva en 16 y negativa en 21; en 71 pacientes con IgG negativa, la IgM fue negativa en 69 y positiva en 2. Por consiguiente, la IgM presentó S 43%, E 97%, VPP 88%, VPN 77%, LR(+) 15 y LR(–) 0,58. Por todo ello, concluyen que una IgM positiva apoya el diagnóstico de enfermedad por arañazo de gato, pero una negativa no permite descartarlo. Por tanto, la IgG debe seguir considerándose como el estándar de oro para el diagnóstico de infecciones por *B. henselae* (26).

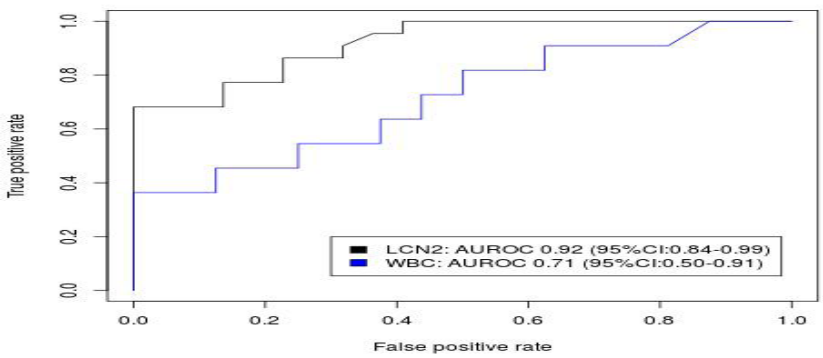
Figura 9. Nomograma de Fagan utilizado para el cálculo de la probabilidad post-test con el LR+ (azul) y LR- (rojo). La prevalencia en la muestra era de 34,3%, punto que se marca como probabilidad pre-test en el nomograma y desde el cual se traza una línea (azul) que cruza el LR+ obtenido (15) con lo que se obtiene una probabilidad post-test de 89% y una línea (roja) que cruza el LR– obtenido (0,58) con lo que se obtiene una probabilidad post-test de 23% (IC95% 19-29). (*Modificado de Abarca y cols*. 2013) (26).



AUC (Área bajo la curva, *area under a curve*) **y curvas ROC (***Receiver Operating Characteristic***)**

La curva ROC se utiliza de forma general para estudiar la capacidad de una prueba para discriminar a los individuos que presentan la condición (verdaderos positivos) de aquellas que se considera que presentan la condición de forma errónea (falsos positivos). La curva ROC se utiliza también para determinar el mejor punto de corte de una prueba (mayor cantidad de verdaderos positivos con menor cantidad de falsos positivos), y finalmente para comparar la eficiencia de dos o más pruebas, a través de sus áreas bajo la curva (AUC, del inglés *Area Under a Curve*). (Ver apartado 3.2.8.).
En un trabajo en el que se evalúa rendimiento diagnóstico de la lipocalina-2 en líquido sinovial para discriminar infección protésica de fracaso aséptico del implante encuentran que el punto de corte óptimo para una máxima sensibilidad (86,3%) y especificidad (77,2%) para discriminar fallo aséptico vs infección fue de 152 ng/mL, lo que produjo un AUC de 0,92 (IC 95%: 0,84-0,99). El rendimiento diagnóstico de la lipocalina-2 fue significativamente superior al del recuento de leucocitos (p=0,011) (**Figura 10**) (27).

Figura 10. AUC para todas las concentraciones de lipocalina-2 en comparación con AUC para el recuento de leucocitos en líquido sinovial para discriminar entre infección y fallo aséptico. (*Con permiso de los autores: Vergara A y cols. 2019*) (27).



Evaluación del rendimiento clínico de un PSDIV sin estándar de referencia

En la práctica, puede ocurrir que no se aplique el estándar de referencia a todos los participantes del estudio, bien porque es muy caro, o invasivo, o los pacientes no dan su consentimiento o los médicos decidieron no realizarla a algunos pacientes por razones médicas. También puede ocurrir que el estándar de referencia no sea muy preciso o directamente que no exista un estándar de referencia. En estos casos, no se podrán calcular los indicadores mencionados hasta ahora, sino que habrá que recurrir a otros métodos. En la revisión sistemática de Umemneku Chikere y cols (2019), se recogen diferentes métodos para evaluar el rendimiento clínico de una prueba en estos supuestos (28).

Otros resultados

Se pueden obtener otros datos que pueden ser útiles para determinar qué impacto puede tener el resultado de la prueba en la atención clínica y cómo de viable y sencillo puede resultar su implementación en la rutina, como por ejemplo:

- **Tiempo** desde que se toma una muestra hasta que se obtiene el resultado de la prueba. En el trabajo de Munrós y cols. (2021) se compararon dos técnicas (PCR estándar y PCR rápida) para la detección de *Chlamydia trachomatis* y *Neisseria gonorrhoeae* en mujeres con sospecha de enfermedad inflamatoria pélvica. La concordancia entre las dos técnicas fue del 100%, pero los resultados con la PCR rápida se obtuvieron mucho antes (2 vs. 24 horas) y su uso fue mucho más sencillo, lo que apoya su uso en centros donde la técnica estándar no está disponible (29).
- **Dificultad** en la interpretación de los resultados.
- **Tasa de resultados inválidos**. Debe informarse el número de resultados indeterminados e inválidos (cuando la prueba no devuelve ningún resultado). ¿Hay un número excesivo de pruebas no válidas? ¿A qué se debe? ¿Prueba, operador o muestras?
- **Impacto del resultado de la prueba en la toma de decisiones clínicas** (Apartado 3.4.)

3.4. FASE 4: EFECTOS DE LOS RESULTADOS EN EL PACIENTE O SISTEMA

Una vez establecidas las características analíticas del dispositivo en evaluación resta estudiar las implicaciones del resultado del test tanto en el paciente como en la sociedad. Por tanto, la última etapa en el proceso de evaluación consiste en investigar la medida en que los resultados del test modifican el razonamiento diagnóstico o la toma de decisiones asistenciales y en última instancia su impacto en la situación clínica del paciente. El diseño adecuado para alcanzar una relación de causalidad que permita evaluar adecuadamente el efecto de un test en la salud o calidad de vida del paciente es el **ensayo clínico controlado aleatorizado** (RCT, del inglés *randomized controlled trial)*. Otra cuestión es si efectivamente un test diagnóstico debe considerarse médicamente útil aun cuando su resultado no tiene un efecto clínicamente evaluable o no produce ningún cambio en la actitud clínica del médico responsable, por ejemplo, un test destinado al diagnóstico de una patología no tratable o no prevenible (5).

Una última etapa en el proceso de evaluación de los resultados en el paciente o el sistema supondría la **evaluación económica de la implantación del dispositivo**. Las evaluaciones económicas se pueden llevar a término atendiendo a cómo se computan los gastos de una intervención determinada y sus consecuencias, ya sea en variables monetarias, variables clínicas, o variables de utilidad como los años de vida ganados ajustados por calidad. Atendiendo a esto se clasifican en estudios de minimización de costes, coste eficacia, coste utilidad o coste beneficio. El procedimiento de la SEIMC número 64 ya revisó profundamente todos estos aspectos por lo que no nos extenderemos más aquí (3).

Los ensayos clínicos de manera genérica se pueden clasificar en dos tipos. El ensayo clínico propiamente dicho o estudios de intervención, en los que nos centraremos a continuación, y los estudios observacionales. Estos últimos se caracterizan porque los pacientes no son asignados a una intervención específica por parte del investigador, aunque sí pueden recibir intervenciones (medicamentos, test diagnósticos, etc.) de acuerdo con las prácticas clínicas asistenciales. Los estudios observacionales se clasifican en

estudios transversales (*cross-sectional studies*) y longitudinales:
• Estudios transversales: aquellos que comparan diferentes grupos de población o muestras en un único momento temporal.
  ◦ Permiten el estudio de diferentes variables en un momento determinado de tiempo.
  ◦ No tienen en consideración todo aquello que pasó antes o lo que pasará después.
• Estudios longitudinales:
  ◦ En estos se recopilan datos de los individuos durante un periodo prolongado de tiempo.
  ◦ Permiten establecer la secuencia temporal de los diferentes eventos ocurridos en los individuos, lo que permite establecer relaciones causa-efecto.
  ◦ En contrapartida permiten el estudio de una única variable y son más costosos.

3.4.1 Ensayo clínico controlado y aleatorizado

De acuerdo con el *National Institutes of Health* (NIH) un ensayo clínico es un estudio de investigación en el cual uno o más humanos son asignados a una o más intervenciones para evaluar el efecto de éstas en los resultados médicos o conductuales de los individuos. El RCT (*randomized clinical trial*) es el *gold standard* para demostrar asociaciones causa-efecto en intervenciones médicas, farmacológicas o diagnósticas. La asignación aleatoria de los individuos a uno u otro brazo del ensayo, preferiblemente en una asignación ciega para el clínico y el paciente, aporta la base para la cuantificación del efecto de la intervención con una interpretación causal y es la base de la medicina basada en la evidencia (30).

Los ensayos clínicos se pueden clasificar en cinco grandes tipos con base en su finalidad:
• I. Ensayos de prevención que investigan cómo prevenir una determinada patología o su recurrencia.
• II- Ensayos de *screening* destinados a detectar ciertas enfermedades o su recurrencia.
• III- Ensayos de diagnóstico, los cuales examinan pruebas o procedimientos para el diagnóstico de enfermedades concretas.
• IV- Ensayos de tratamiento que comparan tratamientos en evaluación con aquellos establecidos. Incluyen tanto fármacos como dispositivos o procedimientos.
• V- Ensayos de calidad de vida, los cuales investigan terapias o intervenciones para mejorar el confort o la calidad de vida de pacientes con patologías crónicas.

Otra posible clasificación de los ensayos clínicos es como fijos o adaptativos dependiendo de si el diseño del estudio permite cambios a medida que se van acumulando datos del ensayo. (https://www.nih.gov/health-information/nih-clinical-research-trials-you/basics).

3.4.2. Consideraciones generales para ensayos clínicos de PSDIV

Por su idiosincrasia en los ensayos clínicos que evalúan PSDIV hay diferentes aspectos a resaltar, como son: su intención de uso, el diseño del ensayo, así como el comparador utilizado o test de referencia.

Intención de uso: atiende a diferentes aspectos tales como el propósito clínico del test, el tipo de prueba, el criterio evaluado, el tipo de muestra, lugar anatómico de toma de muestra y procedimiento para su obtención o la población diana. Todos estos son aspectos a tener en consideración para diseñar el experimento, de manera que el ensayo clínico se adecue tanto como sea posible a la práctica real del test evaluado. (Ver apartado 3.3.1.).

Diseño del estudio: de manera general los RCT se pueden clasificar atendiendo al método elegido para la inclusión de los pacientes (aleatoria o secuencial), al número de grupos en el ensayo y el método de comparación entre grupos, (diseños paralelos o pareados) y al tipo de comparador.

