Aplicación del sistema GRADE a las recomendaciones de pruebas diagnósticas en la Guía Colombiana de Falla Cardíaca

Juan Manuel Sénior[1], Edison Muñoz Ortiz[2], James Samir Díaz Betancur[3]

RESUMEN

Introducción: el ejercicio clínico trae implícito un grado de incertidumbre en cuanto al diagnóstico preciso, en relación con un cúmulo de síntomas y signos en el contexto de los factores de riesgo predisponentes detectados al hacer una historia clínica cuidadosa y completa. Se requiere entonces que los clínicos estén familiarizados con las propiedades de las pruebas diagnósticas para utilizarlas en forma adecuada y no como una pesquisa infructuosa y poco específica.

Metodología: llevamos a cabo una revisión de las características operativas de las pruebas diagnósticas, incluyendo sensibilidad, especificidad, valores predictivos positivo y negativo, cocientes de verosimilitud (LR, por su sigla en inglés) positivo y negativo y curvas ROC, los cuales funcionan como base para establecer las recomendaciones de las pruebas diagnósticas en el sistema GRADE en la Guía Colombiana de Falla Cardíaca.

Resultados: explicamos cómo hacer la calificación de las pruebas diagnósticas con la metodología GRADE, teniendo en cuenta las características operativas de las pruebas, y las variables que influyen en la graduación de las recomendaciones, tales como el riesgo de sesgos, el carácter directo o no de la prueba, la presencia de inconsistencia o imprecisión y el sesgo de publicación.

Conclusión: la metodología GRADE de calificación de la evidencia en las pruebas diagnósticas permite una valoración completa de estas, tanto en sus características operativas como de aplicación en la práctica clínica.

PALABRAS CLAVE

Diagnóstico; Guía; Insuficiencia Cardíaca

[1] Coordinador del posgrado en Cardiología Clínica y Cardiología Intervencionista, Universidad de Antioquia. Cardiólogo intervencionista, Hospital Universitario San Vicente Fundación (HUSVF), Medellín, Colombia.

[2] Cardiólogo clínico, HUSVF. Profesor posgrado de Cardiología Clínica, Universidad de Antioquia, Medellín, Colombia.

[3] Médico internista, HUSVF. *Fellow* en Cardiología Clínica, Universidad de Antioquia, Medellín, Colombia.
Correspondencia: Juan Manuel Sénior; mmbt@une.net.co

Recibido: mayo 11 de 2016
Aceptado: julio 30 de 2016

SUMMARY

Application of the GRADE system to recommendations on diagnostic tests in the Colombian Guideline for Heart Failure

Introduction: Clinical practice has an implicit degree of uncertainty as to the precise diagnosis, in relation to a cluster of symptoms and signs in the context of predisposing risk factors, detected during a careful and complete medical history. It is then required that clinicians are familiar with the properties of diagnostic tests, in order to use them properly and not as a fruitless and unspecific research.

Methodology: We reviewed the operational characteristics of diagnostic tests, including sensitivity, specificity, positive and negative predictive values, positive and negative likelihood ratio (LR), and ROC curves, which serve as bases for establishing the recommendations on diagnostic tests in the GRADE system in the Colombian Guideline for Heart Failure.

Results: We explain how to make the qualification of diagnostic tests with the GRADE methodology, taking into account their operating characteristics, and the variables that influence the grading of recommendations, such as the risk of biases, the character of directness or not of the test, the presence of inconsistency or imprecision, and the publication bias.

Conclusion: GRADE methodology rating of evidence in diagnostic tests allows their complete assessment, both in the operational characteristics and in the application to clinical practice.

KEY WORD

Diagnosis; Guideline; Heart Failure

INTRODUCCIÓN

Pruebas diagnósticas

El ejercicio clínico trae implícito un grado de incertidumbre en cuanto al diagnóstico preciso, en relación con los síntomas, signos y factores predisponentes o de riesgo detectados al hacer una historia clínica cuidadosa y completa (1). Dicho ejercicio permite establecer la presencia de enfermedad o la necesidad de hacer algunas pruebas adicionales que puedan confirmarla o descartarla, pero hay que entenderlo como imperfecto, puesto que se establece la probabilidad, más que la certeza, de padecer cierta enfermedad; por tanto, los clínicos deben estar familiarizados con las propiedades de las pruebas diagnósticas para utilizarlas en forma adecuada y no como una pesquisa infructuosa y poco específica (2). Otro aspecto importante para resaltar es que estas pruebas se pueden utilizar para tamizaje (3), es decir, en individuos asintomáticos que tienen riesgo de padecer la enfermedad o como verdaderas pruebas diagnósticas en personas sintomáticas. Aunque su objetivo primordial es definir el diagnóstico, también son útiles para establecer el pronóstico, evaluar el curso clínico y la gravedad y establecer alternativas terapéuticas adecuadas (4).

Características operativas de las pruebas

Para establecer el diagnóstico de la enfermedad se requiere un patrón o estándar de referencia, que generalmente es muy costoso, difícil de realizar, invasivo o de riesgo para el paciente. En algunos casos se presenta un problema conocido como estándar de referencia imperfecto, puesto que no existe una prueba que determine con exactitud la presencia de enfermedad, por lo que se utilizan algunas estrategias que en general incluyen combinaciones del cuadro clínico y algunos métodos no invasivos o invasivos para definirlo, o incluso por un panel o consenso de expertos, creando una incertidumbre mayor, especialmente cuando la prueba que se va a evaluar hace parte de ese estándar imperfecto (5). De acuerdo con este estándar y con la prueba evaluada definimos cuatro situaciones claras: prueba positiva o negativa y enfermedad presente o ausente, lo que genera *verdaderos positivos* cuando la prueba es positiva y se padece la enfermedad, *verdaderos negativos* cuando la prueba es negativa y la enfermedad está ausente, *falsos positivos* cuando la prueba es positiva y la enfermedad está ausente y *falsos negativos* cuando la prueba es negativa y se padece la enfermedad (tabla 1).