• **Método de inclusión de individuos**:
Un ensayo clínico prospectivo debe incluir pacientes, atendiendo a unos criterios de inclusión y exclusión preestablecidos. La inclusión de pacientes al ensayo puede atender a dos modelos:

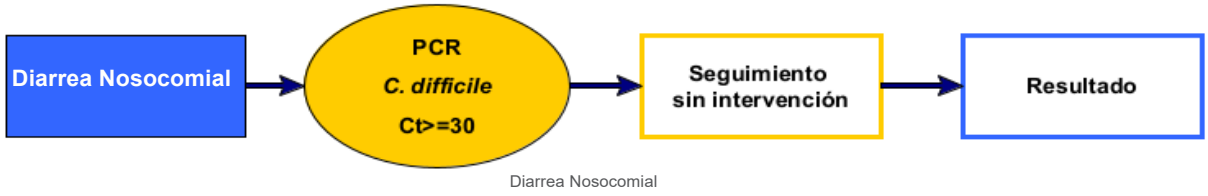
- **Pacientes consecutivos**: se incluye a todos los pacientes que cumplan los criterios de inclusión a medida que se presentan hasta alcanzar el tamaño muestral requerido. Este suele ser el método de elección para la evaluación de test de cribado con una elevada sensibilidad.
- **Pacientes aleatorizados**: mediante este método se incluyen de manera aleatoria a diferentes pacientes siempre y cuando cumplan estos criterios, ya sea de manera global o por clústeres de población, siguiendo una distribución homogénea en cada clúster, proporcional al tamaño de cada uno de ellos en la población general o mixta, con una proporción fija en todos ellos y una variable dependiendo del grado de representación de cada grupo en la población a estudio. Este último método evita fenómenos de sobre- o infrarrepresentación de clústeres poco frecuentes.

- **Diseño de los brazos de análisis:**

Un RCT puede evaluar un único test, sin el uso de un comprador de referencia, para evaluar una característica funcional del test, por ejemplo, su capacidad pronóstica, con entre otros, uno de los siguientes modelos (31):

- **Modelo general para evaluar el valor pronóstico de una técnica analítica** (**Figura 11**): Particularmente útil cuando además de evaluar la concordancia entre el resultado del test y su valor real es necesario conocer su asociación con la gravedad o el pronóstico de una patología determinada, como por ejemplo el modelo que se expone a continuación:

Figura 11. Modelo general para evaluar el valor pronóstico de una técnica analítica. Elaborado con yEd 3.21.

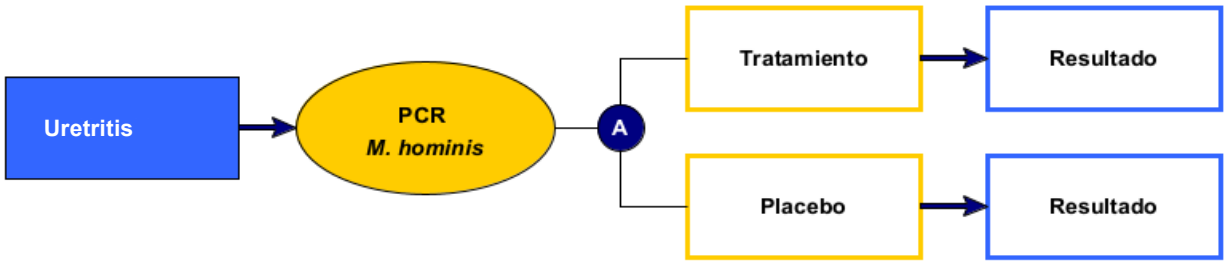


Otros ejemplos bibliográficos: Lewis y cols. (1997) (32).

- **Modelo simple con el procedimiento analítico al inicio del ensayo clínico (Figura 12):**

- Establece el valor pronóstico del test en cada brazo
- Evalúa qué opción terapéutica es más efectiva en los dos brazos de tratamiento, por tanto, en todos los pacientes incluidos en el ensayo.
- Permite evaluar las opciones terapéuticas en pacientes con idéntico resultado en el test y la capacidad del test para predecir el tipo de respuesta a los diferentes brazos de tratamiento.

Figura 12. Modelo simple con el procedimiento analítico al inicio del ensayo clínico. Elaborado con yEd 3.21.

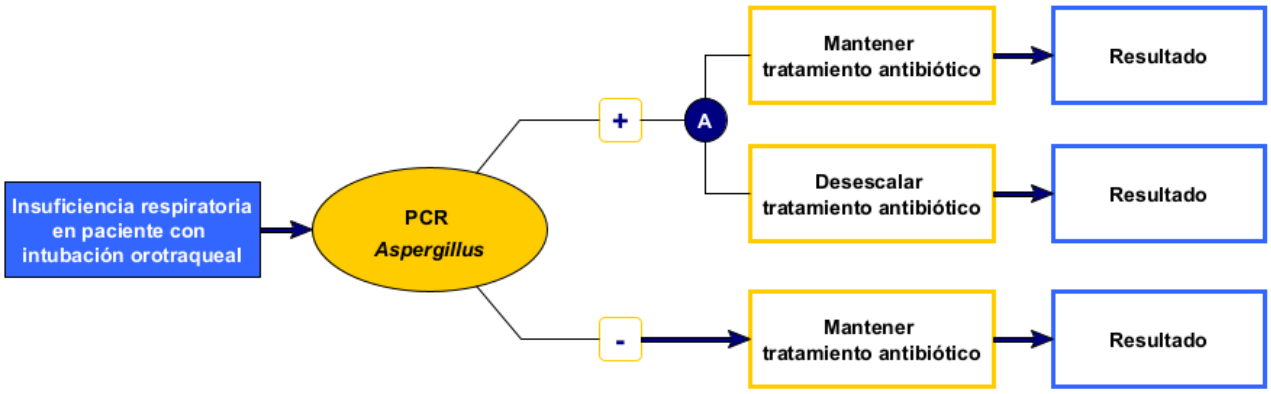


Otros ejemplos bibliográficos: Sargent y cols (2005) (33).

• **Modelo en el cual el resultado del test se utiliza como criterio de inclusión al ensayo**:

- Únicamente los pacientes con resultado positivo se incluyen en el grupo de intervención.
- Permite evaluar el efecto de tu intervención en pacientes con un test positivo.
- Mejora aspectos éticos respecto al modelo anterior.

Figura 13. Modelo en el que el resultado del test se utiliza como criterio de inclusión al ensayo. Elaborado con yEd 3.21.

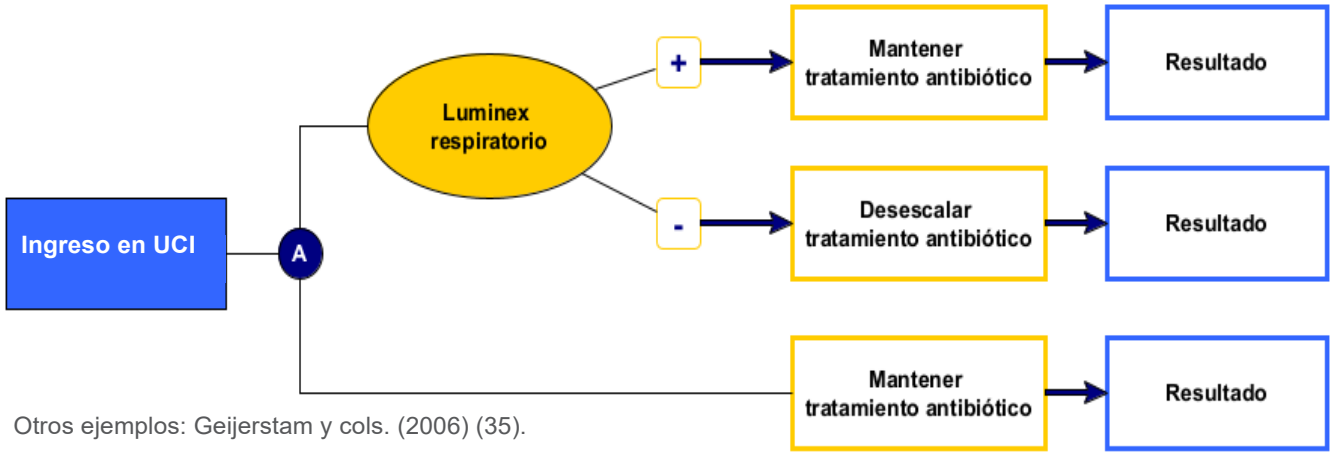


Otros ejemplos bibliográficos: Côté y cols. (1995) (34).

• **Modelo en el que se aleatoriza a los pacientes a realizarse o no el test (Figura 14)**:

- En este tipo de ensayo se aleatoriza a los pacientes a la realización o no del test.
- El modelo evalúa tanto el efecto del test como del tratamiento, pero no se puede diferenciar uno de otro.

Figura 14. Modelo en el que se aleatoriza a los pacientes a realizarse o no el test



Otros ejemplos: Geijerstam y cols. (2006) (35).

Cuando se compara el rendimiento de dos métodos entre sí, existen dos diseños generales: El diseño paralelo y el de grupos pareados (19).

• **Modelo de grupos paralelos**: este modelo se caracteriza por la asignación aleatoria de cada paciente a uno de los dos grupos de análisis, pero no a ambos. Este método es útil para evaluar test invasivos o cuando no sea factible realizar más de un análisis por individuo. En contrapartida requiere de tamaños muestrales grandes para que ambos grupos, si la distribución es aleatoria, estén balanceados en todo el espectro de posibilidades de la población diana.

• **Modelo de grupos pareados**: por este método, cada individuo es evaluado por ambas pruebas, tanto con el método en evaluación cuanto con el test de referencia. Estos diseños son estadísticamente más

eficientes debido a que se evita la variabilidad adicional que aporta el hecho de tener dos grupos diferenciados propio del diseño paralelo. Permite además evaluar la influencia de características individuales de los individuos con uno y otro test. Es el modelo más habitual

Comparación secuencial o simultánea:
Si se evalúa más de un test, la comparación de éstos con su estándar puede ser secuencial, seguida una de otra o simultánea; esta última vertiente posibilita la comparación directa de ambos test de estudio con su referencia en una misma población y resulta más económica y más rápida. Por el contrario, puede dificultar la interpretación independiente de los resultados, dificultar el diseño del estudio, ya sea por la toma de muestra, la aleatorización o el enmascaramiento. Además, puede resultar éticamente cuestionable la toma de varias muestras a un mismo paciente.

Test de referencia:
La evaluación del rendimiento analítico de un test diagnóstico se basa en la comparación con un test de referencia. La elección de un test de referencia adecuado es crucial para la legitimidad de la prueba. (Ver apartado 3.3.1.)

3.4.3. Barreras específicas asociadas a los ensayos clínicos de PSDIV

Existen diferentes aspectos que pueden influir negativamente en la consecución de ensayos clínicos para test diagnósticos debido a las características específicas de este tipo de evaluaciones. A destacar: el momento de la evaluación del test, la aceptabilidad, el enmascaramiento, el grupo comparador, la curva de aprendizaje, entre otros sesgos (36).

- **Momento de la evaluación**: la planificación y desarrollo de un ensayo clínico a menudo abarca varios meses o incluso años. Sin embargo, los dispositivos biomédicos, a menudo sufren cambios de diseño durante los primeros meses tras la introducción en el mercado, lo cual puede mermar la utilidad de los resultados del ensayo.

- **Aceptabilidad:** dificulta la aleatorización de los individuos y por tanto la validez del ensayo. Este problema deriva de las preferencias de los pacientes por elegir un brazo concreto del ensayo (una técnica o proceso diagnóstico en particular respecto al otro).

- **Enmascaramiento**: es un proceso fundamental en los ensayos clínicos; atendiendo a las características de este tipo de ensayos en particular, el enmascaramiento puede ser difícil o imposible, tanto por razones prácticas como por razones éticas.

- **Elección del grupo comparador**: el grupo control es un aspecto fundamental en el ensayo clínico. A menudo, no existe un gold standard con el que comparar o éste no es factible, por las complicaciones técnicas o por tratarse de técnicas alejadas de la práctica asistencial. También existe la posibilidad de que existan varios métodos de comparación reales en la práctica clínica, lo cual hace necesaria la elección del ensayo a utilizar en cada grupo y su justificación.

- **Curva de aprendizaje**: siempre y cuando la experiencia del operador o el facultativo puedan afectar al resultado de la técnica.

- Otros sesgos:
  - **Sesgo temporal**: el momento en el que se realiza la prueba puede ser un factor de confusión, dando a entender que los que se realizan la prueba de manera temprana tienen por ejemplo una mayor supervivencia que aquellos que se realizan la prueba de manera tardía por este sesgo temporal aun siendo la supervivencia independiente del resultado del test.

○ **Sesgo de verificación**: cuando el test de referencia únicamente se aplica al subgrupo de pacientes con resultados positivos en el test en investigación se puede producir este sesgo de verificación.
○ **Sesgo de progresión/regresión de la enfermedad**: cuando los resultados del test en evaluación y el de referencia no se evalúan simultáneamente puede ocurrir que el estado clínico del paciente no sea el mismo y por tanto se produzca este tipo de sesgo.

3.4.4. Factores que pueden afectar a la validez externa de un ensayo clínico

Entre los diferentes factores que pueden restar validez externa a un ensayo clínico de PSDIV destacan los resumidos en la t**abla 13**.

Tabla 13. Factores que pueden afectar a la validez externa de un ensayo clínico

Entorno	Centro de salud, entorno hospitalario, etc.
	Selección de clínicos
	Selección de hospitales
Selección de pacientes	Criterios de inclusión y exclusión
	Métodos de diagnóstico pre-aleatorización
	Ratio
	Proporción de pacientes que declinan el ensayo
Características de aleatorización	Características basales
	Grupos raciales
	Comorbilidades
Diferencias entre el ensayo y la práctica clínica	Tiempos de intervención y resultados
	Exclusión de ciertas intervenciones en el grupo control
Resultados medidos y seguimiento	Frecuencia de seguimiento
	Duración del seguimiento
	Relevancia clínica de los marcadores subrogados elegidos
(Modificado de Rothwell y cols. 2006) (37)	

3.4.5. Control de calidad y monitorización

Los ensayos deben incorporar un **control de calidad** que atienda a todo el proceso analítico y documental de acuerdo con las buenas prácticas de laboratorio clínico (GCLP) y supervisado por una persona externa al equipo de investigación. Las GCLP son un estándar internacional de calidad que surge de la unión de las GCP (buenas prácticas clínicas) que aborda el diseño, metodología y reporte de datos en los ensayos clínicos y los GLP (buenas prácticas de laboratorio) que ofrece un sistema de calidad para los procesos organizativos y las condiciones en las que se planifican, realizan, supervisan, registran, archivan y notifican los estudios de laboratorio no clínicos. Su objetivo es proporcionar un marco unificado para el análisis de muestras con el fin de dar credibilidad a los datos generados y facilitar la aceptación de los datos por las autoridades reguladoras de todo el mundo.

Estos controles de calidad comprenden tres aspectos fundamentales:

- **Control de calidad del estudio**: fundamentalmente se refiere a la generación de procedimientos normalizados de trabajo que recogen detalladamente todos los aspectos del estudio a desarrollar.
- **Monitor externo**: control independiente de calidad llevado a cabo por un evaluador externo.
- **Mejora de calidad del estudio**: pautas para la identificación de puntos críticos, errores y procesos para su corrección.

Para una información más detallada recomendamos consultar el documento de la OMS sobre GCLP (Ver apartado 2 de este procedimiento).

3.4.6. Aspectos fundamentales a considerar para el desarrollo de un ensayo clínico para la evaluación de un PSDIV.

A modo de resumen y de manera esquemática describimos a continuación los aspectos fundamentales a considerar para el desarrollo de un ensayo clínico para la evaluación de dispositivos de diagnóstico in vitro (38):

1. Definir la población a estudio para el test en evaluación
2. Desarrollo metodológico del reclutamiento de pacientes:
    a. Quién reclutará a los individuos
    b. Criterios de inclusión y exclusión
    c. Quién responderá a los pacientes las cuestiones relacionadas con el estudio
    d. Cómo se pedirá el consentimiento informado
3. Diseño de los medios para la consecución del estudio
    a. Selección de datos necesarios para el estudio
    b. Características y tipo de cuestionarios (sencillez, preguntas estructuradas, respuestas predefinidas, lenguaje asequible, diferenciación entre respuestas omitidas y no aplicables)
4. Desarrollo de un estudio piloto
    a. Permitirá evaluar lo medios de recogida de datos
    b. Evaluación de la prevalencia real en la población estudiada para el cálculo del tamaño muestra
5. Cálculo del tamaño muestral
6. Desarrollo de un plan logístico (obtención, transporte, numeración y conservación de las muestras)
7. Definir el método de enmascaramiento del resultado del test de referencia al analizar la muestra mediante el método en evaluación
8. Definir un plan de control de calidad
    a. Para garantizar la competencia en ambos test, el de referencia y el test en evaluación
    b. Garantizar la calidad en la recogida y almacenamiento de datos
9. Diseño del modelo de recogida almacenamiento y análisis estadístico de los datos
10. Método para garantizar la confidencialidad de los datos
11. Definir un plan para asegurar que el estudio reúne los requisitos de calidad necesarios, así como un método para la validación externa del mismo
12. Requisitos éticos y validación por el comité de ética pertinente
13. Definir los métodos para la difusión de resultados.

4. ESTUDIOS DE EVALUACIÓN DE PSDIV

En este apartado se revisan estudios publicados sobre evaluación del rendimiento analítico y clínico de diferentes tipos de PSDIV. Se toman como ejemplo: i) **estudios de evaluación de test rápidos de detección de antígenos**, ii) **estudios de evaluación de técnicas de LAMP *(Loop-mediated Isothermal Amplification),*** y iii) **estudios de evaluación de métodos de determinación de sensibilidad antibiótica**. El propósito de este apartado, es proporcionar por medio de ejemplos de trabajos previos, ideas de diseño de estudios de evaluación de PSDIV. Los ejemplos se han agrupado siguiendo la propuesta de 4 fases diferenciadas en la evaluación de un test diagnóstico.

4.1. ESTUDIOS DE EVALUACIÓN DE TEST RÁPIDOS DE DETECCIÓN DE ANTÍGENOS

Tabla 14. Ejemplos de estudios fase 2 de evaluación de test rápidos de detección de antígenos.

Objetivo	Estándar de referencia	Diseño estudio	Muestras incluidas	Parámetro analizado	Resultado	Ref.
Rendimiento analítico	NA	De sensibilidad analítica	Se generó una muestra artificial mediante un virus inactivado a una concentración de 4.6 x 10^5 $TCID_{50}$/ml, se hicieron diluciones seriadas y se analizó por triplicado.	LOD	Límite de detección: 1.15 x 102 TCD50/ml	(36)
	RT-PCR	Ensayo prospectivo multicéntrico que analizó muestras nasofaríngeas tanto de individuos sintomáticos como de asintomáticos	865 muestras nasofaríngeas.	Sensibilidad Especificidad Exactitud	Sensibilidad relativa: 98.32% (IC95% 94,06- 99,81) Especificidad relativa: 99.60% (IC95% 98,8-99,92) Exactitud: 99.42% (IC95% 98,66- 99,81)	(36)
	RT-PCR	Estudio retrospectivo de reactividad cruzada con distintos organismos Las muestras positivas por PCR-RT para los diferentes organismos resultaron negativas cuando se analizaron con la prueba rápida de antígenos.	32 microorganismos	Especificidad analítica y reactividad cruzada	RSV Tipo A: $5{,}5\times10^7$ PFU/mL RSV Tipo B: $2{,}8\times10^5$ TCID50/mL Influenza A H1N1: 1×10^6 PFU/mL Influenza A H3N2: 1×10^6 PFU/mL Influenza A H5N1: 1×10^6 PFU/mL Rhinovirus: 1×10^6 PFU/mL Adenovirus: $5\times10^{7,5}$ TCID50/mL Adenovirus: $2{,}8\times10^6$ TCID50/mL	(36)

Tabla 15. Ejemplos de estudios fase 3 de evaluación de test rápidos de detección de antígenos.

Objetivo	Estándar de referencia	Diseño estudio	Muestras incluidas	Parámetro analizado	Resultado	Ref.
Inmunocromatografía de malaria en países no endémicos	RT-PCR	Ensayo prospectivo monocéntrico con enmascaramiento para el resultado de la PCR.	256 individuos que retornaban de países endémicos con sospecha de malaria	VPP, VPN y concordancia	Con una prevalencia del 15%: VPP: 89,8% en *P. falciparum* y 88,4% en *P. no-falciparum*. VPN: 97,7% en *P. falciparum* y 93,8% en *P. no-falciparum.* Coeficiente kappa: 0,99 *P. falciparum* y 0,94 Aldolasa panmalárica	(39)
Evaluación de campo de la utilidad de test para el diagnóstico de la infección por COVID-19 en pacientes sintomáticos	RT-PCR	Estudio prospectivo.	412 pacientes sintomáticos	Sensibilidad Especificidad Valores predictivos Coeficiente kappa Curva ROC	Sensibilidad: 79,6% (95%CI 67,0-88,8) Especificidad: 100% (95%CI 98,7-100) Coeficiente kappa: 0,87 (95%CI 0,79-0,94) Prevalencia del 5%: VPN: 99% (95%CI 97,4-99,6) Curva ROC: RT-PCR CT <25 y SARS-CoV-2 RNA loads >5,9 log10 copias/ml discriminan entre PCR+/AG+ y PCR+/AG- con una especificidad y sensibilidad del 100%.	(40)

Tabla 16. Ejemplos de estudios fase 4 de evaluación de test rápidos de detección de antígenos.

Objetivo	Estándar de referencia	Diseño estudio	Muestras incluidas	Parámetro analizado	Resultado	Ref.
Capacidad predictiva de riesgo de exacerbación de EPOC de la prueba de antígeno de galactomanano en suero.	Criterios clínicos	Estudio observacional retrospectivo	191 pacientes	Riesgo de exacerbación de EPOP	Antígeno de galactomanano en suero >0,7 se asoció a mayor incidencia de exacerbación aguda de EPOC. P = 0.0039, Gray's test (*hazard ratio*, 2,162; 95% IC 1,26−3,69; P = 0,005)	(41)

4.2. ESTUDIOS DE EVALUACIÓN DE TÉCNICAS LAMP

Ejemplo Fase 1

La amplificación mediada por bucle (LAMP, del inglés *loop-mediated isothermal amplification*) es una forma de amplificación rápida de ácidos nucleicos que tiene varias ventajas frente a las técnicas de diagnóstico molecular tradicionales, como la PCR, incluida la rapidez, la cantidad de equipo necesario y la tolerancia a los fluidos biológicos (lo que facilita el análisis directo del material clínico, sin necesidad de extracción de los AANN). Actualmente existen múltiples aplicaciones de esta tecnología en el diagnóstico microbiológico. Por ejemplo, en esta revisión (42), justifican el uso de esta tecnología para el diagnóstico de la meningitis bacteriana.