Tabla 1. Tabla de contingencia para evaluación de prueba diagnóstica

		Patrón de referencia o estándar de oro		
		+	-	
Prueba diagnóstica	+	Verdaderos positivos (a)	Falsos positivos (b)	a+b
	-	Falsos negativos (c)	Verdaderos negativos (d)	c+d
		a+c	b+d	

Las dos primeras características operativas de una prueba son la *sensibilidad* y la *especificidad*. Se define la sensibilidad como la proporción de personas con la enfermedad que tienen la prueba diagnóstica positiva -capacidad de detectar enfermos- (a/a+c), y la especificidad como la proporción de personas sanas que tienen la prueba negativa -capacidad de detectar sanos- (d/b+d) (6,7). No necesariamente es bueno tener 100 % de sensibilidad, porque puede darse por obtener siempre un resultado positivo, de lo que se deduce que en pruebas con alta sensibilidad sería ideal obtener un resultado negativo para descartar la enfermedad por la baja tasa de falsos negativos, y en pruebas de alta especificidad sería ideal obtener resultados positivos para confirmarla por la baja ocurrencia de falsos positivos (8,9).

De acuerdo con la finalidad de la prueba se debe pensar en cuál se ajusta mejor en cuanto a sensibilidad y especificidad, es decir, si el objetivo es confirmarla o descartarla; sin embargo, una vez que los resultados de la prueba estén disponibles ya no son relevantes y lo que interesa es saber la probabilidad de tener o no la enfermedad, de acuerdo con el resultado de la prueba, por lo que se introduce el concepto de valor predictivo. El valor predictivo positivo (VPP) es la probabilidad de tener la enfermedad si el resultado de la prueba es positivo (a/a+b) y el valor predictivo negativo (VPN), la probabilidad de estar sano si la prueba es negativa (d/c+d) (10,11). Cuanto más sensible sea la prueba, mayor será el VPN, y cuanto más específica sea, mayor será su VPP; los valores predictivos están influenciados por la prevalencia de la enfermedad (probabilidad pre-prueba), por lo que influye a qué población se les aplican, es decir, si una prueba muy específica se utiliza en una población con baja probabilidad de tener la enfermedad, sus resultados serán predominantemente falsos positivos, mientras que al aumentar la prevalencia disminuye el número de falsos negativos (12); al aumentar la prevalencia el VPP aumenta y viceversa (13). Los valores predictivos también se conocen como probabilidad pos-prueba.

Por lo anteriormente expuesto, la realización e interpretación de las pruebas diagnósticas requiere una aproximación Bayesiana, en la cual se le da importancia a la probabilidad pre-prueba de enfermedad, o sea, a la prevalencia estimada antes de decidir hacer una prueba diagnóstica, pues su rendimiento dependerá de su utilización en la población adecuada. Características clínicas como la disnea de esfuerzo y la disnea paroxística nocturna y signos como galope ventricular, desplazamiento del punto de máximo impulso y congestión, en concordancia con los cambios electrocardiográficos y radiológicos, permiten establecer esa probabilidad en el síndrome de falla cardíaca. De acuerdo con este concepto, podemos observar cómo el resultado negativo o positivo de una prueba no invasiva en pacientes de alta probabilidad no cambia la probabilidad pos-prueba, por lo que se recomienda hacer una ecocardiografía para evaluar la fracción de eyección y otras pruebas específicas para determinar la etiología. En pacientes con diagnóstico dudoso por la presencia de comorbilidades, las pruebas no invasivas, como la medición de péptidos natriuréticos o la propia ecocardiografía, establecerían el diagnóstico (14,15).

Teniendo en cuenta que los conceptos de VPP y VPN son los que en realidad tienen relevancia en la práctica clínica y que dependen de la proporción de enfermos en la muestra estudiada (prevalencia), se desarrolló el concepto de cociente de probabilidad o cociente de verosimilutud (LR) positivo y negativo (16). El LR positivo expresa la relación entre los verdaderos positivos y los falsos positivos (Sensibilidad/1-Especificidad) y el LR negativo, entre los falsos negativos y los

verdaderos negativos (1-Sensibilidad/Especificidad); el cociente no cambia, puesto que depende de la sensibilidad y especificidad. Se consideran ideales un LR + mayor de 10 y un LR – menor de 0,1, aunque se consideran aceptables el positivo mayor de 5 y el negativo menor de 0,2, respectivamente (17).

Los LR también son útiles para establecer la probabilidad pos-prueba, partiendo del conocimiento de la prevalencia. Existen dos enfoques útiles, el primero es utilizando el nomograma de Fagan TJ (18) y el segundo, en forma matemática (16,19).

Para hacerlo en forma matemática se utiliza la fórmula siguiente:

Probabilidad pos-prueba = *odds* pos-prueba/(1 + *odds* pos-prueba); si tenemos una prevalencia del 35 % (probabilidad pre-prueba = a+c/a+b+c+d en la muestra o poblacional si la conocemos) y un LR+ = 5, necesitamos conocer primero los *odds* pre-prueba, que se calculan así:

Odds pre-prueba = probabilidad/(1-probabilidad), o sea, 0,35/0,65 = 0,54, de acuerdo con los datos aportados en el ejemplo; luego lo convertimos a *odds* pos-prueba:

Odds pos-prueba = LR x *odds* pre-prueba, o sea, 5 x 0,54 = 2,7; finalmente regresamos *odds* a la probabilidad que es = *odds*/(1 + *odds*), o sea, 2,7/3,7 = 0,73 (20). En resumen, si aplicamos una prueba con un LR+ de 5 en una población con prevalencia del 35 % (21) y es positiva, la probabilidad pos-prueba será del 73 %; si la aplicamos en una población con prevalencia del 2 %, ¡será tan solo del 9 %!

Si utilizamos el LR- = 0,1 tendremos lo siguiente:

Odds pos-prueba = LR x *odds* pre-prueba, o sea, 0,1 x 0,54 = 0,054; finalmente regresamos *odds* a la probabilidad que es = *odds*/(1 + *odds*), o sea, 0,054/1,054 = 0,51 y 1-0,51= 0,49, que es la probabilidad pos-prueba (49 %).