En este trabajo (43), evalúan en concreto una LAMP para la detección precoz en niños de meningitis meningocócica invasiva a partir de frotis nasofaríngeos. Obtienen muy buenos resultados, pero entre sus limitaciones ya comentan que estos resultados son válidos para su población, pero que habría que analizar la validez clínica de este método en otras poblaciones, adolescentes y adultos, ya que la proporción de portadores asintomáticos en esas poblaciones es superior, lo que supondría más falsos positivos (menor especificidad).

Tabla 17. Ejemplos de estudios fase 2 de evaluación técnicas LAMP.

Objetivo	Estándar de referencia	Diseño estudio	Muestras incluidas	Parámetro analizado	Resultado	Ref.
Identificar los principales microorganismos causantes de neumonía intrahospitalaria directamente en muestras respiratorias	Cultivo	Casos-controles	LBA, BAS y AT procesados en rutina y conservados a -80ºC. LBA negativas inoculadas con microorganismos diana de la LAMP a estudio y diluciones de muestras positivas de BAS y AT.	Límite detección	LBA: 10^2 UFC/mL para *S. aureus*, *E. coli*, *P. aeruginosa* y *K. pneumoniae* y 10^4 UFC/mL para *S. maltophilia* y *A. baumannii*. BAS/AT: 10^2 UFC/mL para *P. aeruginosa*, *K. pneumoniae* y *A. baumannii*, 10^3 UFC/mL para *E. coli*, y 10^4 UFC/mL para *S. aureus* y *S. maltophilia*.	(44)
				Reactividad cruzada	No hubo reacciones cruzadas.	
Detección de ADN de *Trypanosoma cruzi* en sangre humana.	PCR cuantitativa	Casos-controles	Sangre periférica de 23 pacientes con Chagas y 10 controles	Interferencia	El uso de EDTA como anticoagulante resultó adecuado, aunque no se recomendaba para LAMP, debido a que el EDTA compite por el manganeso con el pirofosfato generado una vez iniciada la reacción. Por el contrario, en sangre heparinizada, la sensibilidad fue un orden inferior.	(45)
Detección de SARS-CoV-2	RT-PCR a tiempo real		Ocho réplicas de muestras.	Repetibilidad	Ver tabla 8 del artículo. 100% de las réplicas fueron detectadas para cada muestra por los tres operadores (CV <10%).	(46)
			Ocho réplicas de las muestras en dos plataformas y con tres operadores diferentes.	Reproducibilidad	Ver tabla 8 y 9 del artículo. 100% de las réplicas fueron detectadas para cada muestra por los tres operadores y en las dos plataformas (CV <10% en ambos casos).	

Abreviaturas: LBA, Muestras de lavado broncoalveolar; BAS, muestras de broncoaspirados; AT, muestras de aspirados traqueales (AT); CV, Coeficiente de variación.

Tabla 18. Ejemplos de estudios fase 3 de evaluación técnicas LAMP.

Objetivo	Estándar de referencia	Diseño estudio	Muestras incluidas	Parámetro analizado	Resultado	Ref.
Identificar los principales microorganismos causantes de neumonía intrahospitalaria directamente en muestras respiratorias.	Cultivo	Casos-controles	58 LBA 83 BAS/AT	Resultados individuales	Ver tablas 1 y 3 del artículo	(44)
				S (IC 95%)	BAL: 86,3 (73,7–94,3); BAS/AT: 94,6 (86,7–98,5)	
				E (IC 95%)	BAL: 100 (59–100); BAS/AT: 100 (66,4–100)	
				VPP (IC 95%)	BAL: 100 (89–100); BAS/AT: 100 (93,5–100)	
				VPN (IC 95%)	BAL: 50 (33,5–66,6); BAS/AT: 69,2 (46,5–85,4)	
				Accuracy (IC 95%)	BAL: 87,9 (76,7–95); BAS/AT: 95,2 (88,1–98,7)	
				Coeficiente kappa (IC 95%)	BAL: 70,9 (51,5–90,2); BAS/AT: (89,7 80–99,5)	
Detección de SARS-CoV-2	RT-PCR a tiempo real	Casos-controles	A partir de ARN: 196 Muestra directa: 119	LR+ (IC 95%)	A partir de ARN: 103,39 (14,69–727,57) Muestra directa: 23,57% (5,93–93,68)	(46)
				LR- (IC 95%)	A partir de ARN: 0,03 (0,01–0,10) Muestra directa: 0,34 (0,22–0,50)	
				Tiempo obtención resultado	Todas las muestras con un C_T ≤30 se detectaron en 16 minutos o menos.	
				Probabilidad pre-test	Paciente sintomático sin contacto de riesgo: 0,19 (19%). Paciente asintomático con contacto de riesgo: 0,12 (12%).	
				Probabilidad pre-test	Paciente sintomático sin contacto de riesgo: RT-LAMP positiva 0,81 (81%); RT-LAMP negativa 0,07 (7%). Paciente asintomático con contacto de riesgo: RT-LAMP negativa 0,05 (5%).	
Detección precoz en niños de meningitis meningocócica invasiva a partir de frotis nasofaríngeos.	Cultivo y PCR	Casos-controles	260 frotis naso-faríngeos	Flujo de pacientes incluidos	**Figura 15**	(43)

Abreviaturas: LBA, Muestras de lavado broncoalveolar; BAS, muestras de broncoaspirados; AT, muestras de aspirados traqueales (AT); LR, *Likelihood ratio*.

Figura 15. Ejemplo de flujo de pacientes incluidos en la evaluación del rendimiento clínico de una LAMP. (*Modificado de Waterfield y cols.* 2020) (43).

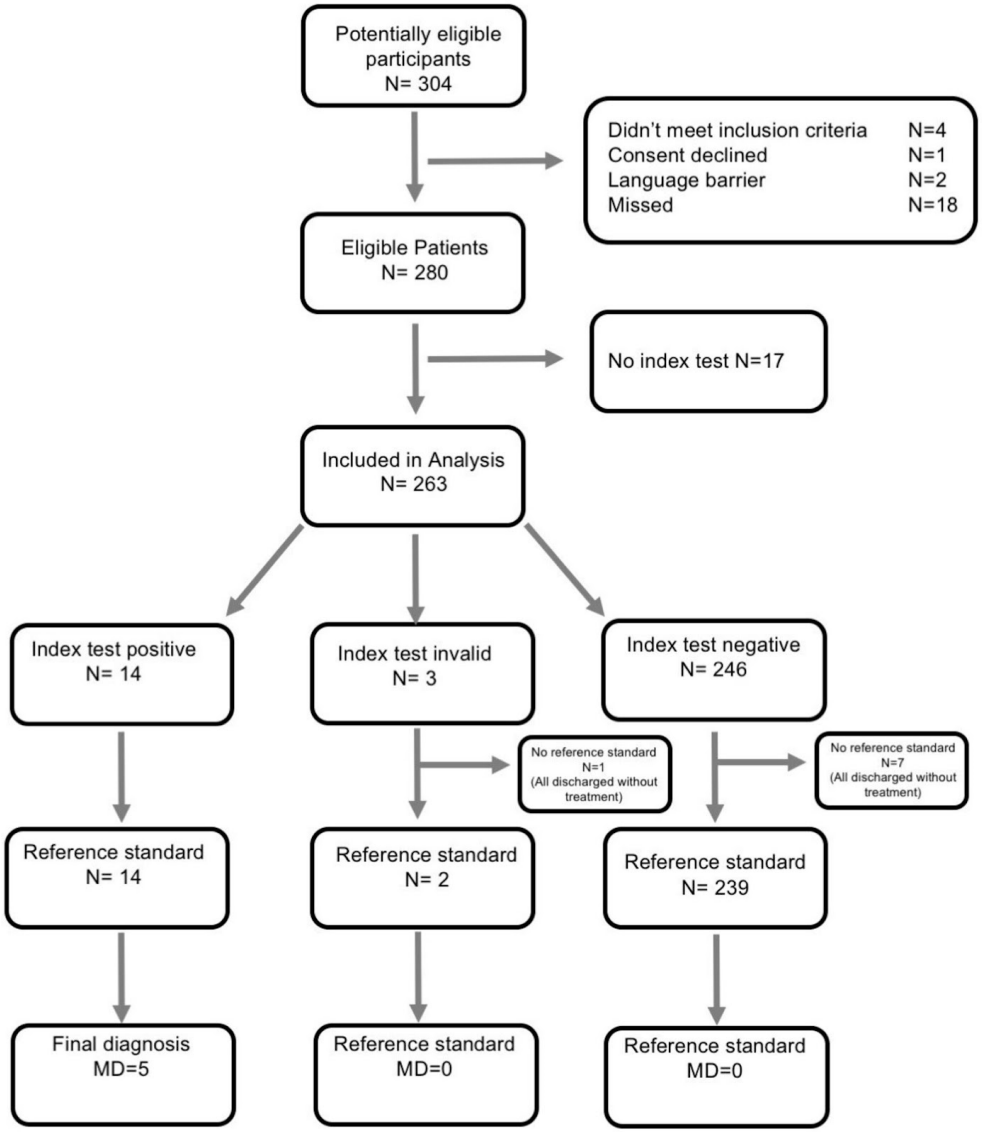


Tabla 19. Ejemplos de estudios fase 4 de evaluación de técnicas LAMP.

Objetivo	Estándar de referencia	Diseño estudio	Muestras incluidas	Parámetro analizado	Resultado	Ref.
Evaluar coste-efectividad del diagnóstico de tuberculosis utilizando pruebas moleculares en la población general tailandesa con sospecha de tuberculosis pulmonar.	Microscopia y cultivo.	Modelo Markov	NA	Coste	La realización de la LAMP para el diagnóstico de tuberculosis fue la opción más barata.	(47)
				Años de vida ganados ajustados por calidad	La realización de la LAMP para el diagnóstico de tuberculosis fue la opción con más años de vida ganados ajustados por calidad.	

4.3. ESTUDIOS DE EVALUACIÓN DE MÉTODOS DE DETERMINACIÓN DE SENSIBILIDAD ANTIBIÓTICA

Tabla 20. Ejemplos de estudios que evalúan el rendimiento de distintos sistemas para el estudio de la sensibilidad a antimicrobianos en distintas especies bacterianas o fúngicas.