Curva ROC

La curva ROC (por la sigla en inglés de *Receiver Operating Characteristic)* fue originada en 1950 con la teoría de la detección electrónica de la señal y del ruido. En las últimas décadas se popularizó para el análisis de la exactitud de pruebas diagnósticas (22). Cuando las pruebas tienen resultados dicotómicos (positivo/negativo), el enfoque convencional es mediante la sensibilidad y especificidad, pero existen situaciones en las que los resultados se dan en escala ordinal o incluso continua, por lo que estas deben ser calculadas a lo largo de posibles valores límites conocidos como *puntos de corte* (2). La sensibilidad y la especificidad varían de acuerdo con el punto de corte y dado que la sensibilidad está inversamente relacionada con la especificad, se puede graficar sensibilidad (señal) versus 1-especificidad (ruido), para determinar el punto de corte adecuado, teniendo en cuenta el área bajo la curva (AUC), que es una medida efectiva de exactitud (23). El concepto no solo es importante para determinar el punto óptimo de corte para diagnóstico, sino que permite hacer comparaciones entre diferentes pruebas en la misma población y, adicionalmente, se utiliza para hacer los metaanálisis diagnósticos; estos últimos se presentan con una gráfica que incluye una curva ROC resumen de todos los estudios incluidos (círculos), un estimativo puntual de resumen (cuadrado), ya sea de sensibilidad o de especificidad, una región de confianza del 95 % para ese estimado puntual (línea de puntos grande) y una región de predicción del 95 %, que predice la verdadera sensibilidad o especificidad en un estudio futuro (línea de puntos pequeños) (figura 1) (24-27).

Sistema GRADE y pruebas diagnósticas

El término *graduación* o calificación hace referencia a la evaluación de la fuerza del conjunto de la evidencia que apoya una determinada recomendación, más que a la calidad de los estudios individuales que conforman dicho conjunto (28). La graduación de la calidad de la evidencia es útil para quienes hacen guías de práctica clínica, las aseguradoras, los clínicos y los propios pacientes, que pueden utilizar una síntesis de la evidencia para adoptar conductas que mejoren los desenlaces clínicos (29,30). Los grados de calidad de la evidencia les permiten a los tomadores de decisiones evaluar si una decisión se basa en una evidencia de calidad alta, moderada o baja.

Cuando se evalúa la fuerza de una cantidad (cuerpo) de evidencia con el enfoque GRADE *(Grading of Recommendations Assessment, Development, and Evaluation)* se deben considerar ocho criterios (31-33); los cinco primeros se utilizan para disminuir la calidad

de la evidencia y son: el riesgo de sesgos, la inconsistencia, el carácter indirecto de la evidencia, la imprecisión de los resultados y el sesgo de publicación (tabla 2). Los otros tres criterios se utilizan para graduar la calidad de la evidencia hacia arriba (aumentarla) y son: la relación dosis-respuesta, la existencia de factores de confusión plausibles no medidos y la fuerza de la asociación (magnitud del efecto estimado) (31,32).

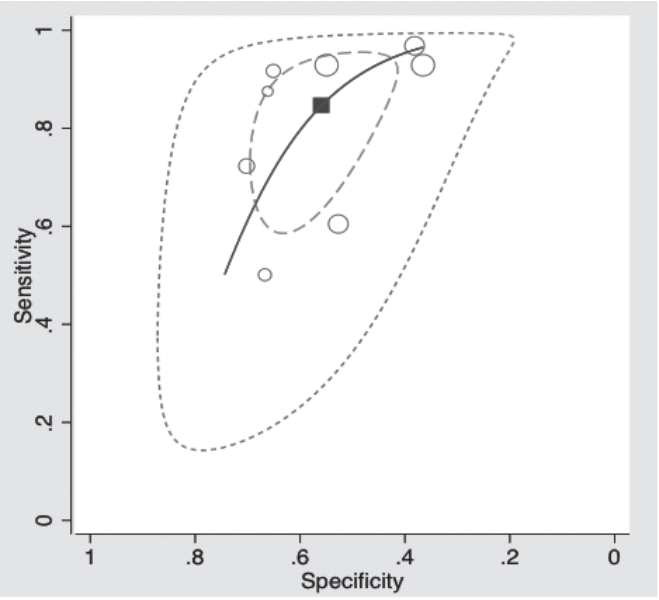

Figura 1. Metaanálisis diagnóstico. BNP en individuos con riesgo de falla cardíaca

La aplicación de la metodología GRADE a la evidencia sobre pruebas de diagnóstico representa un desafío; aunque fue creado como un instrumento genérico, GRADE se utiliza más para graduar la evidencia sobre preguntas clínicas terapéuticas (estudios de intervención) (31,33,34) que para evidencia sobre pruebas diagnósticas. En la tabla 2 se presenta un resumen comparativo de las semejanzas y diferencias del enfoque GRADE para evidencia sobre intervención y para evidencia sobre diagnóstico; este método enfatiza en la importancia de calificar la evidencia y tomar decisiones con base en el impacto que la prueba diagnóstica pueda tener sobre los desenlaces considerados de importancia para los pacientes (desenlaces centrados en el paciente) (30,35,36), más que con los desenlaces llamados intermedios o subrogados (sensibilidad, especificidad y demás características de exactitud de la prueba diagnóstica). Sin embargo, la mayoría de los estudios sobre diagnóstico tienen como desenlace principal el desempeño de las pruebas y no es frecuente encontrar estudios que evalúen el impacto de una prueba sobre desenlaces clínicos como mortalidad, morbilidad, calidad de vida y otros de importancia para los pacientes. Corresponde entonces a los revisores de la evidencia hacer un juicio clínico adecuado que logre conectar los desenlaces de exactitud de una prueba diagnóstica con los desenlaces clínicos de importancia.