Objetivo	Estándar de referencia	Muestras incluidas	Criterios de interpretación empleados	Resultado		Ref.
Evaluar el rendimiento del Phoenix NMIC-413 y la difusión con disco para la determinación de sensibilidad antibiótica a ertapenem, imipenem y meropenem, en *Enterobacterales* sensibles y resistentes a carbapenémicos	Microdilución en caldo	N= 303 aislados clínicos únicos de *Enterobacterales* (195 R a CBP; 108 S a CBP) Aislados QC: -*K. pneumoniae* BAA 1705 (*bla*$_{KPC-2}$) -*E. coli* ATCC 2452 (*bla*$_{NDM-1}$) -*K. pneumoniae* ATCC 700603 -*E. coli* ATCC 25922	CLSI M100	Phoenix-MD: EA y CA >93 % ME <1,75% VME= 0 MinE <5% para los tres CBP	DD-MD: CA>93% ME <3% MinE <5% para los tres CBP	(48)
Comparar los resultados de MD con MD-resazurina y MD-lectura por espectrofotometría para el estudio de sensibilidad antibiótica en *Nocardia* spp.	Microdilución en caldo	N= 15 aislados de *Nocardia* spp. (10 resistentes (R) a ciprofloxacino y 4 R a minociclina) (*N. brasiliensis* n=8; *N. farcinica* n= 2; *N. beijingensis* n=2; *N. ottitidiscaviarum* n=2; *N. asteroides* n=1)	CLSI M24	MD+RSZ-MD Ciprofloxacino: EA: 90,7% CA: 73,7% VME: 0,85% ME: 4,24% MinE: 21,2% Amikacina: EA: 90,8% CA: 100% Minociclina: EA: 80% CA:85,6% VME y ME: 0% MinE:14,4%	MD+OD-MD EA: 89,4% CA: 75% VME: 3,79% ME: 1,52% MinE: 19,7% EA: 99,2% CA: 100% EA: 93,8% CA:87,7% VME y ME:0% MinE: 12,3%	(49)

Tabla 20 continuación. Ejemplos de estudios que evalúan el rendimiento de distintos sistemas para el estudio de la sensibilidad a antimicrobianos en distintas especies bacterianas o fúngicas.

Objetivo	Estándar de referencia	Muestras incluidas	Criterios de interpretación empleados	Resultado	Ref.
Evaluar la precisión de tres sistemas comerciales de microdilución (Sensititre, UMIC, MicroScan) para el estudio de sensibilidad a colistina en bacilos gramnegativos.	Microdilución en caldo	N= 185 aislados clínicos de bacilos gramnegativos (*Enterobacterales* y BNNF) n=52 S a colistina n=19 R intrínseca a colistina n=114 R adquirida a colistina	EUCAST-2018 S ≤ 2 mg/L R > 2 mg/L	Resultados globales para *Enterobacterales* y BGNNF: MD-Sensititre CA: 97,8% VME: 3% ME: 0% MD-UMIC CA: 91,9% VME: 11,3 % ME: 0% MD-MicroScan CA: 91,9% (64,1% BNNF solos) VME: 0,8% ME: 3,1% (*Enterobacterales*) ME: 65% (BNNF)	(50)
Evaluar el rendimiento del sistema de determinación de sensibilidad antifúngica MIRONAUT-AM para *Aspergillus* spp. frente al sistema de microdilución	Microdilución en caldo	N= 78 aislados clínicos (mayoría respiratorios) de *Aspergillus* spp. n=46 *A. fumigatus* (9 no *wild-type* para itraconazol) n=14 *A. flavus* n= 10 *A. terreus* n= 10 *A. nidulans*	CLSI (M27-A2)	Anidulafungina/Anfotericina B EA: 99%/100% CA: 100%/96% Voriconazol/Itraconazol EA: 90%/87% CA: 97%/99%	(51)
Determinar la fiabilidad del sistema YS01 Vitek 2 para el estudio de sensibilidad a la anfotericina B, fluconazol y voriconazol en aislados clínicos de levaduras.	Microdilución en caldo	N= 614 *Candida* spp. n=10 *Criptococcus neoformans* n=1 *Geotrichum* spp. Aislados QC: *C. krusei* (ATCC 6258) *C. parapsilosis* (ATCC 22019)	CLSI (M27-A2)	Anfotericina B CA: 99,5% Fluconazol CA: 92% (*C. glabrata* CA: 51,4%, *C. krusei* CA: 53,3%) Voriconazol: CA: 98,2%	(52)

Abreviaturas: CBP, carbapenémicos; QC, control de calidad; ATCC, *American Type Culture Collection*; EA, acuerdo esencial; CA, acuerdo categórico; VME, errores muy graves; ME, errores graves; MinE, errores menores; DD, difusión con disco; MD, microdilución; MD+RSZ, microdilución con resazurina; MD+OD, microdilución con espectrofotometría; BNNF, bacilos gramnegativos no fermentadores; S, sensibles; R, resistentes.

6. BIBLIOGRAFÍA

1. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing directive 98/ 79/EC and Commission Decision 2010/227/EU. Offic J Eur Union 2017;L 117: 176e331.
2. Camaró-Sala ML, Martínez-García R, Olmos-Martínez P, Catalá-Cuenca V, Ocete-Mochón MD, Gimeno-Cardona C. Validación y verificación analítica de los métodos microbiológicos. 2013. 48. Procedimientos en Microbiología Clínica. Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2018.
3. Gimeno Cardona C, Gómez E, Leiva J, Navarro D, Pérez Sáenz JL. Evaluación económica de las pruebas diagnósticas en Microbiología Clínica. 2018. 64. Navarro D (coordinador). Procedimientos en Microbiología Clínica. Procedimientos en Microbiología Clínica. Cercenado Mansilla E, Cantón Moreno R (editores). Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica (SEIMC). 2018.
4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. HCW, Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326:41–44.
5. Leeflang MMG, Allerberger F. How to: evaluate a diagnostic test. Clin Microbiol Infect. 2019;25:54–59.
6. Rutjes AWS, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PMM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005. 51:1335–1341.
7. Wang H, Zhou H, Jiang R, Qian Z, Wang F, Cao L. Globulin, the albumin-to-globulin ratio, and fibrinogen perform well in the diagnosis of periprosthetic joint infection. BMC Musculoskelet Disord. 2021; 22:583.
8. Kim YJ, Lê HG, Na B-K, Kim BG, Jung Y-K, Kim M, et al. Clinical utility of cerebrospinal fluid vitamin D-binding protein as a novel biomarker for the diagnosis of viral and bacterial CNS infections. BMC Infect Dis. 2021; 21(1):240.
9. Technical Specifications Series for submission to WHO Prequalification – Diagnostic Assessment: Malaria rapid diagnostic tests. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.
10. Technical Guidance Series for WHO Prequalification – Diagnostic Assessment: Principles of performance studies. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.
11. In vitro diagnostic (IVDs) medical devices used for the qualitative and quantitative detection of Hepatitis C RNA. Geneva: World Health Organization; 2021 (Technical specifications series for submission to WHO prequalification – diagnostic assessment, TSS10). Licence: CC BY-NCSA 3.0 IGO.
12. Technical Specifications Series for submission to WHO Prequalification – Diagnostic Assessment: Human Immunodeficiency Virus (HIV) rapid diagnostic tests for professional use and/or selftesting. Geneva: World Health Organization; 2016. Licence: CC BY-NC-SA 3.0 IGO.
13. Chevaliez S, Onelia F, Pacenti M, Goldstein E, Galán J-C, Martínez-García L, et al. Multicenter clinical evaluation of alinity m HCV assay performance. J Clin Virol 2020; 129:104531.
14. Hildenbrand C, Wedekind L, Li G, vonRentzell JE, Shah K, Rooney P, et al. Clinical evaluation of Roche COBAS® AmpliPrep/COBAS® TaqMan® CMV test using nonplasma samples. J Med Virol 2018; 90:1611–1619.
15. Nam M, Song DY, Song SH, Roh EY, Shin S, Park KU, et al. Performance evaluation of immunoassay for infectious diseases on the Alinity i system. J Clin Lab Anal 2021; 35:e23671.
16. Molina Arias M, Ochoa Sangrador C. Pruebas diagnósticas con resultados continuos o politómicos. Curvas ROC. Evid Pediatr. 2017;13:12.
17. Ortiz de la Tabla V, Berruezo M, García Payá E, Fernández M, García JA, Masiá M, et al. Evaluation of the Virclia® automated chemiluminescent immunoassay system for diagnosing pneumonia caused by Mycoplasma pneumoniae. J Clin Lab Anal 2018; 32:e22431.
18. Fidalgo B, Rubio E, Pastor V, Parera M, Ballesté-Delpierre C, Fernández MJ, et al. Improved diagnosis of gastrointestinal infections using a semi-automated multiplex real-time PCR for detection of enteropathogens. J Med Microbiol 2021; 70(9).
19. Biswas B. Clinical performance evaluation of molecular diagnostic tests. J Mol Diagn. 2016; 18:803–812.
20. Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, et al. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. J Clin Epidemiol 2019; 114:38–48.
21. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015; 351:h5527.
22. JuareziSI, Nuñez AE, Aranda MM, Mojica D, Kim AA, Parekh B. Field evaluation of four rapid tests for diagnosis of HIV Infection in Panama. J Clin Microbiol 2016; 54:1127–1129.