Varios de los criterios GRADE para disminuir la calidad de la evidencia merecen una consideración especial cuando se aplican a pruebas diagnósticas.

Riesgo de sesgos: se dispone de múltiples instrumentos para evaluar el riesgo de sesgo de los estudios y los revisores de la evidencia deben elegir el más apropiado. Dos revisiones sistemáticas (37,38) evaluaron los

instrumentos para evaluación de la calidad en el contexto del diagnóstico; en una de ellas se encontraron tres escalas que cumplieron los criterios considerados importantes: la lista de chequeo del *Cochrane Working Group* (39), la herramienta de Lijmer y colaboradores (40) y la lista de chequeo del *National Health and Medical Research* (41). Los autores de la otra revisión no recomiendan ninguna herramienta en particular y la utilizaron como base para desarrollar su propia lista de chequeo conocida como QUADAS *(Quality Assessment of Diagnostic Accuracy Studies)* (42) que actualmente recomiendan algunas agencias para evaluar el riesgo de sesgos de estudios de pruebas diagnósticas (43).

Tabla 2. Comparación de la metodología GRADE para evidencia sobre intervención y para evidencia sobre pruebas diagnósticas

Categoría	GRADE para evidencia sobre intervención	GRADE para evidencia sobre pruebas diagnósticas
Formulación de la pregunta	Estilo PICO (**P**acientes, **I**ntervención, **C**omparación, ***O**utcomes)*	Estilo PICO, pero puede ser inadecuado debido a la falta de distinción en tipo de "pacientes" y estadios de "enfermedad"; ambos elementos pueden tener impacto en el juicio sobre la calidad de la evidencia Las pruebas diagnósticas se deben considerar en el contexto de una estrategia diagnóstica
Desenlaces *(outcomes)*	Desenlaces importantes para el paciente	Desenlaces importantes para el paciente - Medición directa (aleatorización a diferentes estrategias diagnósticas; medición de desenlaces importantes para el paciente) O, en ausencia de evidencia directa: - Medición indirecta usando desenlaces de precisión de la prueba diagnóstica (falsos positivos, falsos negativos) como subrogados de los desenlaces importantes para el paciente
Calificación de la calidad de la evidencia **Criterios GRADE para disminuir la calidad**	Estudios de intervención: por desenlace, entre los estudios	Estudios sobre precisión de pruebas diagnósticas: por desenlaces, entre los estudios
1. Riesgo de sesgos	1. Evaluar el riesgo de sesgos de cada estudio con herramientas como la de Cochrane 2. Disminuir la calidad de la evidencia con base en la calificación de todos los estudios que reporten el desenlace de interés	1. Evaluar el riesgo de sesgos de cada estudio con herramientas como QUADAS 2 2. Disminuir la calidad de la evidencia con base en la calificación de todos los estudios que reporten el desenlace de interés
2. Evidencia indirecta	Disminuir la calidad con base en problemas de aplicabilidad. Ejemplo: diferencias entre la población estudiada y aquella para la que se hace la recomendación, desenlaces subrogados o comparaciones indirectas	Disminuir la calidad con base en problemas de aplicabilidad. Ejemplo: diferencias entre la población estudiada y aquella para la que se hace la recomendación o a la que se aplicará la prueba diagnóstica La comparación de dos o más pruebas no se hace entre ellas, sino contra un estándar diagnóstico: se puede disminuir la calidad de la evidencia por comparaciones indirectas Enfoque en desenlaces importantes para el paciente: la calidad de la evidencia sobre la precisión de una prueba se puede disminuir porque se la considera un subrogado de los desenlaces importantes para el paciente

Tabla 2. Comparación de la metodología GRADE para evidencia sobre intervención y para evidencia sobre pruebas diagnósticas (continuación)

Categoría	GRADE para evidencia sobre intervención	GRADE para evidencia sobre pruebas diagnósticas
3. Inconsistencia	Disminuir la calidad si hay heterogeneidad inexplicada Los criterios incluyen similitud de las estimaciones puntuales, el grado de superposición de los intervalos de confianza y los criterios estadísticos (pruebas de heterogeneidad)	Disminuir la calidad si hay heterogeneidad inexplicada Los criterios son menos claros comparados con los estudios de intervención. Incluyen la similitud de las estimaciones puntuales y el grado de solapamiento de los intervalos de confianza. Faltan métodos estadísticos adecuados para evaluar la heterogeneidad de los estudios
4. Imprecisión	Disminuir la calidad con base en la evaluación de los intervalos de confianza, el tamaño óptimo de la información y el número de eventos	Disminuir la calidad con base en la evaluación de los intervalos de confianza de la sensibilidad y la especificidad. Aún no se dispone de guías y métodos para evaluar este aspecto. Un método podría calcular el número de falsos positivos y falsos negativos con base en una prevalencia definida de la condición que se va a diagnosticar
5. Sesgo de publicación	La calificación de este criterio es un reto Existen diferentes enfoques como el *funnel plot* y las pruebas estadísticas, con varias limitaciones, pero dan alguna orientación sobre como calificar este criterio	La calificación de este criterio es un reto Existen diferentes enfoques para estudios sobre intervenciones que no son aplicables a los estudios de pruebas diagnósticas
Criterio GRADE para graduar hacia arriba la calidad de la evidencia	Existen ejemplos específicos de situaciones en las que es apropiado graduar la calidad de la evidencia hacia arriba, como una gran magnitud del efecto	Aún no se dispone de ejemplos específicos en los que sea apropiado graduar la calidad de la evidencia hacia arriba

Carácter directo: se refiere a que la evidencia evaluada refleja un vínculo directo entre la intervención de interés (prueba diagnóstica) y el desenlace en salud analizado (28); de este modo, se podría considerar indirecta la evidencia cuando el desenlace es la exactitud de la prueba diagnóstica, como ocurre la mayoría de las veces. Sin embargo, los revisores pueden decidir calificar la solidez de la evidencia para un desenlace intermedio como sensibilidad o especificidad cuando se relacionan de manera directa con el diagnóstico de una enfermedad. El carácter directo también se puede aplicar cuando se comparan dos pruebas; por ejemplo, cuando se comparan dos o más pruebas contra un estándar diagnóstico, pero no se comparan entre ellas, se puede disminuir la calidad por comparaciones indirectas (evidencia indirecta) (44).

Inconsistencia: este criterio ofrece especial dificultad en las pruebas de diagnóstico; hace referencia a la homogeneidad en la dirección y la magnitud de los resultados entre los diferentes estudios. En evidencia sobre intervenciones se disminuye la calidad cuando hay heterogeneidad inexplicada que se puede evaluar visualmente en los *forest plot* o con pruebas estadísticas; sin embargo, los criterios estadísticos de heterogeneidad para valores de sensibilidad y especificidad son poco claros. Para algunos (44), se podría usar una representación del resumen de las curvas ROC *(Summary ROC-curve)* que muestra de manera gráfica la sensibilidad y la especificidad de varios estudios (45). De manera adicional, por lo general se observan similitud de las estimaciones puntuales y el grado de solapamiento de los intervalos de confianza. Al igual que con los estudios de intervención, la fuerza de la evidencia se reduce por la heterogeneidad que es inexplicable por diferencias en el diseño o la calidad metodológica, las características de los pacientes o el contexto de los estudios.

Imprecisión: tampoco se dispone de métodos claros para evaluar la imprecisión, que se refiere a la amplitud de los intervalos de confianza para las estimaciones de precisión diagnóstica y está estrechamente relacionada con el tamaño de la muestra (28). Por lo

general se disminuye la calidad de la evidencia con base en la evaluación de los intervalos de confianza de la sensibilidad y especificidad agrupadas; y GRADE propone calcular el número de falsos positivos (FP) y falsos negativos (FN) con base en una prevalencia definida de la condición que se va a diagnosticar (35). Antes de disminuir la calidad de la evidencia por imprecisión, los revisores deben considerar qué tanto la imprecisión de un estimado de precisión diagnóstica puede afectar los desenlaces importantes para los pacientes, y si el impacto es insignificante no se debería castigar la calidad del cuerpo de evidencia.

Sesgo de publicación: se debe comentar cuando las circunstancias sugieren que los resultados negativos o de no diferencia no se han publicado o no están disponibles; hasta la actualidad no existe una manera lo suficientemente confiable para medirlo.

En este artículo se presenta la aplicación del enfoque GRADE diagnóstico a dos preguntas para la Guía Colombiana de Falla Cardíaca (46).

METODOLOGÍA

En la guía se plantearon preguntas específicas tipo PICO que es la sigla para: **P**acientes, **I**ntervención, **C**omparación y **O**utcome (la palabra inglesa que traduce desenlace).

Pregunta 1: en pacientes mayores de 18 años con probable síndrome de falla cardíiaca, ¿cuál es la capacidad diagnóstica del péptido natriurético tipo B (BNP) y de la fracción N-terminal proBNP (NT-proBNP), comparada con el cuadro clínico o la ecocardiografía? La búsqueda de literatura arrojó un total de 104 estudios primarios y 7 revisiones sistemáticas; una de estas incluyó todos los estudios primarios, con metaanálisis para los desenlaces sensibilidad y especificidad para distintos puntos de corte tanto de BNP como de NT-proBNP y separados por escenario de atención médica (ambulatorio o urgencias). Esa revisión se utilizó para responder esta pregunta.

Pregunta 2: en pacientes mayores de 18 años con factores de riesgo para falla cardíaca, ¿cuál es la capacidad del BNP/NT-pro-BNP para el diagnóstico temprano de disfunción ventricular izquierda? En la búsqueda de literatura se identificaron 13 estudios primarios y 2 revisiones sistemáticas que podrían responder la pregunta; uno de los estudios primarios fue un ensayo clínico en el que se evaluó el impacto de la tamización con BNP en desenlaces centrados en los pacientes (aparición de disfunción ventricular y falla cardíaca).

Los revisores del grupo desarrollador de la guía (GDG) analizaron detalladamente el cuerpo de evidencia para ambas preguntas aplicando la metodología GRADE. Para los estudios en que el desenlace fue la precisión de las pruebas diagnósticas, se calificó el riesgo de sesgos con el instrumento QUADAS (42) y se evaluó de manera visual la heterogeneidad con los gráficos de resumen de curvas ROC *(Summary ROC-curve)* presentados en la revisión.

RESULTADOS

Pregunta 1

Para esta pregunta no se encontraron estudios que evaluaran desenlaces centrados en los pacientes. ***Desenlaces:*** para conectar los desenlaces de exactitud diagnóstica del BNP y el NT-proBNP con los desenlaces de importancia clínica se calcularon los FN y los FP de las pruebas usando los datos estimados de sensibilidad y especificidad de los metaanálisis encontrados. Se observó que ambas pruebas son buenas para descartar la falla cardíaca (pocos FN) tanto en urgencias como en el escenario ambulatorio; sin embargo, por el alto número de FP ninguna de las dos pruebas permite confirmar el diagnóstico de manera confiable. ***Riesgo de sesgos:*** se calificó como bajo con el instrumento QUADAS. ***Carácter indirecto:*** la evidencia se consideró directa en todos los casos porque las poblaciones incluidas en los estudios primarios de BNP y NT-proBNP corresponden al foco de la pregunta de la guía; en Colombia se dispone de equipos para procesar ambas pruebas y los conceptos de sensibilidad, especificidad, FN y FP son fácilmente entendidos y aplicables por los médicos. ***Inconsistencia:*** la calidad de la evidencia se graduó hacia abajo por considerarse no consistente para el desenlace especificidad por la presencia de heterogeneidad inexplicada y la poca superposición de los intervalos de confianza de los estudios primarios. ***Imprecisión:*** la evidencia sobre BNP y NT-proBNP en urgencias se

consideró precisa porque los intervalos de confianza de los estimados agrupados de sensibilidad y especificidad fueron estrechos; en cambio, para el BNP en el contexto ambulatorio se consideró imprecisa porque los intervalos de confianza fueron amplios. La evidencia sobre NT-proBNP en el escenario ambulatorio se consideró precisa para sensibilidad, pero imprecisa para especificidad por las mismas razones. ***Sesgo de publicación:*** no se encontró evidencia de sesgo de publicación.