23. Doust JA, Bell KJL, Leeflang MMG, Dinnes J, Lord SJ, Mallett S, et al. Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection. BMJ 2021; 372:n568.
24. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med 1975; 293:257.
25. Aznar-Oroval E, Mancheño-Alvaro A, García-Lozano T, Sánchez-Yepes M. Likelihood ratio and Fagan's nomogram: 2 basic tools for the rational use of clinical laboratory tests. Rev Calid Asist. 2013; 28:390–391.
26. Abarca K, Winter M, Marsac D, Palma C, Contreras AM, Ferrés M. Accuracy and diagnostic utility of IgM in Bartonella henselae infections. Rev Chilena Infectol. 2013; 30:125–128.
27. Vergara A, Fernández-Pittol MJ, Muñoz-Mahamud E, Morata L, Bosch J, Vila J, et al. Evaluation of lipocalin-2 as a biomarker of periprosthetic joint infection. J Arthroplasty. 2019; 34:123–125.
28. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. PLoS One 2019; 14:e0223832.
29. Munrós J, Vergara A, Bataller E, García-Lorenzo B, Álvarez-Martínez MJ, Bosch J. Performance of a rapid molecular test to detect Chlamydia trachomatis and Neisseria gonorrhoeae in women with pelvic inflammatory disease. Enferm Infecc Microbiol Clin (Engl Ed). 2021; S0213-005X(21)00090–2.
30. Minneci PC, Deans KJ. Clinical trials. Semin Pediatr Surg. 2018; 27:332–337.
31. Lijmer JG, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol. 2009; 62:364–373.
32. Lewis RF, Abrahamowicz M, Côté R, Battista RN. Predictive power of duplex ultrasonography in asymptomatic carotid disease. Ann Intern Med. 1997; 127:13–20.
33. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol. 2005; 23:2020–2027.
34. Côté R, Battista RN, Abrahamowicz M, Langlois Y, Bourque F, Mackey A. Lack of effect of aspirin in asymptomatic patients with carotid bruits and substantial carotid narrowing. The Asymptomatic Cervical Bruit Study Group. Ann Intern Med. 1995; 123:649–655.
35. Af Geijerstam JL, Oredsson S, Britton M, OCTOPUS Study Investigators. Medical outcome after immediate computed tomography or admission for observation in patients with mild head injury: randomised controlled trial. BMJ 2006; 333(7566):465.
36. Neugebauer EAM, Rath A, Antoine SL, Eikermann M, Seidel D, Koenen C, et al. Specific barriers to the conduct of randomised clinical trials on medical devices. Trials. 2017; 18:427.
37. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials. 2006; 1:e9.
38. TDR Diagnostics Evaluation Expert Panel, Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, et al. Evaluation of diagnostic tests for infectious diseases: general principles. Nat Rev Microbiol. 2010; 8:S17-29.
39. Farcas GA, Zhong KJY, Lovegrove FE, Graham CM, Kain KC. Evaluation of the Binax NOW ICT test versus polymerase chain reaction and microscopy for the detection of malaria in returned travelers. Am J Trop Med Hyg. 2003; 69:589–592.
40. Albert E, Torres I, Bueno F, Huntley D, Molla E, Fernández-Fuentes MÁ, et al. Field evaluation of a rapid antigen test (PanbioTM COVID-19 Ag Rapid Test Device) for COVID-19 diagnosis in primary healthcare centres. Clin Microbiol Infect. 2021; 27:472.e7-472.e10.
41. Yoshimura K, Suzuki Y, Inoue Y, Nishimoto K, Mori K, Karayama M, et al. Utility of serum Aspergillus-galactomannan antigen to evaluate the risk of severe acute exacerbation in chronic obstructive pulmonary disease. PLoS One. 2018; 13:e0198479.
42. Seki M, Kilgore PE, Kim EJ, Ohnishi M, Hayakawa S, Kim DW. Loop-Mediated Isothermal Amplification methods for diagnosis of bacterial meningitis. Front Pediatr. 2018; 6:57.
43. Waterfield T, Lyttle MD, McKenna J, Maney J-A, Roland D, Corr M, Woolfall K, Patenall B, Shields M, Fairley D, Paediatric Emergency Research in the UK and Ireland (PERUKI). Loop-mediated isothermal amplification for the early diagnosis of invasive meningococcal disease in children. Arch Dis Child. 2020; 105:1151–1156.
44. Vergara A, Boutal H, Ceccato A, López M, Cruells A, Bueno-Freire L et al. Assessment of a loop-mediated isothermal amplification (LAMP) assay for the rapid detection of pathogenic bacteria from respiratory samples in patients with hospital-acquired pneumonia. Microorganisms. 2020; 8:E103.
45. Besuschio SA, Llano Murcia M, Benatar AF, Monnerat S, Cruz I, Picado A, et al. Analytical sensitivity and specificity of a loop-mediated isothermal amplification (LAMP) kit prototype for detection of Trypanosoma cruzi DNA in human blood samples. PLoS Negl Trop Dis. 2017; 11:e0005779.

46. Fowler VL, Armson B, Gonzales JL, Wise EL, Howson ELA, Vincent-Mistiaen Z, et al. A highly effective reverse-transcription loop-mediated isothermal amplification (RT-LAMP) assay for the rapid detection of SARS-CoV-2 infection. J Infect. 2021; 82:117–125.
47. Chitpim N, Jittikoon J, Udomsinprasert W, Mahasirimongkol S, Chaikledkaew U. Cost-utility analysis of molecular testing for tuberculosis diagnosis in suspected pulmonary tuberculosis in Thailand. Clinicoecon Outcomes Res. 2022; 14:61–73.
48. Zhang J, Jia P, Zhu Y, Zhang G, Xu Y, Yang Q. Performance evaluation of BD Phoenix NMIC-413 antimicrobial susceptibility testing panel for imipenem, meropenem, and ertapenem against clinical carbapenem-Resistant and Carbapenem-susceptible Enterobacterales. Front Med (Lausanne). 2020; 8:643194.
49. Caso Coelho V, Pereira Neves SD, Cintra Giudice M, Benard G, Lopes MH, Sato PK. Evaluation of antimicrobial susceptibility testing of Nocardia spp. isolates by broth microdilution with resazurin and spectrophotometry. BMC Microbiol. 2021; 21:331.
50. Jayol A, Nordmann P, André C, Poirel L, Dubois V. Evaluation of three broth microdilution systems to determine colistin susceptibility of Gram-negative bacilli. J Antimicrob Chemother. 2018; 73:1272–1278.
51. Nuh A, Ramadan N, Schelenz S, Armstrong-James D. Comparative evaluation of MIRONAUT-AM and CLSI broth microdilution method for antifungal susceptibility testing of Aspergillus species against four commonly used antifungals. Med Mycol. 2022; myab081.
52. Borghi E, Iatta R, Sciota R, Biassoni C, Cuna T, Montagna MT, Morace G. Comparative evaluation of the Vitek 2 yeast susceptibility test and CLSI broth microdilution reference method for testing antifungal susceptibility of invasive fungal isolates in Italy: the GISIA3 study. J Clin Microbiol. 2010; 48:3153–3157.

Servicio / Unidad de Microbiología Hospital.....	Evaluación del rendimiento analítico y clínico de una PCR a tiempo real para la detección de *Candida auris*	PNT-ERAC-01	
		Edición Nº 01	Página 1 de 5

PNT-ERAC-01
Evaluación del rendimiento analítico y clínico de una PCR a tiempo real para la detección de *Candida auris*

ELABORADO		REVISADO Y APROBADO	
Nombre / Firma	Fecha	Nombre / Firma	Fecha

EDICIÓN	FECHA	ALCANCE DE LAS MODIFICACIONES
01		Edición inicial

Servicio / Unidad de Microbiología Hospital.....	Evaluación del rendimiento analítico y clínico de una PCR a tiempo real para la detección de *Candida auris*	PNT-ERAC-01	
		Edición Nº 01	Página 2 de 5

1. PROPÓSITO Y ALCANCE

El propósito del presente documento es describir el proceso de evaluación analítica y clínica de una PCR a tiempo real para la detección de *Candida auris*.

2. DOCUMENTOS DE CONSULTA

1. Leach L, Zhu Y, Chaturvedi S. Development and validation of a real-time PCR assay for rapid detection of Candida auris from surveillance samples. J Clin Microbiol. 2018;56(2):e01223-17. doi: 10.1128/JCM.01223-17.

3. PROCEDIMIENTO

Se presentan como ejemplo los resultados obtenidos en el estudio de *Leach L, et al*, indicado anteriormente.

3.1. TEST DIAGNÓSTICO EVALUADO

Nombre del test y del fabricante: PCR a tiempo real casera.
Tipo de muestras: ADN extraído de la cepa de *C. auris* M5658, frotis y esponjas.
Diana: Gen ITS2 de *C. auris*.
Fundamento: PCR a tiempo real casera basada en sondas TaqMan.
Plataforma: Termociclador ABI 7500 FAST.

3.2. INTENCIÓN DE USO

Cribado de pacientes portadores de *C. auris.*

3.3. TIPO DE MUESTRA Y USUARIOS

La prueba está pensada para su uso en hospitales y se realiza la evaluación tanto para frotis de pacientes (axila-ingle, axila, ingle, fosas nasales, oído, recto y herida) como para superficies.

3.4 MÉTODO DE REFERENCIA

El estándar de referencia empleado es el cultivo.

3.5 CONDICIONES DE REALIZACIÓN

Cada carrera de PCR también incluye un control positivo de extracción (*C. auris* M5658; 103 UFC/50 l) y un control positivo de amplificación (*C. auris* M5658; 0,02 pg/l), así como controles negativos de extracción (solo reactivos) y de amplificación (agua esterilizada libre de nucleasas). Para evitar cualquier contaminación cruzada, se sigue un flujo de trabajo unidireccional manteniendo las áreas de preparación de reactivos, preparación de muestras y amplificación/detección, separadas.

3.6 TAMAÑO MUESTRAL

Ejemplo (estudio de Leach L, *et al*., ver documentos de consulta):

Se analizan 623 muestras de vigilancia, incluyendo 365 frotis de pacientes y 258 esponjas ambientales. Cada muestra se testa por duplicado.

4. ANÁLISIS Y PRESENTACIÓN DE LOS RESULTADOS

4.1. SENSIBILIDAD ANALÍTICA

La sensibilidad analítica del ensayo se determina a partir de diluciones seriadas de una cantidad inicial conocida de la cepa *C. auris* M5658. Se realizan tres extracciones de ADN independientes.

La PCR en tiempo real de *C. auris* fue lineal en 5 órdenes de magnitud, y el límite de detección del ensayo fue de 1 UFC C. auris / PCR utilizando 45 ciclos de PCR en los tres procesos de extracción.

4.2. REPRODUCIBILIDAD

La reproducibilidad analítica del ensayo se determina a partir de extractos con concentración alta (10^5 CFU/50 µL), moderada (10^3 CFU/50 µL) o baja (10^2 CFU/50 µL) de ADN de *C. auris.* Los mismos extractos se analizan por triplicado en tres días diferentes (interensayo), así como en el mismo día (intraensayo).

La reproducibilidad y la especificidad también se evaluan en frotis y esponjas de vigilancia negativas testadas previamente. Estas muestras negativas se enriquecieron al azar con 5 µL de una solución con alta (10^5 CFU/50 µL), moderada (10^3 CFU/50 µL), o baja (10^2 CFU/50 µL) concentración de *C. auris*. Diez alícuotas de frotis y esponjas también se enriquecieron con una cantidad moderada (10^3 CFU/50 µL) de otras especies de *Candida.*

El ensayo fue altamente reproducible, ya que produjo valores de *Cycle threshold* (Ct) consistentes para una concentración dada en tres días diferentes de realización de la prueba, así como en los replicados el mismo día de la prueba. El coeficiente de variación (CV) fue inferior al 5%, lo que confirma la alta reproducibilidad del ensayo.

Reproducibilidad interensayo:

Concentración	UFC/ reac-ción	Día 1			Día 2			Día 3			Media Ct ± DS	% CV
		Ct1	**Ct2**	**Ct3**	**Ct1**	**Ct2**	**Ct3**	**Ct1**	**Ct2**	**Ct3**		
Alta												
Moderada												
Baja												

Reproducibilidad intraensayo:

Concentración	UFC/ reacción	Ct1	Ct2	Ct3	Media Ct ± DS	% CV
Alta						
Moderada						
Baja						

Todas las muestras de frotis y esponjas enriquecidas con baja cantidad y una elevada cantidad de *C. auris* fueron positivos mediante el ensayo de PCR en tiempo real, con un CV de menos del 5 %.

4.3. ESPECIFICIDAD ANALÍTICA

La especificidad analítica del ensayo se realiza analizando aproximadamente 1 ng de ADN genómico de un panel de referencia y de aislados clínicos de hongos (levaduras y hongos filamentosos), bacterias, parásitos y virus.

El ensayo fue altamente específico, ya que ninguno de los otros microorganismos reaccionaron de forma cruzada.

Todas las muestras de frotis y esponjas enriquecidas con otras especies de *Candida* fueron negativas.