Pregunta 2

Para esta pregunta se encontró un ensayo clínico con desenlaces centrados en el paciente; en este ensayo con asignación aleatoria se estudió una población con factores de riesgo bien definidos para falla cardíaca y se utilizó la medición periódica del BNP para definir la necesidad de ecocardiografía, evaluación por médico especialista y "cuidado colaborativo". ***Desenlaces:*** el desenlace primario consistió en la aparición de falla cardíaca, disfunción sistólica o diastólica del ventrículo izquierdo durante 4 años de seguimiento. ***Riesgo de sesgos:*** el ensayo se consideró con bajo riesgo de sesgos. ***Carácter indirecto:*** la evidencia se calificó como indirecta porque, a pesar de evaluar desenlaces centrados en los pacientes, el estudio en mención se hizo en una población irlandesa donde el "cuidado colaborativo" de los pacientes con factores de riesgo para falla cardíaca es diferente del cuidado en Colombia. ***Inconsistencia:*** la evidencia se consideró consistente porque los resultados del experimento clínico van en el mismo sentido que los de otros estudios que evaluaron desenlaces de precisión de la prueba diagnóstica. ***Imprecisión:*** la evidencia se consideró precisa por los intervalos de confianza estrechos de los estimados puntuales. ***Sesgo de publicación:*** no se encontraron indicios de sesgo de publicación.

DISCUSIÓN

En este documento se presentan las principales diferencias del enfoque GRADE para evidencia sobre intervención y pruebas diagnósticas, y a manera de ejemplo se presenta la aplicación de este enfoque a dos preguntas PICO sobre pruebas diagnósticas de la Guía Colombiana de Falla Cardiaca (46). La metodología GRADE resalta la importancia de los desenlaces centrados en los pacientes y en la evaluación de pruebas diagnósticas este aspecto cobra mayor importancia. Para ello hay que diferenciar la precisión diagnóstica de una prueba del impacto que puede tener su utilización en los desenlaces clínicos; cuando se encuentra evidencia sobre pruebas diagnósticas para este tipo de desenlaces la aplicación de GRADE es sencilla porque se asemeja a la de estudios de intervención. Para la pregunta 2 se encontró un ensayo clínico con desenlaces centrados en el paciente, pero para la pregunta 1 no se encontraron estudios que evaluaran el impacto en desenlaces clínicos de la utilización del BNP y el NT-proBNP.

Varios criterios GRADE, como inconsistencia, imprecisión y sesgo de publicación pueden ser difíciles de interpretar y aplicar a la evidencia sobre exactitud de una prueba diagnóstica (47). En el juicio sobre la inconsistencia para evidencia sobre intervenciones terapéuticas se afirma explícitamente que la heterogeneidad inexplicada entre los estudios es una razón para disminuir la calidad; sin embargo, la exploración de la heterogeneidad es problemática en las revisiones sobre la exactitud de una prueba diagnóstica. La falta de métodos estadísticos adecuados para evaluar la heterogeneidad en este tipo de estudios dificulta la aplicación del criterio inconsistencia (39). Para la pregunta 1, los revisores de la Guía Colombiana utilizaron el método de los gráficos de resumen de curvas ROC *(Summary ROC-curve)* (45) además del grado de superposición de los intervalos de confianza de los estudios en los metaanálisis.

La evaluación de la imprecisión generalmente incluye la percepción de qué tan amplios son los intervalos de confianza de la sensibilidad y especificidad agrupadas. Además, se pueden obtener elementos para calificar este criterio calculando el rango de individuos que se clasifican mal con la prueba diagnóstica, es decir el rango de FP y FN para una prevalencia dada de la enfermedad. Para la Guía Colombiana, los revisores hicieron uso de ambos elementos para calificar la imprecisión.

El tercer criterio GRADE que podría resultar difícil de aplicar es el sesgo de publicación, pero esto no es exclusivo de los estudios sobre pruebas diagnósticas (48) y aún falta consenso sobre los métodos para evaluar

el sesgo de publicación para un cuerpo de evidencia. A diferencia de la práctica creciente del registro de los protocolos de ensayos clínicos antes de su ejecución (49) que permite un seguimiento del sesgo de publicación, esta conciencia científica aún no existe para los estudios de pruebas de desempeño diagnóstico. En los ejemplos presentados en este documento, los revisores de la evidencia no encontraron elementos que sugirieran sesgo de publicación.

Además de todo lo anterior, la evaluación de la calidad de la evidencia sobre pruebas diagnósticas ofrece otra dificultad: las pruebas casi nunca se aplican de manera aislada en la práctica clínica y por lo general hacen parte de estrategias de diagnóstico; por ejemplo: algoritmos para diagnóstico de tromboembolia pulmonar, algoritmos para diagnóstico de falla cardíaca, entre muchos otros. En esas estrategias la realización de un examen conduce a la realización o no de otro y eso dificulta el juicio sobre el impacto clínico real de la prueba. Generalmente no se encuentran estudios que evalúen una estrategia diagnóstica más que una prueba aislada (50); aunque la Colaboración Cochrane exige la inclusión de una estrategia diagnóstica en todas las revisiones sobre desempeño de pruebas diagnósticas, no hay una guía explícita para desarrollar tal estrategia.

Es incierta la aplicabilidad de los criterios GRADE a pruebas de diagnóstico para graduar hacia arriba la calidad de la evidencia (relación dosis-respuesta, factores de confusión plausibles no medidos y fuerza de la asociación). A diferencia de lo que ocurre con la evidencia sobre intervenciones, no se dispone de ejemplos específicos en los que sea apropiado graduar hacia arriba la calidad de la evidencia sobre la exactitud de una prueba diagnóstica por estos criterios GRADE. Para la Guía Colombiana, los revisores decidieron no considerar estos aspectos para aumentar la calidad de la evidencia, pero se pueden dar algunos ejemplos teóricos de cuándo serían aplicables estos criterios. La **relación dosis-respuesta** podría apoyar la importancia potencial de algunas pruebas que miden resultados continuos y con múltiples puntos de corte (por ejemplo, la expresión de algunos genes, los niveles séricos del antígeno prostático específico o la gammagrafía de ventilación/perfusión pulmonar). El impacto de los **factores de confusión plausibles no medidos** puede ser relevante para probar estrategias que predicen desenlaces. Un estudio puede encontrar baja precisión diagnóstica debido al sesgo del espectro de gravedad de la enfermedad, aunque realmente la prueba tenga una alta precisión diagnóstica para el espectro clínico adecuado. La **magnitud de la asociación** puede ser relevante cuando se compara la precisión de dos pruebas diagnósticas y una es más precisa que el otra.

Finalmente, para dar una recomendación integral con base en la evidencia disponible, los revisores de la evidencia deben considerar los beneficios y potenciales daños de la prueba diagnóstica, los valores y preferencias de los pacientes y los recursos disponibles. Para la Guía Colombiana, además del cuerpo de evidencia, se tuvieron en cuenta la disponibilidad del BNP y el NT-proBNP, los riesgos potenciales de estas pruebas y los análisis de costo-efectividad publicados en otros países para hacer las recomendaciones sobre el uso de las pruebas para diagnóstico y tamizaje de la falla cardíaca (51).

CONCLUSIONES

La aplicación de una metodología transparente y bien estructurada como el enfoque GRADE para describir la evidencia facilita el proceso para el desarrollo, interpretación y aplicación de las recomendaciones sobre el diagnóstico clínico. GRADE es una herramienta aplicable a la evidencia sobre pruebas de diagnóstico, pero se deben tener en cuenta algunas consideraciones especiales. Aunque el cuerpo de evidencia esté centrado en el desempeño de la prueba para diagnosticar una enfermedad, se deben preferir los desenlaces centrados en los pacientes, cuando estén disponibles. Al analizar desenlaces de precisión diagnóstica, se deben enmarcar en el contexto clínico y tratar de establecer la relación entre la prueba diagnóstica y los desenlaces clínicos. Finalmente, se debe tener cuidado especial al evaluar el riesgo de sesgos, la inconsistencia, el carácter indirecto, la imprecisión y el sesgo de publicación de la evidencia sobre desempeño de pruebas para diagnóstico.

REFERENCIAS BIBLIOGRÁFICAS

1. Jaimes F. Pruebas diagnósticas: uso e interpretación. Acta Med Colomb. 2007 Ene-Mar;32(1):29-33.

2. Feltcher R, Fletcher S. Diagnosis. In: Clinical Epiemidology: The Essentials. 4th ed. Philadelphia: Lippincot Williasm and Wilkins; 2005. p. 35-58.

3. Grimes DA, Schulz KF. Uses and abuses of screening tests. Lancet. 2002 Mar;359(9309):881-4. Erratum in: Lancet. 2008 Jun;371(9629):1998.

4. Knottnerus JA, Buntinx F, van Weel C. General introduction: evaluation of diagnostic procedures. In: Knottnerus JA, Buntinx F, editors. The Evidence Base of Clinical Diagnosis: Theory and methods of diagnostic research. 2nd ed. New Jersey: Wiley; 2011. p. 1-19.

5. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007 Dec;11(50):iii, ix-51.

6. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008 Jan-Feb;56(1):45-50.

7. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994 Jun;308(6943):1552.

8. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ. 2003 Sep;327(7417):716-9. Erratum in: BMJ. 2003 Nov;327(7422):1043.

9. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr. 2007 Mar;96(3):338-41.

10. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994 Jul;309(6947):102.

11. Stojanovic M, Apostolovic M, Stojanovic D, Milosevic Z, Toplaovic A, Lakusic V, et al. Understanding sensitivity, specificity and predictive values. Vojnosanit Pregl. 2014 Nov;71(11):1062-5.

12. Collier J, Huebscher R. Sensitivity, specificity, positive and negative predictive values: diagnosing purple mange. J Am Acad Nurse Pract. 2010 Apr;22(4):205-9. DOI 10.1111/j.1745-7599.2010.00496.x.

13. Chu K. An introduction to sensitivity, specificity, predictive values and likelihood ratios. Emerg Med. 1999 Sep;11(3):175-81. DOI 10.1046/j.1442-2026.1999.00041.x.

14. McMurray JJ, Adamopoulos S, Anker SD, Auricchio A, Böhm M, Dickstein K, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. Eur Heart J. 2012 Jul;33(14):1787-847. DOI 10.1093/eurheartj/ehs104. Erratum in: Eur Heart J. 2013 Jan;34(2):158.

15. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2013 Oct;62(16):e147-239. DOI 10.1016/j.jacc.2013.05.019.

16. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet. 2005 Apr 23-29;365(9469):1500-5.

17. Parikh R, Parikh S, Arun E, Thomas R. Likelihood ratios: clinical application in day-to-day practice. Indian J Ophthalmol. 2009 May-Jun;57(3):217-21. DOI 10.4103/0301-4738.49397.

18. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med. 1975 Jul;293(5):257.

19. McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002 Aug;17(8):646-9.

20. Spitalnic S. Test properties 2: Likelihood ratios, Bayes' formula, and receiver operating characteristic curves. Hosp Physician. 2004 Oct:53-8.

21. Ledwidge M, Gallagher J, Conlon C, Tallon E, O'Connell E, Dawkins I, et al. Natriuretic peptide-based screening and collaborative care for heart failure: the STOP-HF randomized trial. JAMA. 2013 Jul;310(1):66-74. DOI 10.1001/jama.2013.7588.

22. Lusted LB. Logical analysis in roentgen diagnosis. Radiology. 1960 Feb;74:178-93.

23. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013 Spring;4(2):627-35.

24. Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. Evid Based Ment Health. 2015 Nov;18(4):103-9. DOI 10.1136/eb-2015-102228.

25. Putter H, Fiocco M, Stijnen T. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. Biom J. 2010 Feb;52(1):95-110. DOI 10.1002/bimj.200900073.

26. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. BMC Med Res Methodol. 2009 Nov;9:73. DOI 10.1186/1471-2288-9-73.

27. Riley RD, Ahmed I, Ensor J, Takwoingi Y, Kirkham A, Morris RK, et al. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. Syst Rev. 2015 Feb;4:12. DOI 10.1186/2046-4053-4-12.

28. Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions--agency for healthcare research and quality and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):513-23. DOI 10.1016/j.jclinepi.2009.03.009.

29. Atkins D, Fink K, Slutsky J; Agency for Healthcare Research and Quality; North American Evidence-based Practice Centers. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med. 2005 Jun;142(12 Pt 2):1035-41.

30. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ [Internet]. 2008 May [cited 2016 Feb 17];336(7653):[1106-10]. Available from:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2386626&tool=pmcentrez&rendertype=abstract

31. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ [Internet]. 2008 Apr [cited 2014 Jul 29];336(7650): [924-6]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2335261&tool=pmcentrez&rendertype=abstract

32. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. DOI 10.1016/j.jclinepi.2010.07.015.

33. Thornton J, Alderson P, Tan T, Turner C, Latchem S, Shaw E, et al. Introducing GRADE across the NICE clinical guideline program. J Clin Epidemiol. 2013 Feb;66(2):124-31. DOI 10.1016/j.jclinepi.2011.12.007.

34. Treweek S, Oxman AD, Alderson P, Bossuyt PM, Brandt L, Brożek J, et al. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results. Implement Sci [Internet]. 2013 Jan [cited 2016 Apr 15];8:[6]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3553065&tool=pmcentrez&rendertype=abstract

35. Hsu J, Brożek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implement Sci. 2011 Jun;6:62. DOI 10.1186/1748-5908-6-62.

36. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009 Aug;64(8):1109-16. DOI 10.1111/j.1398-9995.2009.02083.x.

37. West S, King V, Carey T, Lohr K, McKoy N, Sutton S, et al. 47 Systems to Rate the Strength of Scientific Evidence: Summary. In: AHRQ Evidence Report Summaries [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 1998-2005 [cited 2016 Apr 17]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK11930/

38. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol [Internet]. 2005 Jan [cited 2016 Apr 17];58(1):[1-12]. Available from: http://www.sciencedirect.com/science/article/pii/S0895435604001659

39. Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. Syst Rev [Internet]. 2013 Oct [cited 2016 Mar 26];2:[82]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3851548&tool=pmcentrez&rendertype=abstract

40. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA [Internet]. 1999 Sep [cited 2016 Apr 17];282(11):[1061-6]. Available from: http://jama.jamanetwork.com/article.aspx?articleid=191668

41. Glasziou PP, Irwig L, Bain CJ, Colditz GA. How to review the evidence: systematic identification and review of the scientific literature. Canberra: National Health and Medical Research Council; 2000.

42. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol [Internet]. 2003 Nov [cited 2015 Mar 8];3:[25]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=305345&tool=pmcentrez&rendertype=abstract

43. Santaguida PL, Riley CM, Matchar DB. Assessing Risk of Bias as a Domain of Quality in Medical Test Studies. In: Chang SM, Matchar DB, Smetana GW, Umscheid CA, editors. Methods Guide for Medical Test Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 [cited 2015 Mar 8]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK98233/

44. Singh S, Chang SM, Matchar DB, Bass EB. Chapter 7: grading a body of evidence on diagnostic tests. J Gen Intern Med [Internet]. 2012 Jun [cited 2016 Apr 17];27 Suppl 1:[S4755]. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3364356&tool=pmcentrez&rendertype=abstract

45. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001 Oct;20(19):2865-84.

46. Ministerio de Salud y Protección Social, Departamento Administrativo de Ciencia, Tecnología e Innovación, Instituto de evaluación Tecnológica en Salud. Guía de Práctica Clínica para la prevención, diagnóstico, tratamiento y rehabilitación de la falla cardíaca en población mayor de 18 años clasificación B, C y D. [Internet]. Bogotá: Ministerio de Salud y Protección Social; 2015 [consultado 2016 Abr 17]. Disponible en: http://gpc.minsalud.gov.co/guias/Documents/Cardiaca/GPC_FallaCardiaca_Socializacion08052015.pdf

47. Gopalakrishna G, Mustafa RA, Davenport C, Scholten RJ, Hyde C, Brozek J, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. J Clin Epidemiol [Internet]. 2014 Jul [cited 2016 Mar 1];67(7):[760-8]. Available from: http://www.sciencedirect.com/science/article/pii/S0895435614000444

48. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. J Clin Epidemiol [Internet]. 2011 Dec [cited 2016 Apr 15];64(12):[1277-82]. Available from: http://www.sciencedirect.com/science/article/pii/S0895435611001818

49. Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, Godlee F, et al. Clinical trial registration--looking back and moving ahead. N Engl J Med [Internet]. 2007 Jun [cited 2016 Apr 15];356(26):[2734-6]. Available from: http://www.nejm.org/doi/full/10.1056/NEJMe078110#t=article

50. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol [Internet]. 2011 Apr [cited 2016 Apr 13];64(4):[395-400]. Available from: http://www.sciencedirect.com/science/article/pii/S0895435610003318

51. Singh S, Chang SM, Matchar DB, Bass EB. Grading a body of evidence on diagnostic tests. In: Chang SM, Matchar DB, editors. Methods guide for medical test reviews [Internet]. Rockville, MD: Agency for Healthcare Research and Quality; 2012 [cited 2016 May 2]. Available from: https://www.effectivehealthcare.ahrq.gov/ehc/products/246/558/Methods-Guide-for-Medical-Test-Reviews_Full-Guide_20120530.pdf