Especificidad:

ID Cepa	Microorganismo	Fuente	Resultado PCR
			…

4.4 RENDIMIENTO CLÍNICO

Resultados para frotis:

Real-time PCR result	No. of swabs with indicated culture result: Positive	No. of swabs with indicated culture result: Negative	Accuracy (%)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (%)
Positive	46	3	98	89 (77–96)	99 (97–100)	94
Negative	6	310				

[a]CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

Resultados para esponjas:

Real-time PCR result	No. of sponges with indicated culture result: Positive	No. of sponges with indicated culture result: Negative	Accuracy (%)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (%)	NPV (%)
Positive	32	26	90	100 (89–100)	89 (84–92)	55	100
Negative	0	200					

Las áreas bajo las curvas ROC para el ensayo de PCR en tiempo real fueron 0,940 para frotis y 0,978 para esponjas. El índice de Youden mostró que el valor de corte óptimo de positividad era Ct ≤37,0 y ≤38,0 para frotis y esponjas, respectivamente.

5. CONCLUSIONES

El ensayo de PCR en tiempo real desarrollado detecta de forma precisa y rápida *C. auris* a partir de muestras de vigilancia, lo que permitiría un control más eficaz para prevenir la propagación de este patógeno fúngico multirresistente emergente en hospitales.

6. BIBLIOGRAFÍA

1. Leach L, Zhu Y, Chaturvedi S. Development and Validation of a Real-Time PCR Assay for Rapid Detection of Candida auris from Surveillance Samples. J Clin Microbiol. 2018 Jan 24;56(2):e01223-17. doi: 10.1128/JCM.01223-17. PMID: 29187562; PMCID: PMC5786737.

Servicio / Unidad de Microbiología Hospital.....	Evaluación del rendimiento clínico de pruebas rápidas para el diagnóstico de SARS-CoV-2	PNT-ERAC-02	
		Edición Nº 01	Página 1 de 5

PNT-ERAC-02
Evaluación del rendimiento clínico de pruebas rápidas para el diagnóstico de SARS-CoV-2

ELABORADO		REVISADO Y APROBADO	
Nombre / Firma	Fecha	Nombre / Firma	Fecha

EDICIÓN	FECHA	ALCANCE DE LAS MODIFICACIONES
01		Edición inicial

Servicio / Unidad de Microbiología Hospital.....	Evaluación del rendimiento clínico de pruebas rápidas para el diagnóstico de SARS-CoV-2	PNT-ERAC-02	
		Edición Nº 01	Página 2 de 5

1. PROPÓSITO Y ALCANCE

El propósito del presente procedimiento es describir el proceso seguido por el Servicio de Microbiología para evaluar el rendimiento clínico de una prueba rápida para la detección de antígeno de SARS-CoV-2 en muestras de frotis oro-nasofaríngeo

2. DOCUMENTOS DE CONSULTA

1. Möckel M, Corman VM, Stegemann MS, Hofmann J, Stein A, Jones TC, et al. SARS-CoV-2 antigen rapid immunoassay for diagnosis of COVID-19 in the emergency department. Biomarkers. 2021;26:213-220. doi:10.1080/1354750X.2021.18767

3. PROCEDIMIENTO

Se presentan como ejemplo los resultados obtenidos en el estudio de Möckel M, *et al,* indicado anteriormente.

3.1. TEST DIAGNÓSTICO EVALUADO

Nombre del test y del fabricante: se evalúa el test rápido de detección de antígeno de Roche SARS-CoV-2, fabricado por SD Biosensor.
Tipo de muestras: la evaluación se realiza empleando muestras oro-nasofaríngeas. Se toman dos muestras a cada paciente. La primera se emplea para realizar la RT-PCR. La segunda, para realizar el test de detección de antígenos. La recogida de muestras la lleva a cabo personal de enfermería del Servicio de Urgencias.
Diana: el test diagnóstico se basa en la detección de la proteína de la nucleocápside del virus SARS-CoV-2.
Fundamento: inmunocromatografía.
Plataforma: no necesaria.
Resultados de evaluación analítica: se pueden consultar en el *insert* del fabricante.
Interpretación: la interpretación de los resultados se basa en la lectura visual de la inmunocromatografía, siguiendo las indicaciones del fabricante. La lectura de la inmunocromatografía se realiza por dos sanitarios de manera independiente. Ambos deben alcanzar un consenso en la lectura

3.2.INTENCIÓN DE USO

El test evaluado se empleará como método diagnóstico en pacientes sintomáticos con sospecha de COVID-19.

3.3 POBLACIÓN OBJETIVO

Ejemplo (estudio de Möckel M, *et al*, ver documentos de consulta):

La población objetivo son pacientes sintomáticos. Los pacientes se incluirán de manera consecutiva, de acuerdo a los siguientes criterios de inclusión:
• Sintomatología respiratoria aguda y/o pérdida del gusto u olfato.
• Contacto con un caso confirmado de COVID-19 hasta un máximo de 14 días antes de la aparición de cualquier síntoma COVID-19.
• Signos clínicos o radiológicos de neumonía viral en el contexto de un brote en residencias o hospitales.

Las contraindicaciones para el empleo del test de antígeno son las siguientes:
• Cribado de pacientes o personal sanitario asintomáticos.
• Cribado de personas asintomáticas que regresen de zonas de riesgo.

El periodo de tiempo durante el cual se reclutarán los pacientes será entre el día de mes de año y el día de mes de año.

3.4. REQUERIMIENTOS

La toma de muestra será realizada por personal sanitario cualificado y entrenado.

3.5. PREVALENCIA DEL PROBLEMA EN LA COMUNIDAD

Debido a la cambiante situación de pandemia, la prevalencia de SARS-CoV-2 es muy fluctuante. Este dato se obtendrá tras el estudio de evaluación, indicándose en resultados la prevalencia en la población objetivo.

3.6. LUGAR DE REALIZACIÓN DEL TEST

El estudio se llevará a en cuatro unidades de urgencias de adultos y una unidad de urgencias pediátricas.

El test de antígeno se realizará en los servicios de urgencias.

La RT-PCR se procesará en el Servicio de Microbiología.

3.7. TEST DE REFERENCIA

El estándar de referencia en el diagnóstico de SARS-COV-2 es la RT-PCR. En el presente estudio de evaluación, se emplea la PCR de Roche Cobas SARS-CoV-2 (Penzberg, Germany) en el instrumento de Roche Cobas 6800. Los datos de rendimiento analítico de la RT-PCR se pueden consultar en insert del fabricante.

La RT-PCR se realizará en el Servicio de Microbiología, sin conocer el resultado del test de antígeno.

3.8. TAMAÑO MUESTRAL

El tamaño muestral será el número de pacientes incluidos durante el periodo de estudio. Se informará de los pacientes excluidos y sus motivos.

3.9. TIPOS DE MUESTRAS

Las muestras empleadas serán frotis oro-nasofaríngeos de pacientes atendidos en los Servicios de Urgencias con diagnóstico de sospecha de SARS-CoV-2. Los pacientes serán incluidos de manera consecutiva durante el período de estudio.

No será necesario el almacenamiento de las muestras previo a su procesamiento. Las muestras para realizar el test de antígeno se procesarán de manera inmediata en el Servicio de Urgencias, y la muestra para realizar la PCR, se enviará al Servicio de Microbiología para realizar la PCR en la misma jornada de trabajo.

Una vez procesadas, las muestras se archivarán en congeladores del Servicio de Microbiología.

3.10 ASPECTOS ÉTICOS.

Se pedirá la aprobación del comité de ética.

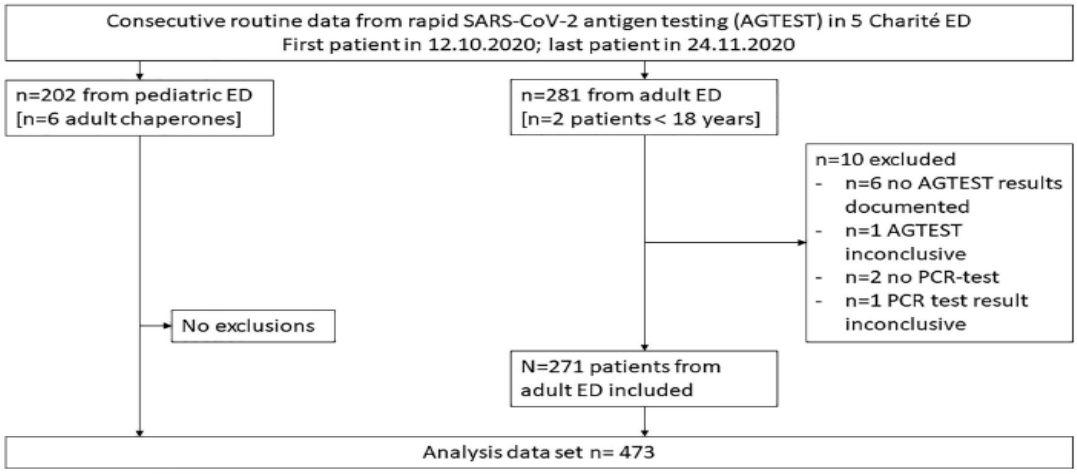
4. ANÁLISIS Y PRESENTACIÓN DE LOS RESULTADOS

4.1 DIAGRAMA DE FLUJO.

Siguiendo las directrices STARD, se incluye un diagrama de flujo para informar a las personas incluidas y excluidas en el estudio y a las personas incluidas, pero cuyas muestras no se analizaron, o no se analizaron tanto por el método evaluado como por el de referencia u otros motivos de exclusión.

En la figura 1 se muestran los pacientes incluidos. Son 202 pacientes pediátricos, sin pacientes excluidos. Y 271 pacientes adultos, con 10 excluidos por distintas causas. En total se incluyen 473 pacientes.

Figura 1. Diagrama de flujo estudio de evaluación rendimiento clínico.



(Tomado de Möckel M, *et al*, ver documentos de consulta)

4.2.RENDIMIENTO CLÍNICO

Los valores del rendimiento clínico junto con los intervalos de confianza al 95% fueron los siguientes:

En población pediátrica:

- Sensibilidad: 72,0% [53,3/86,7]
- Especificidad: 99,4% [97,3/99,9]
- Valor predictivo positivo (VPP): 94,7% [78,3/99,7]
- Valor predictivo negativo (VPN): 96,2% [92,7/98,3]
- La prevalencia en esta población objetivo fue: 12,4% [8,3/17,4]
- Coeficiente kappa: 0,79 [65,8/93,5]

En población adulta:

- Sensibilidad: 75,3% [65,8/83,4]
- Especificidad: 100% [98,4/100]
- VPP: 100% [95,7/100]
- VPN: 89,2% [84,5/93,9]
- La prevalencia en esta población objetivo fue: 32,8% [27,4/38,6]
- Coeficiente kappa: 0,80 [72,5/88,2]

5. CONCLUSIONES

El empleo del test de detección de antígeno de SARS-CoV-2 es útil en pacientes sintomáticos que acuden al Servicio de Urgencias. Por ello, se recomienda su empleo para la detección rápida de pacientes con COVID-19, pero no como método de cribado en personas asintomáticas. Debido a los falsos negativos del test de antígeno, se recomienda una RT-PCR de confirmación en esta población de estudio.

6. LIMITACIONES DEL PROCEDIMIENTO

El empleo de una doble muestra, una para la realización del test de antígeno y otra para la realización de la RT-PCR puede producir resultados discrepantes.

7.BIBLIOGRAFÍA

1. Möckel M, Corman VM, Stegemann MS, et al. SARS-CoV-2 antigen rapid immunoassay for diagnosis of COVID-19 in the emergency department. Biomarkers. 2021;26(3):213-220. doi:10.1080/1354750X.2021.1876769

Servicio / Unidad de Microbiología Hospital.....	Evaluación del rendimiento de un sistema de determinación de sensibilidad antibiótica	PNT-ERAC-03	
		Edición Nº 01	Página 1 de 6

PNT-ERAC-03
Evaluación del rendimiento de un sistema de determinación de sensibilidad antibiótica

ELABORADO		REVISADO Y APROBADO	
Nombre / Firma	Fecha	Nombre / Firma	Fecha

EDICIÓN	FECHA	ALCANCE DE LAS MODIFICACIONES
01	2022	Edición inicial

1. PROPÓSITO Y ALCANCE

El propósito del presente documento es definir la metodología para la evaluación del rendimiento de un sistema de determinación de sensibilidad antibiótica utilizando como comparador el método de referencia.

2. FUNDAMENTO

Para la evaluación de sistemas de determinación de sensibilidad antibiótica basada en la concentración mínima inhibitoria (CMI), la norma ISO 20776-2:2021 recomienda el cálculo del **acuerdo esencial (EA)** y el **sesgo**. El EA permite conocer el grado de concordancia entre los resultados del método en evaluación y el de referencia. Además, los sistemas que producen algún tipo de sesgo generan resultados que de manera sistemática se encuentran por encima o por debajo del resultado proporcionado por el método de referencia.

3. DOCUMENTOS DE CONSULTA

1. ISO 20776-2:2021. Clinical laboratory testing and *in vitro* diagnostic test systems — Susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices — Part 2. Evaluation of performance of antimicrobial susceptibility test devices against reference broth microdilution.(https://www.iso.org/standard/79377.html)
2. CLSI. *Performance standards for antimicrobial susceptibility testing*. 31th ed. CLSI supplement M100. Wayne, PA: Clinical and Laboratory Standars Institute; 2021.
3. The European Committee on Antimicrobial Susceptibility Testing. Routine and extended internal quality control for MIC determination and disk diffusion as recommended by EUCAST. Version 12.0, 2022. http://www.eucast.org.
4. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters. Version 12.0, 2022. (http://www.eucast.org).

4. PROCEDIMIENTO

4.1. TIPO Y NÚMERO DE MUESTRAS

Incluir en la evaluación al menos 300 aislados bacterianos de origen clínico: al menos 100/laboratorio participante en la evaluación, que representen el mayor número de géneros y especies posible. Una distribución posible puede ser:

- **N=25 Aislados contemporáneos**: recuperados de muestras clínicas en los seis meses previos y que apenas se hayan subcultivado.

- **N=75 Aislados de *stock***: recuperados de una muestra clínica que se ha guardado o que pertenece a una colección de archivo sin importar el tiempo transcurrido desde su aislamiento primario.

El empleo de aislados de stock permite que se incluyan microorganismos con mecanismos de resistencia conocidos, poco frecuentes o pertenecientes a un género o especie que no se aísla habitualmente en el laboratorio.

Cuando sea posible se ensayarán al menos 25 aislados cuya CMI se encuentre en escala.

4.2. MÉTODO DE REFERENCIA

El método de referencia empleado es el de microdilución en caldo.

4.3. CÁLCULO DEL ACUERDO ESENCIAL (EA)

El acuerdo esencial se define como el porcentaje de CMIs respecto del total de determinaciones obtenidas por el sistema de prueba que se encuentran dentro del rango de ± 1 dilución doble respecto la CMI del sistema de referencia. Se consideran aceptables los **valores ≥ 90%**.

Para el cálculo del EA es necesario que el rango de concentraciones del sistema en evaluación sea el mismo o se adecúe al del método de referencia. Si el sistema de referencia tiene una escala de concentraciones más amplia que el test de prueba, los valores de CMIs (del método de referencia) inferiores al valor más pequeño informado por el test de prueba deben contabilizarse de manera combinada en la CMI más baja que proporciona el método de prueba. De igual forma, las CMIs del método de referencia que sean mayores que la CMI más alta proporcionada por el test de prueba deben combinarse para el cálculo del EA y el sesgo.

Ejemplo: el método de prueba proporciona las CMIs ≤2, 4, 8, 16, 32 y >32 para un determinado antibiótico mientras que las del método de referencia se encuentran en la escala ≤0,5 hasta >128. En la Tabla 1 y tabla 2 se muestra cómo se realizaría el ajuste de los valores del método de referencia a la escala del test de prueba para el cálculo del EA.

Tabla 1. Ejemplo de frecuencia y distribución de CMIs proporcionadas por el método de referencia

CMI	≤0,5	1	2	4	8	16	32	64	128	>128
Nº aislados	44	85	92	48	13	3	8	3	1	3

Tabla 2. Ajuste de CMIs del método de referencia a la escala del método de prueba

CMI	≤2	4	8	16	32	>32
Nº de aislados	221	48	13	3	8	7

De manera práctica el cálculo de EA puede llevarse a cabo representando en una tabla la distribución de CMIs de los aislados ensayados y las CMIs obtenidas por el método de referencia.

Tabla 3. Ejemplo de comparación de CMIs, para un antibiótico y microorganismo/grupo de microorganismos, por el método de prueba y el de referencia

		Método de referencia						Total
		≤2	4	8	16	32	>32	
Método en prueba	≤ 2	154	17	0	0	0	0	171
	4	66	30	8	1	1	0	106
	8	1	1	1	0	0	0	3
	16	0	0	3	0	2	0	5
	32	0	0	1	2	3	3	9
	>32	0	0	0	0	2	4	6
Total		221	48	13	3	8	7	300

A continuación, se contabiliza el número de valores de CMI coincidentes, los que varían en ± 1 dilución, ± 2 diluciones y ± 3 diluciones. Con estos datos se calcula el acuerdo esencial expresado en porcentaje.

Tabla 4. Representación del número de microorganismos con una CMI concordante (0) o con acuerdos 1, 2 o 3 diluciones del método de referencia (± 1, ± 2 o ≤ -3 o ≥ +3 diluciones).

Nº de acuerdos ± dilución por encima, debajo o en la misma CMI del método de referencia							**Acuerdo esencial (EA)**
≤ -3	**-2**	**-1**	**0**	**+1**	**+2**	**≥ +3**	296/300= **98,7 %**
1	1	30	192	74	2	0	

4.4. CÁLCULO DEL SESGO/DESVIACIÓN

Para evaluar la desviación es necesario comparar los porcentajes de resultados del test de prueba que son mayores y menores que el test de referencia.

La desviación se calcula para todos los microorganismos ensayados, considerándose aceptable una desviación (diferencia entre el porcentaje superior e inferior) que se encuentre en el intervalo de **– 30%** hasta **+ 30%.**

Para calcular la desviación es necesario que el número de aislados en escala (cuya CMI se encuentre en el rango de CMIs medidas por el método de prueba y referencia) sea al menos 25.

4.4.1. CÁLCULO DEL PORCENTAJE DE VALORES DEL TEST DE PRUEBA CON CMIs SUPERIORES AL MÉTODO DE REFERENCIA

Si usamos como ejemplo los datos de la tabla 3, el rango de diluciones en los que es posible que el test de prueba proporcione resultados de CMI superiores a los de referencia es ≤2 hasta 32.

De manera práctica y utilizando los valores de la tabla 3:

1. Calcular el nº de aislados cuyas CMIs obtenidas por el método de prueba sean superiores a las obtenidas por el método de referencia para el rango de concentraciones de ≤2 hasta 32.

   • Ejemplo: (66+1+1+3+1+2+2)= 76

2. Calcular el número total de aislados cuyos valores de CMI por el método de referencia se encuentren en el rango ≤2 hasta 32.

   • Ejemplo: (221+48+13+3+8)= 293

3. Calcular el porcentaje de desviación por encima del método de referencia

   • Ejemplo: (76/293) x100= **25,9%**

Servicio / Unidad de Microbiología Hospital.....	Evaluación del rendimiento de un sistema de determinación de sensibilidad antibiótica	PNT-ERAC-03	
		Edición Nº 01	Página 5 de 6

4.4.2. CÁLCULO DEL PORCENTAJE DE VALORES DEL TEST DE PRUEBA CON CMIs INFERIORES AL MÉTODO DE REFERENCIA

El rango de diluciones en los que es posible que el test de prueba proporcione resultados de CMI inferiores a los de referencia es **4 hasta >32.**

De manera práctica y utilizando los valores de la tabla 3:

1. Calcular el nº de aislados cuyas CMIs obtenidas por el método de prueba sean inferiores a las obtenidas por el método de referencia para el rango de concentraciones de 4 hasta >32.

- Ejemplo: (17+8+1+2+1+3)= 32

2. Calcular el número total de aislados cuyos valores de CMI por el método de referencia se encuentren en el rango 4 hasta >32.

- Ejemplo: (48+13+3+8+7)=79

3. Calcular el porcentaje de desviación por debajo del método de referencia

- Ejemplo: (32/79) x100= **40,5%**

4.4.3. CÁLCULO DE LA DIFERENCIA ENTRE DESVIACIONES SUPERIOR E INFERIOR

Siguiendo con los valores calculados en base a las distribuciones de la tabla 3:

- **% desv sup-% desv inf= 25,9 -40,5= -14,6%**

5. OBTENCIÓN Y EXPRESIÓN DE RESULTADOS

El resultado de la evaluación deberá constar al menos del cálculo del acuerdo esencial y el sesgo.

Los valores de acuerdo esencial pueden calcularse para todos los antibióticos y microorganismos de manera combinada. Sin embargo, es recomendable que se calcule de manera independiente, al menos, para microorganismos grampositivos, gramnegativos y gramnegativos no fermentadores.

Si el sesgo se encuentra fuera del rango de valores establecido (**– 30%** hasta **+ 30%**) se puede:

a) Llevar a cabo pruebas adicionales (resolución de discrepancias) para determinar si la desviación o sesgo puede deberse al propio test o a una variación aleatoria (de acuerdo a lo descrito en el documento científico de este procedimiento apartado 3.2.9).

b) Si no puede encontrarse solución a la excesiva desviación, deberá indicarse la fuente de desviación en el etiquetado, en el informe o publicación derivada del estudio (por ejemplo, rango de CMIs en escala o grupo de microorganismos afectados).

<table>
<tr><td rowspan="2">Servicio / Unidad de Microbiología
Hospital.....</td><td rowspan="2">Evaluación del rendimiento de un sistema de determinación de sensibilidad antibiótica</td><td colspan="2">PNT-ERAC-03</td></tr>
<tr><td>Edición Nº 01</td><td>Página 6 de 6</td></tr>
</table>

6. BIBLIOGRAFÍA

1. ISO 20776-2:2021. Clinical laboratory testing and in vitro diagnostic test systems — Susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices — Part 2: Evaluation of performance of antimicrobial susceptibility test devices against reference broth micro-dilution.
2. Sharp SE, Clark RB. Verification and validation of procedures in the clinical microbiology laboratory. Cumitech 31 a. American Society for Microbiology 2009.